Psychonomic Bulletin & Review
2005, 12 (4), 573-604

THEORETICAL AND REVIEW ARTICLES

An introduction to Bayesian hierarchical models with an application in the theory of signal detection

JEFFREY N. ROUDER
University of Missouri, Columbia, Missouri

and

JUN LU
American University, Washington, D.C.

Although many nonlinear models of cognition have been proposed in the past 50 years, there has been little consideration of corresponding statistical techniques for their analysis. In analyses with nonlinear models, unmodeled variability from the selection of items or participants may lead to asymptotically biased estimation. This asymptotic bias, in turn, renders inference problematic. We show, for example, that a signal detection analysis of recognition memory data leads to asymptotic underestimation of sensitivity. To eliminate asymptotic bias, we advocate hierarchical models in which participant variability, item variability, and measurement error are modeled simultaneously. By accounting for multiple sources of variability, hierarchical models yield consistent and accurate estimates of participant and item effects in recognition memory. This article is written in tutorial format; we provide an introduction to Bayesian statistics, hierarchical modeling, and Markov chain Monte Carlo computational techniques.

In experimental science, it is desirable to hold all factors constant except those intentionally manipulated. In psychology, however, this ideal is often not possible. Elements such as participants and items vary, in addition to the intended factors. For example, a researcher interested in the psychology of reading might manipulate the part of speech and observe reading times. In this case, there is unintended variability from the selection of both participants and items. In his classic article, "The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research," H. H. Clark (1973) discussed how unintended variability from the simultaneous selection of participants and items leads to underestimation of confidence intervals and inflation of Type I error rates in conventional analysis. Type I error rate inflation, or an increased tendency to find a significant effect when none exists, is highly undesirable.

To demonstrate the problem, consider the question of whether nouns and verbs are read at the same rate. To answer this question, a researcher could randomly select suitable verbs and nouns and ask a number of participants to read them. Each participant produces a set of reading time scores for both nouns and verbs. A common approach is to tabulate for each participant one mean reading time for nouns and another for verbs. To test the hypothesis of the equality of reading rates, these pairs of mean reading times may be submitted to paired *t* tests. This analytic approach is often used in memory research. For example, Riefer and Rouder (1992) used this analysis to determine whether bizarre sentences are better remembered than common ones. Clark (1973), however, argued that using *t* tests to analyze means tabulated across different items leads to Type I error rate inflation.

In the following demonstration, we show by simulation that this inflation is not only real, but also surprisingly large. We generate data for a standard ANOVA-style model (discussed below) with no part-of-speech effects. We analyze these data by first computing participant means for each part of speech and then submitting these means to a paired *t* test. This process is performed repeatedly, and the proportion of significant results is reported. If the test has no Type I error inflation, the proportion should be the nominal Type I error rate, which is set to the conventional value of .05.

Consider the following ANOVA-style model for nouns: It is reasonable to expect that each participant has a unique effect on reading time; some participants are fast at reading, but others are slow. This effect for the *i*th participant

This research is supported by NSF Grant SES-0095919 to J.N.R., Dongchu Sun, and Paul Speckman. We thank Dongchu Sun and Paul Speckman for many intensive conversations, and Andrew Heathcote, Trisha Van Zandt, and Richard Morey for helpful comments on a previous draft. Correspondence relating to this article may be sent to J. N. Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211 (e-mail: rouderj@missouri.edu).

is denoted α_i. Likewise, it is reasonable to expect that each item has a unique effect on reading time; some items are read quickly, and others are read slowly. This effect for the jth item is denoted β_j. Reading times reflect both the participant and item effects, as well as noise:

$$N_{ij} = \mu_{\mathrm{n}} + \alpha_i + \beta_j + \varepsilon_{ij}^{(\mathrm{n})}, \tag{1}$$

where N_{ij} is the ith participant's reading time on the jth noun, μ_{n} is a grand reading time for nouns, α_i and β_j are participant and item effects, respectively, and $\varepsilon_{ij}^{(\mathrm{n})}$ is any additional noise. Random variables $\varepsilon_{ij}^{(\mathrm{n})}$ are independent normals centered around 0, with equal variances. Equation 1 is a familiar additive form that underlies both ANOVA and regression. For this paradigm, participant and item effects should be treated as random. It is reasonable to model them as independent random draws from normal distributions:

$$\alpha_i \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma_1^2\right), \tag{2}$$

$$\beta_j \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma_2^2\right), \tag{3}$$

and

$$\varepsilon_{ij}^{(\mathrm{n})} \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma^2\right). \tag{4}$$

In these equations, the symbol "~" is used to denote a distribution and may be read *is distributed as*. When more than one random variable is assigned, as is the case above, and they are independent, $\overset{\text{ind}}{\sim}$ will be used to denote this relationship. The model for nouns is similar to conventional "within-subjects" or repeated measures models used in a standard ANOVA. The difference is that whereas in conventional models only participants are treated as random effects, in the present model both participants and items are simultaneously treated as random effects.

An analogous model is placed on verbs:

$$V_{ij} = \mu_{\mathrm{v}} + \alpha_i + \gamma_j + \varepsilon_{ij}^{(\mathrm{v})}, \tag{5}$$

where V_{ij} is the ith participant's reading time on the jth verb, μ_{v} is a grand reading time for verbs, α_i and γ_j are participant and item effects, respectively, and $\varepsilon_{ij}^{(\mathrm{v})}$ is any additional noise. These random effects are modeled analogously:

$$\gamma_j \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma_2^2\right) \tag{6}$$

and

$$\varepsilon_{ij}^{(\mathrm{v})} \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma^2\right). \tag{7}$$

We simulated data from this model and performed the conventional analysis on aggregated means. Vector notation is helpful in describing the simulations. Let $\boldsymbol{\alpha}$ denote the vector of all subject random effects, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_I)$. (Boldface type is reserved for vectors and matrices.) The goal is to assess Type I error rates; consequently, data were simulated with no true difference in reading times between nouns and verbs ($\mu_{\mathrm{n}} = \mu_{\mathrm{v}}$). Each replicate of the simulation starts with simulating participant random effects ($\boldsymbol{\alpha}$), noun-item random effects ($\boldsymbol{\beta}$), and verb-item random effects ($\boldsymbol{\gamma}$) as draws from normal distribution. In our first simulation, there were 50 hypothetical participants, each observing 50 nouns and 50 verbs. Hence, there were 100 values of $\boldsymbol{\alpha}$ and 50 values each of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. These random effects were kept constant throughout a replication. Next, the values of the noise, $\boldsymbol{\varepsilon}^{(\mathrm{n})}$ and $\boldsymbol{\varepsilon}^{(\mathrm{v})}$, were sampled. There was a total of 5,000 samples for each type of noise, one for each participant-by-item combination. Then, scores **N** and **V** were computed by adding the grand means, random effects, and noise in accordance with Equations 1 and 5. Mean scores for each participant in each part-of-speech condition were tabulated and submitted to a paired t test. There were 500 independent replicates per simulation.

We performed several simulations by manipulating σ_1^2 and σ_2^2 (σ^2 was set to 1 and serves to scale the other parameter values). The main result is that the real Type I error rate is a function of item variability (σ_2^2). Figure 1 shows the proportion of Type I errors (significant t test results) as a function of item variability. The filled circles are error rates for 50 hypothetical participants observing 50 nouns and 50 verbs; the circles with hatched lines are error rates for 20 hypothetical participants observing 20 nouns and 20 verbs. With no item variability, the Type I error rate equals the nominal value of .05. As item variability is increased, however, the Type I error rate increases, and does so dramatically. For example, for the simulation of the larger experiment, when item variability is only one third that of σ^2, the real Type I error rate is around .40. This is a surprisingly high rate.

The intuitive reason for the increased Type I error rate goes as follows. For each replicate, the aggregate item scores, mean($\boldsymbol{\beta}$) and mean($\boldsymbol{\gamma}$), vary. This variation affects all participants equally. In effect, this variation induces a correlation across participants. If a sampled set of items is a bit quicker than usual, all participants will be equally affected. This correlation violates the independent-observations assumption of the t test. It is not surprising, then, that there is an increase in Type I error rate.

The analysis above is termed *participant analysis*, since the data were aggregated across items to produce participant-specific scores (Baayen, Tweedie, & Schreuder, 2002). One alternative would be *item analysis*, in which data are aggregated across participants. A mean reading score is then tabulated for each item, and the mean scores are submitted to an appropriate t test for inference. Unfortunately, the Type I error rate of this t test is inflated by participant variability. Another alternative is to perform both item and participant analyses. Unfortunately, this alternative is also flawed, for if there is both item and participant variability, each of these tests has an inflated Type I error rate.

There are valid statistical procedures for this problem. Clark (1973) proposed a correction, a quasi-F statistic, that accounts for item variability. This correction works well (Forster & Dickinson, 1976; Raaijmakers, Schrijnemakers, & Gremmen, 1999). More recently, Baayen

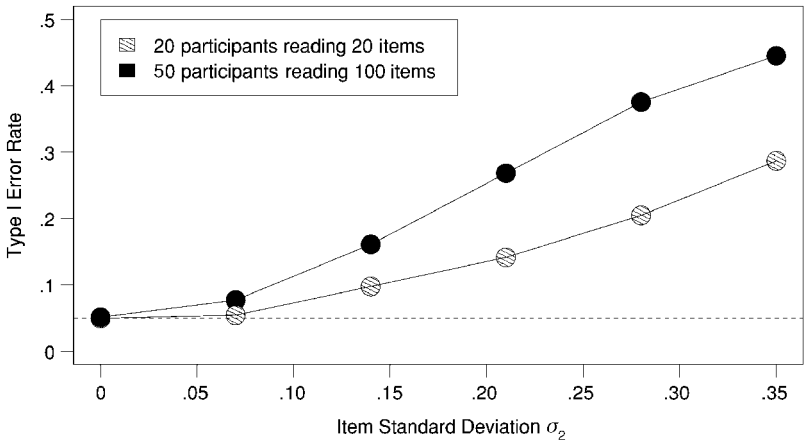

Figure 1. The effect of unmodeled item variability (σ_2) on Type I error rate when data are aggregated across items. All error rates were computed for a nominal Type I error rate of .05.

et al. (2002) proposed a mixed linear model that accounts for both participant and item variation. Experimentalists, however, have a more intuitive approach: replication. The more a finding is replicated, the lower the chance that it is due to a Type I error. For example, consider two independent researchers who replicate each other's experiment at a nominal Type I error rate of .05. Assume that due to unaccounted item variability, the actual Type I error rate for each replication is .2. If both replications are significant, the combined Type I error rate is .04, which is below the nominal criterion. Psychologists do not often use strict replication, but instead use *near replications* in which there are only minor procedural differences across replicates. Consider the bizarre memory example above. Although Riefer and Rouder (1992) used aggregation to conclude that bizarre sentences are better recalled than common ones, the basic finding has been obtained repeatedly (see, e.g., Einstein, McDaniel, & Lackey, 1989; Hirshman, Whelley, & Palij, 1989; Pra Baldi, de Beni, Cornoldi, & Cavedon, 1985; Wollen & Cox, 1981), so it is surely not the result of a Type I error. Oft-replicated phenomena, such as the Stroop effect and semantic priming effects, are certainly not spurious.

The reason that replication is feasible in linear contexts (such as those underlying both *t* tests and ANOVA) is that population means can be estimated without bias, even when there is unmodeled variability. For example, in our simulation of Type I error rates, estimates of true condition means were quite accurate and showed no apparent bias. Consequently, the true difference between two groups can be estimated without bias and may be obtained with greater accuracy by increasing sample size. In the case in which there is no true difference, increasing the sample size yields increasingly better estimates of the null group difference.

The situation is not nearly so sanguine for nonlinear models. Examples of nonlinear models include signal detection (Green & Swets, 1966), process dissociation (Egan, 1975; Jacoby, 1991), the diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998), the fuzzy logical model of perception (Massaro & Oden, 1979), the hybrid deadline model (Rouder, 2000), the similarity choice model (Luce, 1963), the generalized context model (Medin & Schaffer, 1978; Nosofsky, 1986), and the interactive activation model (McClelland & Rumelhart, 1981). In fact, almost all models used for cognition and perception outside the ANOVA/regression framework are nonlinear. Nonlinear models are postulated because they are more realistic than linear models. Even though psychologists have readily adopted nonlinear models, they have been slow to acknowledge the effects of unmodeled variability in these contexts. These effects are not good: Unmodeled variability often yields distorted parameter estimates in many nonlinear models. For this reason, the assessment of true differences in these models is difficult.

EFFECTS OF VARIABILITY IN SIGNAL DETECTION

To demonstrate the effects of unmodeled variability on a nonlinear model, we performed a set of simulations in which the theory of signal detection (Green & Swets, 1966) was applied to a recognition memory paradigm. Consider the analysis of a recognition memory experiment that entails both randomly selected participants and items. In the signal detection model, participants monitor the familiarity of a target. If familiarity is above criterion, participants report the target as an old item; otherwise, they report it as a new item. The distribution of familiarity is assumed to be greater for old than for new items, and the degree of this difference is the *sensitivity*. Hits occur when old items are judged as *old*; false alarms occur when new items are judged as *old*. The model is shown in Figure 2. It is reasonable to expect that there is

participant-level variability in sensitivity; some participants have better mnemonic abilities than others. It is reasonable also to expect item-level variability; some items are easier to remember than others.

To explore the effects of variability, we implemented a simulation similar to the previous one. Let d'_{ij} denote the *i*th individual's sensitivity to the *j*th item; likewise, let c_{ij} denote the *i*th individual's criterion when assessing the familiarity of the *j*th item (c_{ij} is measured as the distance from the mean of the new-item distribution; see Figure 2). Consider the following model:

$$d'_{ij} = \mu_d + \alpha_i + \beta_j \tag{8}$$

and

$$c_{ij} = \mu_c + \alpha_i/2 + \beta_j/2. \tag{9}$$

Parameters μ_d and μ_c are grand means, parameter α_i denotes the effect of the *i*th participant, and parameter β_j denotes the effect of the *j*th item. The model on d' is additive for items and participants; in this sense, it is analogous to the previous model for reading times. One odd-looking feature of this simulation model is that participant and item effects are half as large on criteria as they are on sensitivity. This feature is motivated by Glanzer, Adams, Iverson, and Kim's (1993) mirror effect. The *mirror effect* refers to an often-observed pattern in which manipulations that increase sensitivity do so by both increasing hit rates and decreasing false alarms. Consider the case of unbiased responding shown in Figure 2. In the figure, the criterion is half of the value of the sensitivity. If a manipulation were to increase sensitivity by increasing the hit rate and decreasing the false alarm rate in equal increments, then criterion must increase half as much as the sensitivity gain (like sensitivity, criterion is measured from 0, the mean of the new-item distribution). In Equations 8 and 9, the particular effect of a participant or item is balanced in hit and false alarm rates. If a particular participant or item is associated with high sensitivity, the effect is equally strong in hit and false alarm rates.

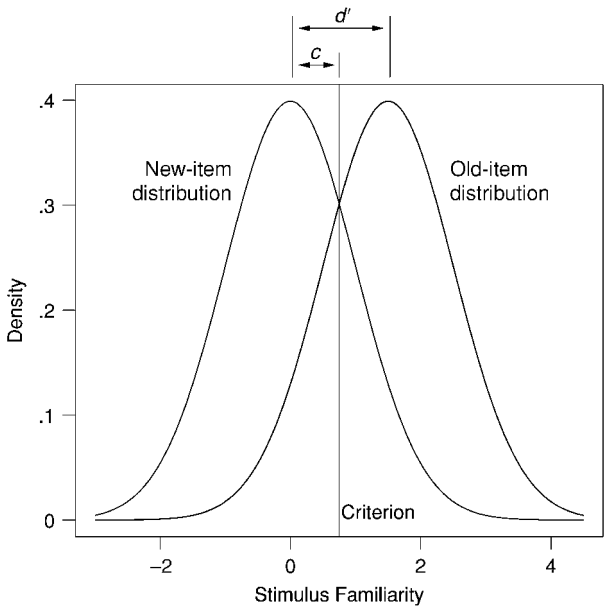

Figure 2. The signal detection model. Hit and false alarm probabilities are the areas of the old-item and new-item familiarity distributions that are greater than the criterion, respectively. Values of sensitivity (d') and criterion (c) are measured from 0, the mean of the new-item familiarity distribution.

Because α_i and β_j denote the effects from randomly selected participants and items, it is reasonable to model them as random effects. These are given by

$$\alpha_i \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma_1^2\right) \tag{10}$$

and

$$\beta_j \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma_2^2\right). \tag{11}$$

We simulated data in a manner similar to the previous simulation. First, item and participant effects were sampled according to Equations 10 and 11. Then these random effects were combined according to Equations 8 and 9 in order to produce underlying sensitivities and criteria for each participant/item combination. On the basis of these true values, hypothetical responses were produced in accordance with the signal detection model shown in Figure 2. These hypothetical responses are dichotomous: A specific participant judges a specific item as either *old* or *new*. These responses were aggregated across items to produce hit and false alarm rates for each individual. From these aggregated rates, d' and c were computed for each participant. Because individualized estimates of d' were computed, variability across participants was not problematic (σ_1 was set equal to .5). The left panel of Figure 3 shows the results of a simulation with no item effects ($\sigma_2 = 0$). As expected, d' estimates appear unbiased. The right panel shows a case with item variability ($\sigma_2 = 1.5$). The estimates systematically underestimate true values. This underestimation is an asymptotic bias; it does not diminish as the numbers of participants and items increase. We implemented variability in the context of a mirror effect, but the underestimation is also obtained in other implementations. Simulations with item variability exclusively in hits result in the same underestimation. Likewise, simulations with item variability in c alone result in the same underestimation (see Wickelgren, 1968, who makes a comparable argument).

The presence of asymptotic bias is disconcerting. Unlike analysis with ANOVA, replication does not guarantee correct inference. In the case of signal detection, one cannot tell whether a difference in overall estimated sensitivity between two conditions is due to a true sensitivity difference or to an increase in unmodeled variability in one condition relative to the other. One domain in which asymptotic underestimation is particularly perni-

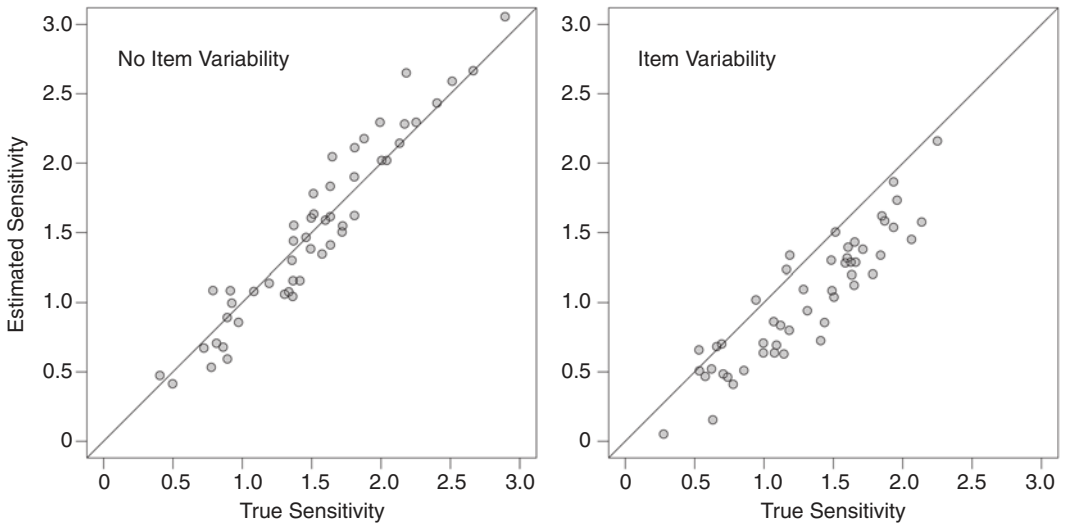

Figure 3. The effect of unmodeled item variability (σ_2) on the estimation of sensitivity (d'). True values of d' are the means of d'_{ij} across items. The left and right panels, respectively, show the results without and with item variability ($\sigma_2 = 1.5$). For both plots, the value of σ_1 was .5.

cious is signal detection analyses of subliminal semantic activation. Greenwald, Draine, and Abrams (1996), for example, used signal detection sensitivity aggregated across items to claim that a set of word primes was subliminal ($d' \approx 0$). These subliminal primes then affect response times to subsequent stimuli. The asymptotic downward bias from item variability, however, renders suspect the claim that the primes were truly subliminal.

The presence of asymptotic bias with unmodeled variability is not unique to signal detection. In general, unmodeled variability leads to asymptotic bias in nonlinear models. Psychologists have long known about problems associated with aggregation in specific contexts. For example, Estes (1956) warned about aggregating learning curves across participants. Aggregate learning curves may be more graded than those of individuals, leading researchers to misdiagnose underlying mechanisms (Haider & Frensch, 2002; Heathcote, Brown, & Mewhort, 2000). Curran and Hintzman (1995) critiqued aggregation in Jacoby's (1991) process dissociation procedure. They showed that aggregating responses over participants, items, or both possibly leads to asymptotic underestimation of automaticity estimates. Ashby, Maddox, and Lee (1994) showed that aggregating data across participants possibly distorts estimates from similarity choice-based scaling (such as those from Gilmore, Hersh, Caramazza, & Griffin, 1979). Each of the examples above is based on the same general problem: Unmodeled variability distorts estimates in a nonlinear context. In all of these cases, the distortion does not diminish with increasing sample sizes. Unfortunately, the field has been slow to grasp the general nature of this problem. The critiques above were made in isolation and appeared as specific critiques of specific models. Instead, we see them as different instances of the negative effects of aggregating data in analyses with nonlinear models.

The solution is to model both participant and item variability simultaneously. We advocate Bayesian hierarchical models for this purpose (Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder, Sun, Speckman, Lu, & Zhou, 2003). In a hierarchical model, variability on several levels, such as from the selection of items and participants as well as from measurement error, is modeled simultaneously. As was mentioned previously, hierarchical models have been used in linear contexts to improve power and to better control Type I error rates (see, e.g., Baayen et al., 2002; Kreft & de Leeuw, 1998). Although better statistical control is attractive to experimentalists, it is not critical. Instead, it is often easier (and wiser) to replicate results than to learn and implement complex statistical methodologies. For nonlinear models, however, hierarchical models are critical, because bias is asymptotic.

The main drawback to nonlinear hierarchical models is tractability. They are difficult to implement. Starting in the 1980s and 1990s, however, new computational techniques, based on Bayesian analysis, emerged in statistics. The utility of these new techniques is immense; they allow for the analysis of previously intractable models, including nonlinear hierarchical ones (Gelfand & Smith, 1990; Gelman, Carlin, Stern, & Rubin, 2004; Gill, 2002; Tanner, 1996). These new Bayesian-based techniques have had tremendous impact in many quantitatively oriented disciplines, including statistics, engineering, bioinformatics, and economics.

The goal of this article is to provide an introduction to Bayesian hierarchical modeling. The focus is on a relatively new computational technique that has made

Bayesian analysis more tractable: Markov chain Monte Carlo sampling. In the first section, basic issues of estimation are discussed. Then this discussion is expanded to include the basics of Bayesian estimation. Following that, Markov chain Monte Carlo integration is introduced. Finally, a series of hierarchical signal detection models is presented. These models provide unbiased and relatively efficient sensitivity estimates for the standard equal-variance signal detection model, even in the presence of both participant and item variability.

ESTIMATION

It is easier to discuss Bayesian estimation in the context of a simple example. Consider the case in which a single participant performs N trials. The result of each trial is either a success or a failure. We wish to estimate the underlying true probability of a success, denoted p, for this single participant. Estimating a probability of success is essential to many applications in cognitive psychology, including signal detection analysis. In signal detection, we typically estimate hit and false alarm rates. Hits are successes on old-item trials, and false alarms are failures on new-item trials.

A formula that provides an estimate is termed an *estimator*. A reasonable estimator for this case is

$$\hat{p}_0 = \frac{y}{N}, \tag{12}$$

where y is the number of successes out of N trials. Estimator $\hat{p}_0$ is natural and straightforward.

To evaluate the usefulness of estimators, statisticians usually discuss three basic properties: bias, efficiency, and consistency. Bias and efficiency are illustrated in Table 1. The data are the results of weighing a hypothetical person of 170 lb on two hypothetical scales four separate times. *Bias* refers to the mean of repeated estimates. Scale A is unbiased because the mean of the estimates equals the true value of 170 lb. Scale B is biased. The mean is 172 lb, which is 2 lb greater than the true value of 170 lb. Scale B, however, has a smaller degree of error than does Scale A, so Scale B is termed more *efficient* than Scale A. Efficiency is the inverse of the expected error of an observation and may be indexed by the reciprocal of root mean squared error. Bias and efficiency have the same meaning for estimators as they do for scales. *Bias* refers to the difference between the average value of an estimator and a true value. *Efficiency* refers to the average magnitude of the difference between an estimator and a true value. In many situations, efficiency determines the quality of inference more than bias does.

Table 1
Weight of a 170-Pound Person Measured on Two Hypothetical Scales

	Scale A	Scale B
Data (lb)	180, 160, 175, 165	174, 170, 173, 171
Mean	170	172
Bias	0	2.0
RMSE	7.91	2.55

How biased and efficient is estimator $\hat{p}_0$? To provide a context for evaluation, consider the following two alternative estimators:

$$\hat{p}_1 = \frac{y + .5}{N + 1} \tag{13}$$

and

$$\hat{p}_2 = \frac{y + 1}{N + 2}. \tag{14}$$

These two alternatives may seem unprincipled, but, as is discussed in the next section, they are justified in a Bayesian framework.

Figure 4 shows sampling distributions for the three probability estimators when the true probability is $p = .7$ and there are $N = 10$ trials. Estimator $\hat{p}_0$ is not biased, but estimators $\hat{p}_1$ and $\hat{p}_2$ are. Surprisingly, estimator $\hat{p}_2$ has the lowest average error; that is, it is the most efficient. The reason is that the tails of the sampling distribution are closer to the true value of $p = .7$ for $\hat{p}_2$ than for $\hat{p}_1$ or $\hat{p}_0$. Figure 4 shows the case for a single true value of $p = .7$. Figure 5 shows bias and efficiency for all three estimators for the full range of p. The conventional estimator $\hat{p}_0$ is unbiased for all true values of p, but the other two estimators are biased for extreme probabilities. None of the estimators is always more efficient than the others. For intermediate probabilities, estimator $\hat{p}_2$ is most efficient; for extreme probabilities, estimator $\hat{p}_0$ is most efficient. Typically, researchers have some idea of what type of probability of success to expect in their experiments. This knowledge can therefore be used to help pick the best estimator for a particular situation.

The other property of estimators is consistency. A *consistent* estimator converges to its true value. As the sample size is increased, the estimator not only becomes unbiased, the overall variance shrinks toward 0. Consistency should be viewed as a necessary property of a good estimator. Fortunately, many estimators used in psychology are consistent, including sample means, sample variances, and sample correlation coefficients. The three estimators, $\hat{p}_0$, $\hat{p}_1$, and $\hat{p}_2$, are consistent; with sufficient data, they converge to the true value of p. In fact, all estimators presented in this article are consistent.

BAYESIAN ESTIMATION OF A PROBABILITY

In this section, we provide a Bayesian approach to estimating the probability of success. The datum in this application is the number of successes (y), and this quantity may be modeled as a binomial:

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

where N is known and p is the unknown parameter of interest.

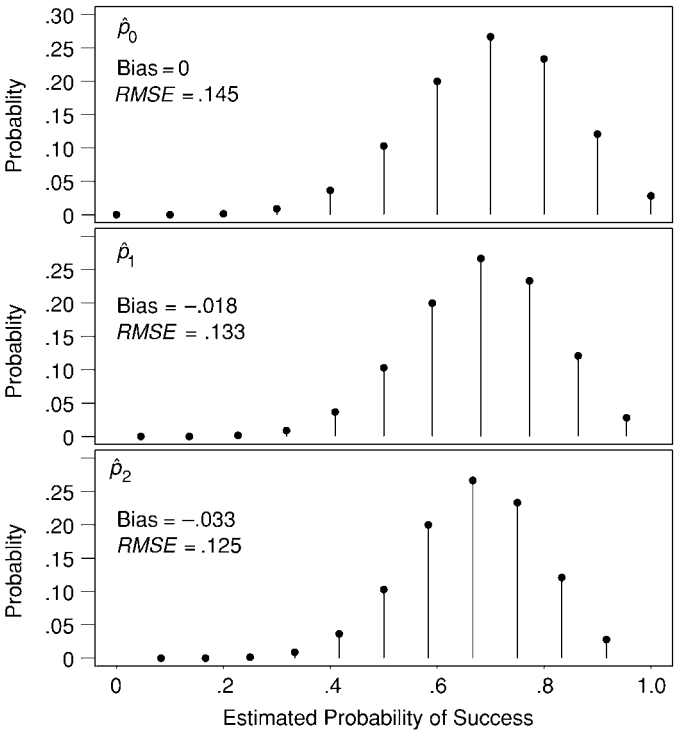

Figure 4. Sampling distributions of $\hat{p}_0$, $\hat{p}_1$, and $\hat{p}_2$ for $N = 10$ trials with a $p = .7$ true probability of success on any one trial. Bias and root mean squared error (*RMSE*) are included.

Bayes's (1763) insightful contribution was to use the law of conditional probability to estimate the parameter p. He noted that

$$f(p \mid y) = \frac{f(y \mid p) f(p)}{f(y)}. \tag{15}$$

We describe each of these terms in order.

The main goal is to find $f(p|y)$, the quantity on the left-hand side of Equation 15. The term $f(p|y)$ is the distribution of the parameter p given the data y. This distribution is referred to as the *posterior distribution*. It is viewed as a function of the parameter p. The mean of the posterior distribution, termed the *posterior mean*, serves as a suitable point estimator for p.

The term $f(y|p)$ plays two roles in statistics, depending on context. When it is viewed as a function of y for known p, it is known as the probability mass function. The *probability mass function* describes the probability of any outcome y given a particular value of p. For a binomial random variable, the probability mass function is given by

$$f(y \mid p) = \begin{cases} \binom{N}{y} p^y (1-p)^{N-y}, & y = 0, \ldots, N, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

For example, this function can be used to compute the probability of observing a total of two successes on three trials when $p = .5$:

$$f(y = 2 \mid p = .5) = \binom{3}{2} (.5)^2 (1 - .5)^1 = 3/8.$$

The term $f(y|p)$ can also be viewed as a function of parameter p for fixed y. In this case, it is termed the *likelihood function* and describes the likelihood of a particular parameter value p given a fixed value of y. For the binomial, the likelihood function is given by

$$f(y \mid p) = \binom{N}{y} p^y (1-p)^{N-y}, \quad 0 \le p \le 1 \tag{17}$$

In Bayes's theorem, we are interested in the posterior as a function of the parameter p. Hence, the right-hand terms of Equation 15 are viewed as a function of parameter p. Therefore, $f(y|p)$ is viewed as the likelihood of p rather than the probability mass of y.

The term $f(p)$ is the *prior distribution* and reflects the experimenter's a priori beliefs about the true value of p.

Finally, the term $f(y)$ is the distribution of the data given the model. Although its interpretation is important in some contexts, it plays a minimal role in the develop-

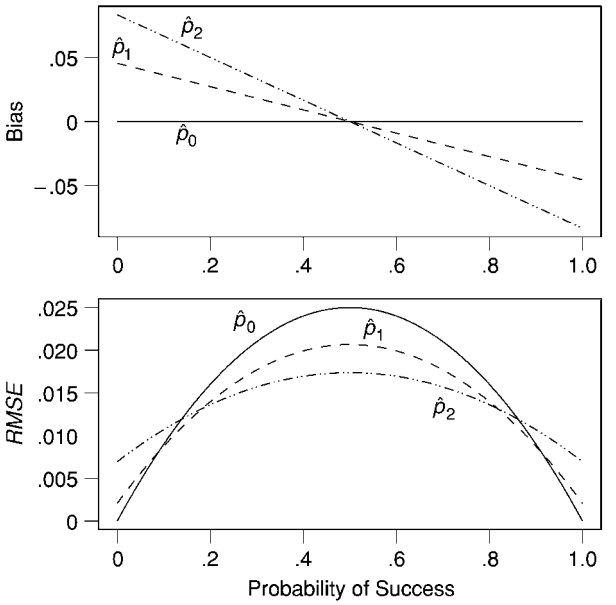

Figure 5. Bias and root mean squared error (*RMSE*) for the three estimators as functions of true probability of success. The solid, dashed, and dashed/dotted lines denote the characteristics of $\hat{p}_0$, $\hat{p}_1$, and $\hat{p}_2$, respectively.

ment presented here for the following reason: The posterior is a function of the parameter of interest, p, but $f(y)$ does not depend on p. The term $f(y)$ is a constant with respect to p and serves as a normalizing constant that ensures that the posterior density integrates to 1. As will be shown in the examples, the value of this normalizing constant usually becomes apparent in analysis.

To perform Bayesian analysis, the researcher must choose a prior distribution. For this application of estimating a probability from binomially distributed data, the *beta distribution* serves as a suitable prior distribution for p. The beta is a flexible two-parameter distribution; see Figure 6 for examples. In contrast to the normal distribution, which has nonzero density on all numbers, the beta has nonzero density between 0 and 1. The beta distribution is a function of two parameters, a and b, that determine its shape. The notation for specifying a beta distribution prior for p is

$$p \sim \text{Beta}(a, b).$$

The density of a beta random variable is given by

$$f(p) = \frac{p^{a-1}(1-p)^{b-1}}{\text{Be}(a,b)}. \tag{18}$$

The denominator, Be(a, b), is termed the *beta function*.[1] Like $f(y)$, it is not a function of p and plays the role of a normalizing constant in analysis.

In practice, researchers must completely specify the prior distribution before analysis; that is, they must choose suitable values for a and b. This choice reflects researchers' beliefs about the possible values of p before data are collected. When $a = 1$ and $b = 1$ (see the middle panel of Figure 6), the beta distribution is flat; that is, there is equal density for all values in the interval [0, 1]. By choosing $a = b = 1$, the researcher is committing to all values of p being equally likely before data are collected. Researchers can choose values of a and b to best match their beliefs about the expected data. For example, consider a psychophysical experiment in which the value of p will almost surely be greater than .5. For this experiment, priors with $a \geq b$ are appropriate.

The goal is to derive the posterior distribution using Bayes's theorem (Equation 15). Substituting the likelihood function (Equation 17) and prior (Equation 18) into Bayes's theorem yields

$$f(p \mid y) = \binom{N}{y} \frac{p^y(1-p)^{N-y}\, p^{a-1}(1-p)^{b-1}}{\text{Be}(a,b) \times f(y)}.$$

Collecting terms in p yields

$$f(p \mid y) = k p^{(y+a-1)}(1-p)^{(N-y+b-1)}, \tag{19}$$

where

$$k = [\text{Be}(a,b) \times f(y)]^{-1} \binom{N}{y}.$$

The term k is constant with respect to p and serves as a normalizing constant. Substituting

$$a' = y + a, \quad b' = N - y + b \tag{20}$$

into Equation 19 yields

$$f(p \mid y) = k p^{a'-1}(1-p)^{b'-1}.$$

This equation is proportional to a beta density for parameters a' and b'. To make $f(p \mid y)$ a proper probability density, k has a value such that $f(p \mid y)$ integrates to 1.

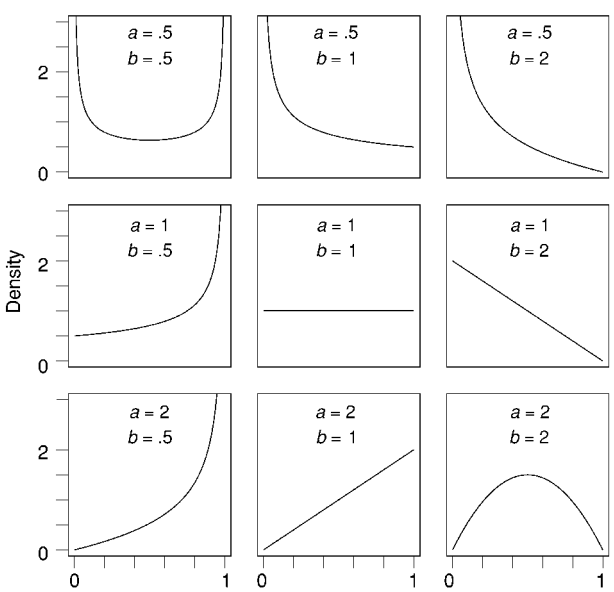

Figure 6. Probability density function of the beta distribution for various values of parameters *a* and *b*.

This occurs only if the value of k is a beta function—for example, $k = 1/\mathrm{Be}(a', b')$. Consequently,

$$f(p \mid y) = \frac{p^{a'-1}(1-p)^{b'-1}}{\mathrm{Be}(a', b')}. \qquad (21)$$

The posterior, like the prior, is a beta, but with parameters a' and b' given by Equation 20.

The derivation above was greatly facilitated by separating those terms that depend on the parameter of interest from those that do not. The latter terms serve as a normalizing constant and can be computed by ensuring that the posterior integrates to 1.0. In most derivations, the value of the normalizing constant becomes apparent, much as it did above. Consequently, it is convenient to consider only those terms that depend on the parameter of interest and to lump the rest into a proportionality constant. Hence, a convenient form of Bayes's theorem is

Posterior distribution $\propto$
Likelihood function $\times$ Prior distribution.

The symbol "$\propto$" denotes proportionality and is read *is proportional to*. This form of Bayes's theorem is used throughout this article.

Figure 7 shows both the prior and posterior of p for the case of $y = 7$ successes in $N = 10$ trials. The prior corresponds to a beta with $a = b = 1$, and the observed data are 7 successes in 10 trials. The posterior in this case is a beta distribution with parameters $a' = 8$ and $b' = 4$. As can be seen, the posterior is considerably more narrow than the prior, indicating that a large degree of information has been gained from the data. There are two valuable quantities derived from the posterior distribution: the posterior mean and the 95% highest density region (also termed the *95% credible interval*). The former is the point estimate of p, and the latter serves analogously to a confidence interval. These two quantities are also shown in Figure 7.

For the case of a beta-distributed posterior, the expression for the posterior mean is

$$\bar{p} = \frac{a'}{a' + b'} = \frac{y + a}{N + a + b}.$$

The posterior mean reduces to $\hat{p}_1$ if $a = b = .5$; it reduces to $\hat{p}_2$ if $a = b = 1$. Therefore, estimators $\hat{p}_1$ and $\hat{p}_2$ are theoretically justified from a Bayesian perspective. The estimator $\hat{p}_0$ may be obtained for values of $a = b = 0$. For these values, however, the integral of the prior distribution is infinite. If the integral of a distribution is infinite, the distribution is termed *improper*. Conversely, if the integral is finite, the distribution is termed *proper*. In Bayesian analysis, it is essential that posterior distributions be proper. However, it is not necessary for the prior to be proper. In some cases, an improper prior will still yield a proper posterior. When this occurs, the analysis is perfectly valid. For estimation of p, the posterior is proper even when $a = b = 0$.

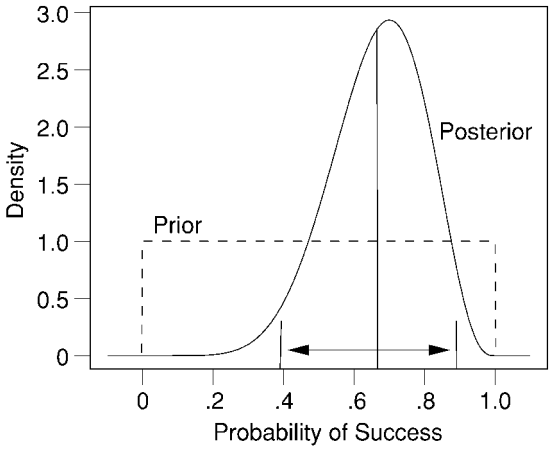

Figure 7. Prior and posterior distributions for 7 successes out of 10 trials. Both are distributed as betas. Parameters for the prior are $a = 1$ and $b = 1$; parameters for the posterior are $a' = 8$ and $b' = 4$. The posterior mean and 95% highest density region are also indicated.

When the prior and posterior are from the same distribution, the prior is termed *conjugate*. The beta distribution is the conjugate prior for binomially distributed data because the resulting posterior is also a beta distribution. Conjugate priors are convenient; they often facilitate the derivation of posterior distributions. Conjugacy, however, is not at all necessary for Bayesian analysis. A discussion about conjugacy is provided in the General Discussion.

BAYESIAN ESTIMATION WITH NORMALLY DISTRIBUTED DATA

In this section, we present Bayesian analysis for normally distributed data. The normal plays a critical role in the hierarchical signal detection model. We will assume that participant and item effects on sensitivity and bias are distributed as normals. The results and techniques developed in this section are used directly in analyzing the subsequent hierarchical signal detection models.

Consider the case in which a sequence of observations, $\mathbf{w} = (w_1, \ldots, w_N)$, are independent and identically distributed as normals:

$$w_j \overset{\text{ind}}{\sim} \text{Normal}\left(\mu, \sigma^2\right).$$

The goal is to estimate parameters μ and σ^2. In this section, we discuss a more limited goal—the estimation of one of the parameters assuming that the other is known. In the next section, we will introduce a form of Markov chain Monte Carlo sampling for estimation when both parameters are unknown.

Estimating μ With Known σ^2

Assume that σ^2 is known and that the estimation of μ is of primary interest. The goal then is to derive the pos-

terior, $f(\mu|\sigma^2; \mathbf{w})$. According to the proportional form of Bayes's theorem,

$$f\left(\mu \mid \sigma^2; \mathbf{w}\right) \propto f\left(\mathbf{w} \mid \mu, \sigma^2\right) f\left(\mu \mid \sigma^2\right).$$

All terms are conditioned on σ^2 to reflect the fact that it is known.

The first step is to choose a prior distribution for μ. The normal is a suitable choice:

$$\mu \sim \text{Normal}\left(\mu_0, \sigma_0^2\right).$$

Parameters μ_0 and σ_0^2 must be specified before analysis according to the researcher's beliefs about μ. By choosing σ_0^2 to be increasingly large, the researcher can make the prior arbitrarily variable. In the limit, the prior approaches having equal density across all values. This prior is considered *noninformative* for μ.

The density of the prior is given by

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(\frac{-\left(\mu - \mu_0\right)^2}{2\sigma_0^2}\right).$$

Note that the prior on μ does not depend on σ^2; therefore, $f(\mu) = f(\mu|\sigma^2)$.

The next step is multiplying the likelihood, $f(\mathbf{w}|\mu, \sigma^2)$, by the prior density, $f(\mu|\sigma^2)$. For normally distributed data, the likelihood is given by

$$f\left(\mathbf{w} \mid \mu, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(\sum_j \frac{-\left(w_j - \mu\right)^2}{2\sigma^2}\right). \quad (22)$$

Multiplying the equation above by the prior yields

$$f\left(\mu \mid \sigma^2; \mathbf{w}\right) \propto \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(\sum_j \frac{-\left(w_j - \mu\right)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(\frac{-\left(\mu - \mu_0\right)^2}{2\sigma_0^2}\right).$$

We concern ourselves with only those terms that involve μ, the parameter of interest. Other terms are used to normalize the posterior density and are absorbed into the constant of proportionality.

$$f\left(\mu \mid \sigma^2; \mathbf{w}\right) \propto \exp\left(\sum_j \frac{-\left(w_j - \mu\right)^2}{2\sigma^2}\right) \times \exp\left(\frac{-\left(\mu - \mu_0\right)^2}{2\sigma_0^2}\right),$$

$$f\left(\mu \mid \sigma^2; \mathbf{w}\right) \propto \exp\left(-\frac{1}{2}\left[\frac{\sum_j\left(w_j^2 - 2\mu w_j + \mu^2\right)}{\sigma^2} + \frac{\mu^2 - 2\mu_0\mu + \mu_0^2}{\sigma_0^2}\right]\right)$$

$$f(\mu \mid \sigma^2; \mathbf{w}) \propto \exp\left(-\frac{1}{2}\left[\left\{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right\}\mu^2 - 2\left\{\frac{\sum_j w_j}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right\}\mu + \left\{\frac{\sum_j w_j^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\right\}\right]\right),$$

$$f\left(\mu \mid \sigma^2; \mathbf{w}\right) \propto \exp\left(\frac{-\left(\mu - \mu'\right)^2}{2\sigma^{2'}}\right),$$

where

$$\sigma^{2'} = \left(N\sigma^{-2} + \sigma_0^{-2}\right)^{-1} \quad (23)$$

and

$$\mu' = \left(\sigma^{-2}\sum_j w_j + \sigma_0^{-2}\mu_0\right)\sigma^{2'}. \quad (24)$$

The posterior is proportional to the density of a normal with mean and variance given by μ' and $\sigma^{2'}$, respectively. Because the conditional posterior distribution must integrate to 1.0, this distribution must be that of a normal with mean μ' and variance $\sigma^{2'}$:

$$\mu \mid \sigma^2; \mathbf{w} \sim \text{Normal}\left(\mu', \sigma^{2'}\right). \quad (25)$$

Unfortunately, the notation does not make it clear that both μ' and $\sigma^{2'}$ depend on the value of σ^2. When this dependence is critical, it will be made explicit: $\mu'(\sigma^2)$ and $\sigma^{2'}(\sigma^2)$.

Because the posterior and prior are from the same family, the normal distribution is the conjugate prior for the population mean with normally distributed data (with known variance). In the limit that the prior is diffuse (i.e., as its variance is made increasingly large), the posterior of μ is a normal with $\mu' = \bar{\mathbf{w}}$ and $\sigma^{2'} = \sigma^2/N$. This distribution is also the sampling distribution of the sample mean in classical statistics. Hence, for the diffuse prior, the Bayesian and conventional approaches yield the same results.

Figure 8 shows the estimation for an example in which the data are $\mathbf{w} = (112, 106, 104, 111)$ and the variance is assumed known to be 16. The mean of these four observations is 108.25. For demonstration purposes, the prior

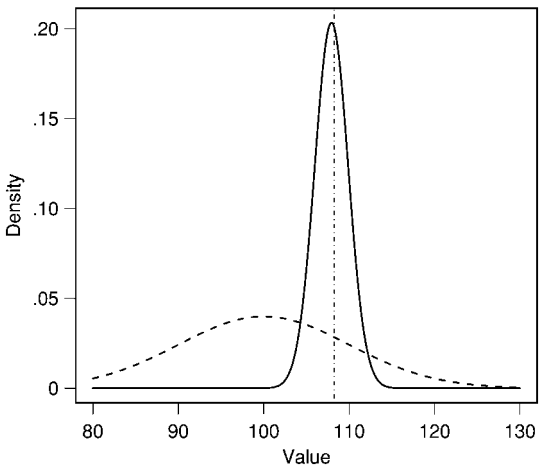

Figure 8. Prior and posterior distributions for estimating the mean of normally distributed data with known variance. The prior and posterior are the dashed and solid distributions, respectively. The vertical line shows the sample mean. The prior "pulls" the posterior slightly downward in this case.

is constructed to be fairly informative, with $\mu_0 = 100$ and $\sigma_0^2 = 100$. The posterior is narrower than the prior, reflecting the information gained from the data. It is centered at $\mu' = 107.9$, with variance 3.85. Note that the posterior mean is not the sample mean but is modestly influenced by the prior. Whether the Bayesian or the conventional estimate is better depends on the accuracy of the prior. For cases in which much is known about the dependent measure, it may be advantageous to reflect this information in the prior.

Estimating σ^2 With Known μ

In the last section, the posterior for μ was derived for known σ^2. In this section, it is assumed that μ is known and that the estimation of σ^2 is of primary interest. The goal then is to estimate the posterior $f(\sigma^2 | \mu; \mathbf{w})$. The *inverse-gamma* distribution is the conjugate prior for σ^2 with normally distributed data. The inverse-gamma distribution is not widely used in psychology. As its name implies, it is related to the gamma distribution. If a random variable g is distributed as a gamma, then $1/g$ is distributed as an inverse gamma. An inverse-gamma prior on σ^2 has density

$$f\left(\sigma^2\right) = \frac{b^a}{\Gamma(a)\left(\sigma^2\right)^{a+1}} \exp\left(\frac{-b}{\sigma^2}\right), \; a, b, \sigma^2 \geq 0.$$

The parameters of the inverse gamma are a and b, and in application, these are chosen beforehand. As a and b approach 0, the inverse gamma approaches $1/\sigma^2$, which is considered the appropriate noninformative prior for σ^2 (Jeffreys, 1961).

Multiplying the inverse-gamma prior by the likelihood given in Equation 22 yields

$$f\left(\sigma^2 \mid \mu; \mathbf{w}\right) \propto \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(\sum_j \frac{-\left(w_j - \mu\right)^2}{2\sigma^2}\right) \times \frac{b^a}{\Gamma(a)\left(\sigma^2\right)^{a+1}} \exp\left(\frac{-b}{\sigma^2}\right).$$

Collecting only those terms dependent on σ^2 yields

$$f\left(\sigma^2 \mid \mu; \mathbf{w}\right) \propto \frac{1}{\left(\sigma^2\right)^{N/2}} \exp\left(\sum_j \frac{-\left(w_j - \mu\right)^2}{2\sigma^2}\right) \times \frac{1}{\left(\sigma^2\right)^{a+1}} \exp\left(\frac{-b}{\sigma^2}\right),$$

$$f\left(\sigma^2 \mid \mu; \mathbf{w}\right) \propto \frac{1}{\left(\sigma^2\right)^{N/2+a+1}} \times \exp\left(-\frac{.5\sum_j \left(w_j - \mu\right)^2 + b}{\sigma^2}\right),$$

$$f\left(\sigma^2 \mid \mu; \mathbf{w}\right) \propto \frac{1}{\left(\sigma^2\right)^{a'+1}} \exp\left(\frac{-b'}{\sigma^2}\right),$$

where

$$a' = N/2 + a \tag{26}$$

and

$$b' = \left[\sum_j \left(w_j - \mu\right)^2\right] / 2 + b. \tag{27}$$

The posterior is proportional to the density of an inverse gamma with parameters a' and b'. Because this posterior integrates to 1.0, it must also be an inverse gamma:

$$\sigma^2 | \mu; \mathbf{w} \sim \text{Inverse Gamma } (a', b'). \tag{28}$$

Note that posterior parameter b' depends explicitly on the value of μ.

MARKOV CHAIN MONTE CARLO SAMPLING

The preceding development highlights a problem: The use of conjugate priors allowed for the derivation of posteriors only when some of the parameters were assumed to be known. The posteriors in Equations 25 and 28 are known as *conditional posteriors* because they depend on other parameters. The quantities of primary interest are the marginal posteriors: $\mu | \mathbf{w}$ and $\sigma^2 | \mathbf{w}$. The straightfor-

ward method of obtaining marginals is to integrate conditionals:

$$f(\mu \mid \mathbf{w}) = \int f\left(\mu \mid \sigma^2, \mathbf{w}\right) f\left(\sigma^2 \mid \mathbf{w}\right) d\sigma^2.$$

In this case, the integral is tractable, but in many cases it is not. When an integral is symbolically intractable, however, it can often still be evaluated numerically. One of the recent developments in numeric integration is *Markov chain Monte Carlo* (MCMC) sampling. MCMC sampling has become quite popular in the last decade and is described in detail in Gelfand and Smith (1990). We show here how MCMC can be used to estimate the marginal posterior distributions of μ and σ^2 without assuming known values. We use a particular type of MCMC sampling known as *Gibbs sampling* (Geman & Geman, 1984). Gibbs sampling is a restricted form of MCMC; the more general form is known as Metropolis–Hastings MCMC.

The goal is to find the marginal posterior distributions. MCMC provides a set of random samples from the marginal posterior distributions (rather than a closed-form derivation of the posterior density or cumulative distribution function). Obtaining a set of random samples from the marginal posteriors is sufficient for analysis. With a sufficiently large sample from a random variable, one can compute its density, mean, quantiles, and any other statistic to arbitrary precision.

To explain Gibbs sampling, we will digress momentarily using a different example. Consider the case of generating samples from an ex-Gaussian random variable. The ex-Gaussian is a well-known distribution in cognitive psychology (Hohle, 1965). It is the sum of normal and exponential random variables. The parameters are μ, σ, and τ, the location and scale of the normal and the scale of the exponential, respectively. An ex-Gaussian random variable, x, can be expressed in conditional form:

$$x \mid \eta \sim \text{Normal}\left(\mu + \eta, \sigma^2\right)$$

and

$$\eta \sim \text{Exponential}(\tau).$$

We can take advantage of this conditional form to generate ex-Gaussian samples. First, we generate a set of samples from η. Samples will be denoted with square brackets; $[\eta]_1$ denotes the first sample, $[\eta]_2$ denotes the second, and so on. These samples can then be used in the conditional form to generate samples from the marginal random variable x. For example, the first two samples are given by

$$[x]_1 \sim \text{Normal}\left(\mu + [\eta]_1, \sigma^2\right)$$

and

$$[x]_2 \sim \text{Normal}\left(\mu + [\eta]_2, \sigma^2\right).$$

In this manner, a whole set of samples from the marginal x can be generated from the conditional random variable $x \mid \eta$. Figure 9 shows that the method works. A histogram of [**x**] generated in this manner converges to the true ex-Gaussian density. This is an example of *Monte Carlo integration* of a conditional density.

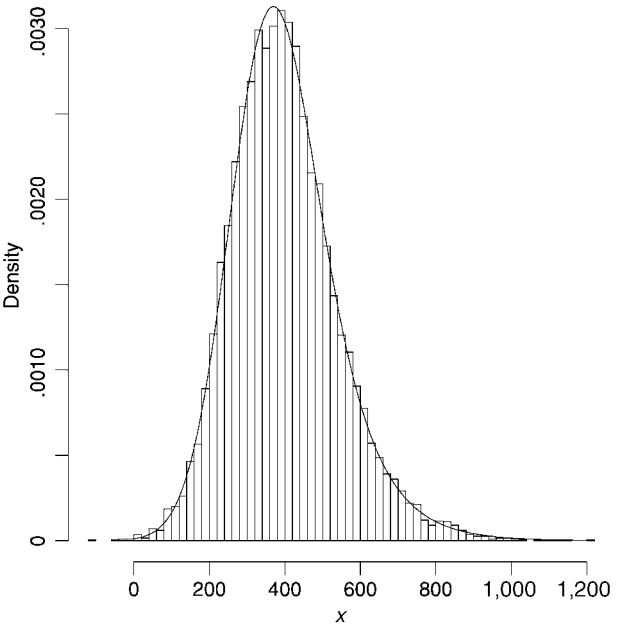

Figure 9. An example of Monte Carlo integration of a normal conditional density to yield ex-Gaussian distributed samples. The histogram presents the samples from a normal whose mean parameter is distributed as an exponential. The line is the density of the appropriate ex-Gaussian distribution.

We now return to the problem of deriving marginal posterior distributions of $\mu \mid \mathbf{w}$ from the conditional $\mu \mid \sigma^2; \mathbf{w}$. If there was a set of samples from $\sigma^2 \mid \mathbf{w}$, Monte Carlo integration could be used to sample $\mu \mid \mathbf{w}$. Likewise, if there was a set of samples of $\mu \mid \mathbf{w}$, Monte Carlo integration could be used to sample $\sigma^2 \mid \mathbf{w}$. In Gibbs sampling, these relations are used iteratively, as follows: First, a value of $[\sigma^2]_1$ is picked arbitrarily. This value is then used to sample the posterior of μ from the conditional in Equation 25:

$$[\mu]_1 \sim \text{Normal}\left(\mu'\left(\left[\sigma^2\right]_1\right), \sigma^{2'}\left(\left[\sigma^2\right]_1\right)\right),$$

where μ' and $\sigma^{2'}$ are the explicit functions of σ^2 given in Equations 23 and 24. Next, the value of $[\mu]_1$ is used to sample σ^2 from the conditional posterior in Equation 28. In general, for iteration m,

$$[\mu]_m \sim \text{Normal}\left(\mu'\left(\left[\sigma^2\right]_m\right), \sigma^{2'}\left(\left[\sigma^2\right]_m\right)\right) \qquad (29)$$

and

$$\left[\sigma^2\right]_m \sim \text{Inverse Gamma}\left(a', b'\left([\mu]_{m-1}\right)\right). \qquad (30)$$

In this manner, sequences of random samples $[\boldsymbol{\mu}]$ and $[\boldsymbol{\sigma^2}]$ are obtained.

The initial samples of $[\boldsymbol{\mu}]$ and $[\boldsymbol{\sigma^2}]$ certainly reflect the choice of starting values $[\sigma^2]_1$ and are not samples from the desired marginal posteriors. It can be shown, however, that under mild technical conditions, the later samples of $[\boldsymbol{\mu}]$ and $[\boldsymbol{\sigma^2}]$ are samples from the desired marginal posteriors (Tierney, 1994). The initial region in

which the samples reflect the starting value is termed the *burn-in period*. The later samples are the *steady-state period*. Formally speaking, the samples form irreducible Markov chains with stationary distributions that equal the marginal posterior distribution (see Gilks, Richardson, & Spiegelhalter, 1996). On a more informal level, the first set of samples have random and arbitrary values. At some point, by chance, the random sample happens to be from a high-density region of the true marginal posterior. From this point forward, the samples are from the true posteriors and may be used for estimation.

Let m_0 denote the point at which the samples are approximately steady state. Samples after m_0 can be used to provide both posterior means and credible regions of the posterior (as well as other statistics such as posterior variance). Posterior means are estimated by the arithmetic means of the samples:

$$\bar{\mu} = \frac{\sum_{m>m_0} [\mu]_m}{M - m_0}$$

and

$$\overline{\sigma^2} = \frac{\sum_{m>m_0} \left[\sigma^2\right]_m}{M - m_0},$$

where M is the number of iterations. Credible regions are constructed as the centermost interval containing 95% of the samples from m_0 to M.

To use MCMC to estimate posteriors of μ and σ^2, it is necessary to sample from a normal and an inverse-gamma distribution. Sampling from a normal is provided in many software packages, and routines for low-level languages such as C or Fortran are readily available. Samples from an inverse-gamma distribution may be obtained by taking the reciprocal of samples from a gamma distribution. Ahrens and Dieter (1974, 1982) provide algorithms for sampling the gamma distribution.

To illustrate Gibbs sampling, consider the case of estimating IQ. The hypothetical data for this example are four replicates from a single participant, $\mathbf{w} = (112, 106, 104, 111)$. The classical estimation of μ is the sample mean (108.25), and confidence intervals are constructed by multiplying the standard error ($s_{\bar{w}} = 1.93$) by the appropriate critical t value with three degrees of freedom. From this multiplication, the 95% confidence interval for μ is [102.2, 114.4]. Bayesian estimation was done with Gibbs sampling. Diffuse priors were placed on μ ($\mu_0 = 0$, $\sigma_0^2 = 10^6$) and σ^2 ($a = 10^{-6}$, $b = 10^{-6}$). Two chains of $[\boldsymbol{\mu}]$ are shown in the left panel of Figure 10, each corresponding to a different initial value for $[\sigma^2]_1$. As can be seen, both chains converge to their steady state rather quickly. For this application, there is very little burn-in. The histogram of the samples of $[\boldsymbol{\mu}]$ after burn-in is also shown (right panel). The histogram conforms to a t distribution with three degrees of freedom. The posterior mean is 108.26, and the 95% credible interval is [102.2, 114.4]. For normally distributed data, the diffuse priors used in Bayesian analysis yield results that are numerically equivalent to their frequentist counterparts.

The diffuse priors used in the previous case represent the situation in which researchers have no previous knowl-

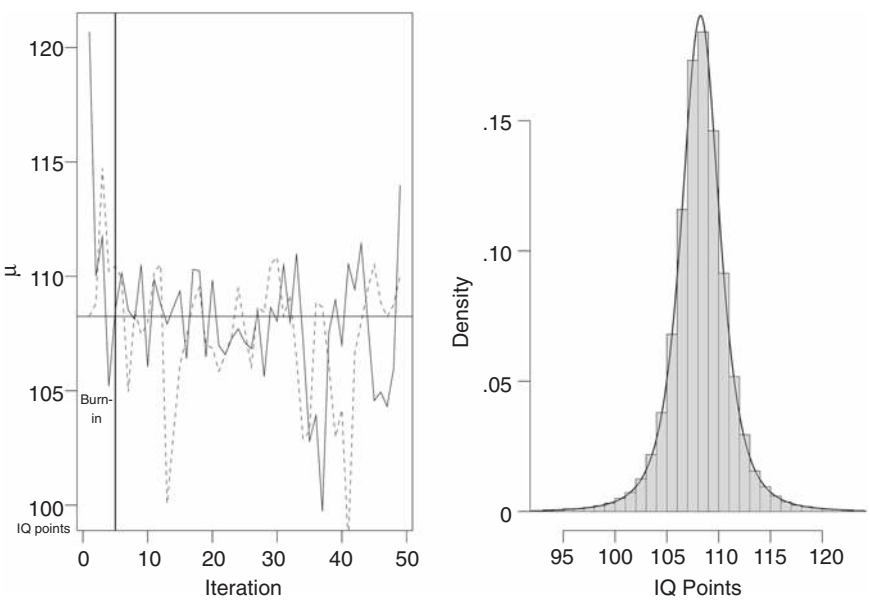

Figure 10. Samples of μ from Markov chain Monte Carlo integration for normally distributed data. The left panel shows two chains, each from a poor choice of the initial value of σ^2 (the solid and dotted lines are for $[\sigma^2]_1 = 10{,}000$ and $[\sigma^2]_1 = .0001$, respectively). Convergence is rapid. The right panel shows the histogram of samples of μ after burn-in. The density is the appropriate frequentist t distribution with three degrees of freedom.

edge. In the IQ example, however, much is known about the distribution of IQ scores. Suppose our test was normed for a mean of 100 and a standard deviation of 15. This knowledge could be profitably used by setting $\mu_0 = 100$ and $\sigma_0^2 = 15^2 = 225$. We may not know the particular test–retest correlation of our scales, but we can surmise that the standard deviations of replicates should be less than the population standard deviation. After experimenting with the values of a and b, we chose $a = 3$ and $b = 100$. The prior on σ^2 corresponding to these values is shown in Figure 11. It has much of its density below 100, reflecting our belief that the test–retest variability is smaller than the variability in the population. Figure 11 also shows a histogram of the posterior of μ (the posterior was obtained with Gibbs sampling as outlined above; $[\sigma^2]_1 = 1$, burn-in of 100 iterations). The posterior mean is 107.95, which is somewhat below the sample mean of 108.25. This difference is from the prior; people, on average, score lower than the obtained scores. Although it is a bit hard to see in the figure, the posterior distribution is more narrow in the extreme tails than is the corresponding t distribution. The 95% credible interval is [102.1, 113.6], which is modestly different from the corresponding frequentist confidence interval. This analysis shows that the use of reasonable prior information can lead to a result different from that of the frequentist analysis. The Bayesian estimate is not only different but often more efficient than the frequentist counterpart.

The theoretical underpinnings of MCMC guarantee that infinitely long chains will converge to the true posterior. Researchers must decide how long to burn in chains and then how long to continue in order to approximate the posterior well. The most important consideration in this decision is the degree of correlation from one sample to another. In MCMC sampling, consecutive samples are not necessarily independent of one another. The value of a sample on cycle $m + 1$ is, in general, dependent on that of m. If this dependence is small, then convergence happens quickly, and relatively short chains are needed. If this dependence is large, then convergence is much slower. One informal, graphical method of assessing the degree of dependence is to plot the autocorrelation functions of the chains. The bottom right panel of Figure 11 shows that for the normal application, there appears to be no autocorrelation; that is, there is no dependence from one sample to the next. This fact implies that convergence is relatively rapid: Good approximations may be obtained with short burn-ins and relatively short chains.

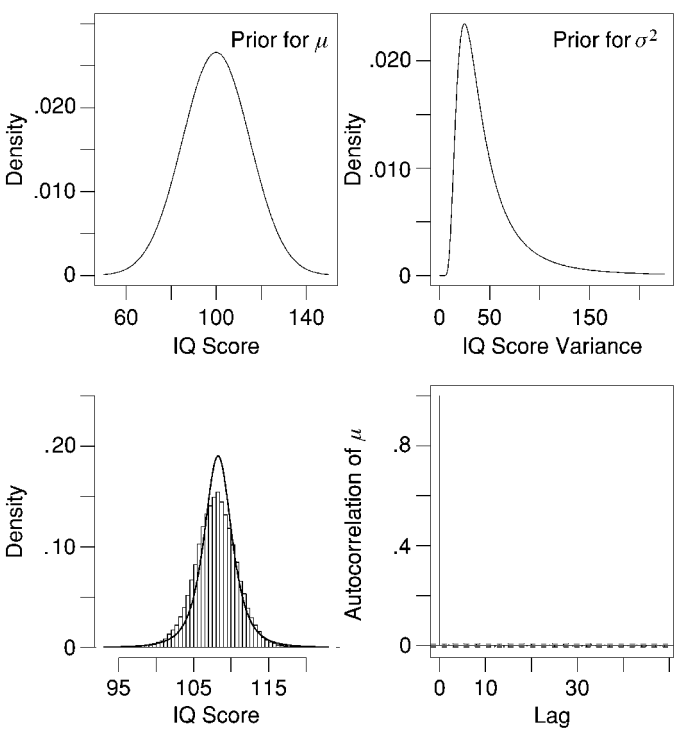

Figure 11. Analysis with informative priors. The two top panels show the prior distributions on μ and σ^2. The bottom left panel shows the histogram of samples of μ after burn-in. It differs from a t distribution from frequentist analysis in that it is shifted toward 100 and has a smaller upper tail. The bottom right panel shows the autocorrelation function for μ. There is no discernible autocorrelation.

There are a number of formal procedures for assessing convergence in the statistical literature (see, e.g., Gelman & Rubin, 1992; Geweke, 1992; Raftery & Lewis, 1992; see also Gelman et al., 2004).

A PROBIT MODEL FOR ESTIMATING A PROBABILITY OF SUCCESS

The next step toward a hierarchical signal detection model is to revisit estimation of a probability from individual trials. Once again, y successes are observed in N trials, and $y \sim \text{Binomial}(N, p)$, where p is the parameter of interest. In the previous section, a beta prior was placed on parameter p and the conjugacy of the beta with the binomial directly led to the posterior. Unfortunately, a beta distribution prior is not convenient as a prior for the signal detection application.

The signal detection model relies on the probit transform. Figure 12 shows the probit transform, which maps probabilities into z values using the normal inverse cumulative distribution function. Whereas p ranges from 0 to 1, z ranges across the real number line. Let Φ denote the standard normal cumulative distribution function. With this notation, $p = \Phi(z)$ and, conversely, $z = \Phi^{-1}(p)$.

Signal detection parameters are defined in terms of probit transforms:

$$d' = \Phi^{-1}\left(p^{(\mathrm{h})}\right) - \Phi^{-1}\left(p^{(\mathrm{f})}\right)$$

and

$$c = -\Phi^{-1}\left(p^{(\mathrm{f})}\right),$$

where $p^{(\mathrm{h})}$ and $p^{(\mathrm{f})}$ are hit and false alarm probabilities, respectively. As a precursor to models of signal detection, we develop estimation of the probit transform of a probability of success, $z = \Phi^{-1}(p)$. The methods for doing so will then be used in the signal detection application.

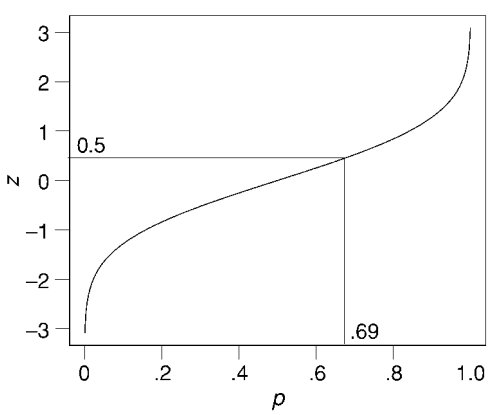

Figure 12. Probit transform of probability (p) into z scores (z). The transform is the inverse cumulative distribution function of the normal distribution. Lines show the transform of $p = .69$ into $z = 0.5$.

When using a probit transform, we model the y successes in N trials as

$$y \sim \text{Binomial}[N, \Phi(z)],$$

where z is the parameter of interest.

When using a probit, it is convenient to place a normal prior on z:

$$z \sim \text{Normal}\left(\mu_0, \sigma_0^2\right). \tag{31}$$

Figure 13 shows various normal priors on z and, by transform, the corresponding prior on p. Consider the special case when a standard normal is placed on z. As is shown in the figure, the corresponding prior on p is flat. A flat prior also corresponds to a beta with $a = b = 1$ (see Figure 6). From the previous development, if a beta prior is placed on p, the posterior is also a beta with parameters $a' = a + y$ and $b' = b + (N - y)$. Noting for a flat prior that $a = b = 1$, it is expected that the posterior of p is distributed as a beta with parameters $a' = 1 + y$ and $b' = 1 + (N - y)$.

The next step is to derive the posterior, $f(z \mid Y)$, for a normal prior on z. The straightforward approach is to note that the posterior is proportional to the likelihood multiplied by the prior:

$$f(z \mid Y) \propto \Phi(z)^Y [1 - \Phi(z)]^{N-Y} \times \exp\left[-\left(z - \mu_0\right)^2 \left(2\sigma_0^2\right)^{-1}\right].$$

This equation is intractable. Albert and Chib (1995) provide an alternative for analysis, and the details of this alternative are provided in Appendix A. The basic idea is that a new set of latent variables is defined upon which z is conditioned. Conditional posteriors for both z and these new latent variables are easily derived and sampled. These samples can then be used in Gibbs sampling to find a marginal posterior for z.

In this application, it may be desirable to estimate the posterior of p as well as that of z. This estimation is easy in MCMC. Because p is the inverse probit transform of z, samples can also be inversely transformed. Let $[\mathbf{z}]$ be the chain of samples from the posterior of z. A chain of posterior samples of p is given by $[p]_m = \Phi([z]_m)$.

The top right panel of Figure 14 shows the histogram of the posterior of p for $y = 7$ successes on $N = 10$ trials. The prior parameters were $\mu_0 = 0$ and $\sigma_0^2 = 1$, so the prior on p was flat (see Figure 13). Because the prior is flat, it corresponds to a beta with parameters ($a = 1, b = 1$). Consequently, the expected posterior is a beta with density $a' = y + 1 = 8$ and $b' = (N - y) + 1 = 4$. The chains in the Gibbs sampler were started from a number of values of $[z]_1$, and convergence was always rapid. The resulting histogram for 10,000 samples is plotted; it approximates the expected distribution. The bottom panel shows a small degree of autocorrelation. This autocorrelation can be overcome by running longer chains and using a more conservative burn-in period. Chains of

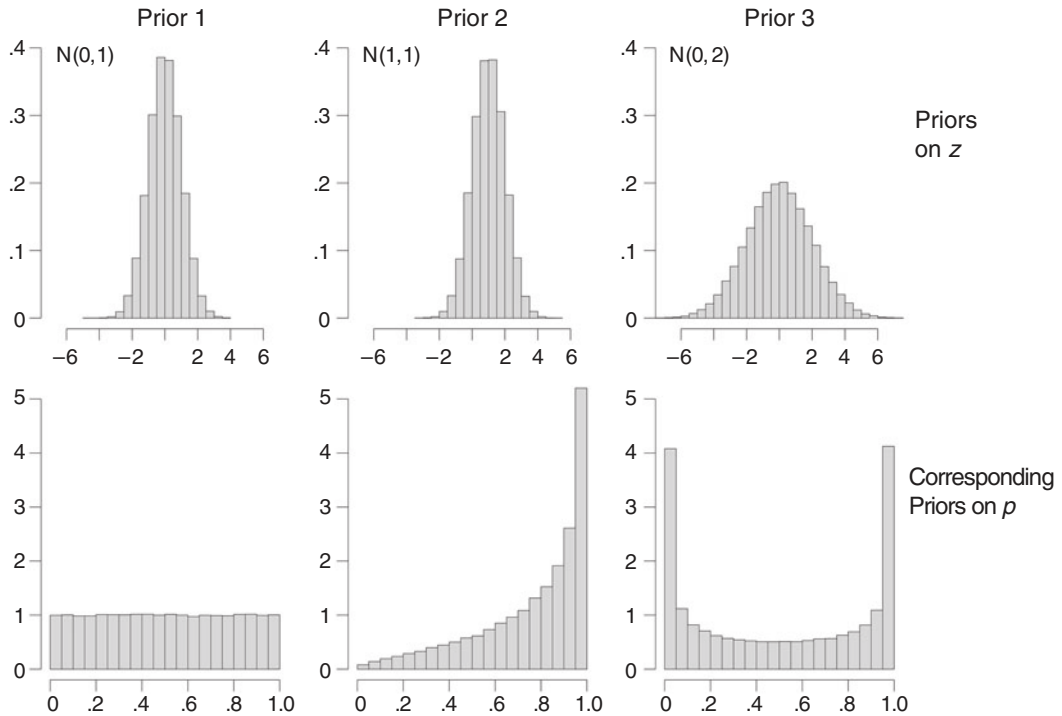

Figure 13. Three normally distributed priors on *z* and the corresponding priors on *p*. Priors on *z* were obtained by drawing 100,000 samples from the normal. Priors on *p* were obtained by transforming each sample of *z*.

1,000 samples with 100 samples of burn-in are more than sufficient for this application.

ESTIMATING PROBABILITY FOR SEVERAL PARTICIPANTS

In this section, we propose a model for estimating several participants' probability of success. The main goal is to introduce hierarchical models within the context of a relatively simple example. The methods and insights discussed here are directly applicable to the hierarchical model of signal detection. Consider data from a number of participants, each with his or her own unique probability of success, p_i. In conventional estimation, each of these probabilities is estimated separately, usually by the estimator $\hat{p}_i = y_i/\mathrm{N}_i$, where y_i and N_i are the number of successes and trials, respectively, for the *i*th participant. We show here that hierarchical modeling leads to more accurate estimation.

In a hierarchical model, it is assumed that although individuals vary in ability, the distribution governing individuals' true abilities is orderly. We term this distribution the *parent distribution*. If we knew the parent distribution, this prior information would be useful in estimation. Consider the example in Figure 15. Suppose 10 individuals each performed 10 trials. Hypothetical performance is indicated with *X*s. Let's suppose the task is relatively easy, so that individuals' true probabilities of success range from .7 to 1.0, as indicated by the curve in the figure. One participant does poorly, however, succeeding on only 3 trials. The conventional estimate for this poor-performing participant is .3, which is a poor estimate of the true probability (between .7 and 1.0). If the parent distribution is known, it would be evident that this poor score is influenced by a large degree of sample noise and should be corrected upward, as indicated by the arrow. The new estimate, which is more accurate, is denoted by an *H* (for *hierarchical*).

In the example above, we assumed a parent distribution. In actuality, such assumptions are unwarranted. In hierarchical models, parent distributions are estimated from the data; these parents, in turn, affect the estimates from outlying data. The estimated parent for the data in Figure 15 would have much of its mass in the upper ranges. When serving as a prior, the estimated parent would exert an upward tendency in the estimation of the outlying participant's probability, leading to better estimation. In Bayesian analysis, the method of implementing a hierarchical model is to use a *hierarchical prior*. The parent distribution serves as a prior for each individual, and this parent is termed the *first stage* of the prior. The parent distribution is not fully specified, but instead, free parameters of the parent are estimated from the data. These parameters also need a prior on them, which is termed the *second stage* of the prior.

The probit transform model is ideally suited for a hierarchical prior. A normally distributed parent may be placed on transformed probabilities of success. The *i*th

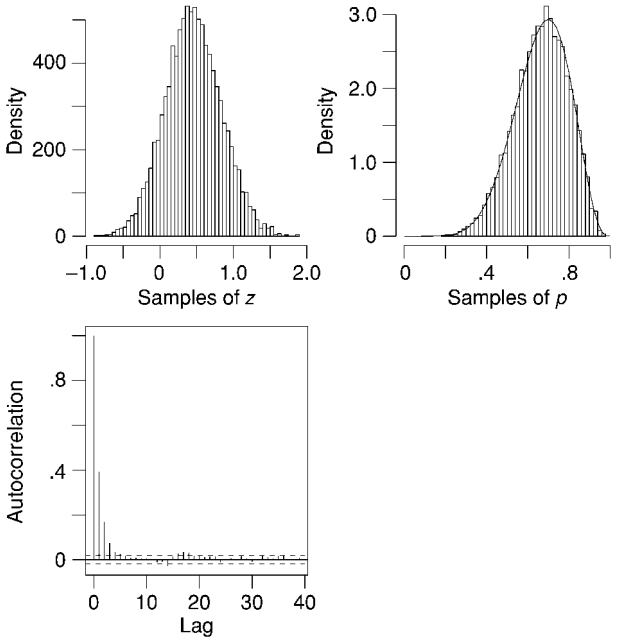

Figure 14. (Top) Histograms of the post-burn-in samples of z and p, respectively, for 7 successes out of 10 trials. The line fit for p is the true posterior, a beta distribution with $a' = 8$ and $b' = 4$. (Bottom) Autocorrelation function for z. There is a small degree of autocorrelation.

participant's true ability, z_i, is simply a sample from this normal. The number of successes is modeled as

$$y_i \mid z_i \overset{\text{ind}}{\sim} \text{Binomial}\left[N_i, \Phi\left(z_i\right)\right], \tag{32}$$

with a parent distribution (first-stage prior) given by

$$z_i \mid \mu, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left(\mu, \sigma^2\right). \tag{33}$$

In this case, parameters μ and σ^2 describe the location and variance of the parent distribution and reflect group-level properties. Priors are needed on μ and σ^2 (second-stage priors). Suitable choices are

$$\mu \sim \text{Normal}\left(\mu_0, \sigma^2\right). \tag{34}$$

and

$$\sigma^2 \sim \text{Inverse Gamma } (a, b). \tag{35}$$

Values of μ_0, σ_0^2, a, and b must be specified before analysis.

We performed an analysis on hypothetical data in order to demonstrate the hierarchical model. Table 2 shows hypothetical data for a set of 20 participants performing 50 trials. These success frequencies were generated from a binomial, but each individual had a unique true probability of success. These true probabilities,[2] which ranged between .59 and .80, are shown in the second column of Table 2. Parameter estimators $\hat{p}_0$ (Equation 12) and $\hat{p}_2$ (Equation 14) are shown.

All of the theoretical results needed to derive the conditional posterior distributions have already been presented. The Albert and Chib (1995) analysis for this application is provided in Appendix B. A coded version of MCMC for this model is presented in Appendix C.

The MCMC chain was run for 10,000 iterations. Priors were set to be fairly diffuse ($\mu_0 = 0$, $\sigma_0^2 = 1{,}000$, $a = b = .1$). All parameters converged to steady-state behavior quickly (burn-in was conservative, at 100 iterations). Autocorrelation of samples of p_i for 2 select participants is shown in the bottom panels of Figure 16. The autocorrelation in these plots was typical of other parameters. Autocorrelation for this model was no worse than that for the nonhierarchical Bayesian model of the preceding section. The top left panel of Figure 16 shows posterior means of the probability of success as a function of the true probability of success for the hierarchical model (filled circles). Also included are classical estimates from $\hat{p}_0$ (filled circles). Points from the hierarchical model are, on average, closer to the diagonal. As is indicated in Table 2, the hierarchical model has lower root mean squared error than does its frequentist counterpart, $\hat{p}_0$. The top right panel shows the hierarchical estimates as a function of $\hat{p}_0$. The effect of the hierarchical structure is clear: The extreme estimates in the frequentist approach are moderated in the Bayesian analysis.

Estimator $\hat{p}_2$ also moderates extreme values. The tendency is to pull them toward the middle; specifically, estimates are closer to .5 than they are with $\hat{p}_0$. For the data of Table 2, the difference in efficiency between $\hat{p}_2$ and $\hat{p}_0$ is exceedingly small (about .0001). The reason that there was little gain for $\hat{p}_2$ is that it pulled estimates toward a value of .5, which is not the mean of the true probabilities. The mean was $\bar{p} = .69$, and for this value and these sample sizes, estimators $\hat{p}_2$ and $\hat{p}_0$ have about the same efficiency. The hierarchical estimator is supe-

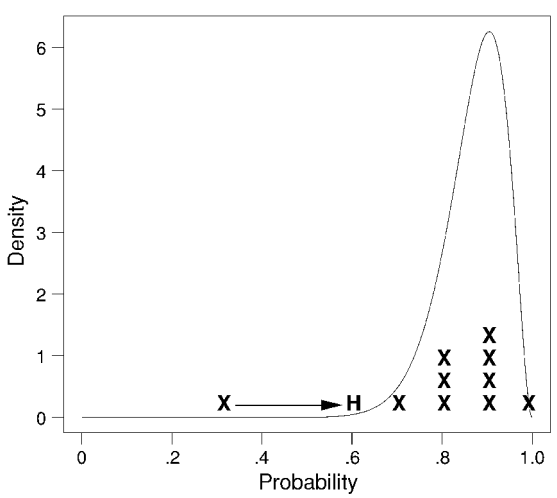

Figure 15. Example of how a parent distribution affects estimation. The outlying observation is incompatible with the parent distribution. Bayesian estimation allows for the influence of the parent as a prior. The resulting estimate is shifted upward, reducing error.

Table 2
Data and Estimates of a Probability of Success

Participant	True Prob.	Data*	$\hat{p}_0$	$\hat{p}_2$	Hierarchical
1	.587	33	.66	.654	.672
2	.621	28	.56	.558	.615
3	.633	34	.68	.673	.683
4	.650	29	.58	.577	.627
5	.659	32	.64	.635	.660
6	.665	37	.74	.731	.716
7	.667	30	.60	.596	.638
8	.667	29	.58	.577	.627
9	.687	34	.68	.673	.684
10	.693	34	.68	.673	.682
11	.696	34	.68	.673	.683
12	.701	37	.74	.731	.715
13	.713	38	.76	.750	.728
14	.734	38	.76	.750	.727
15	.736	37	.74	.731	.717
16	.738	33	.66	.654	.672
17	.743	37	.74	.731	.716
18	.759	33	.66	.654	.672
19	.786	38	.76	.750	.728
20	.803	42	.84	.827	.771
RMSE			.053	.053	.041

*Data are numbers of successes in 50 trials.

rior to these nonhierarchical estimators because it pulls outlying estimates toward the mean value, $\overline{p}$. Of course, the advantage of the hierarchical model is a function of sample size. With increasing sample sizes, the amount of pulling decreases. In the limit, all three estimators converge to the true individual's probability.

A HIERARCHICAL SIGNAL DETECTION MODEL

Our ultimate goal is a signal detection model that accounts for both participant and item variability. As a precursor, we present a hierarchical signal detection model without item variability; the model will be expanded in the next section to include item variability. The model here is applicable to psychophysical experiments in which signal and noise trials are identical replicates, respectively.

Signal detection parameters are simple subtractions of probit-transformed probabilities. Let d'_i and c_i denote signal detection parameters for the ith participant:

$$d'_i = \Phi^{-1}\left(p_i^{(\mathrm{h})}\right) - \Phi^{-1}\left(p_i^{(\mathrm{f})}\right)$$

and

$$c_i = -\Phi^{-1}\left(p_i^{(\mathrm{f})}\right),$$

where $p_i^{(\mathrm{h})}$ and $p_i^{(\mathrm{f})}$ are individuals' true hit and false alarm probabilities, respectively. Let h_i and f_i be the probit transforms: $h_i = \Phi^{-1}(p_i^{(\mathrm{h})})$ and $f_i = \Phi^{-1}(p_i^{(\mathrm{f})})$. Then, sensitivity and criteria are given by

$$d'_i = h_i - f_i$$

and

$$c_i = -f_i.$$

The tactic we follow is to place hierarchical models on h and f rather than on signal detection parameters directly. Because the relationship between signal detection parameters and probit-transformed probabilities is linear, accurate estimation of h_i and f_i results in accurate estimation of d'_i and c_i.

The model is developed as follows: Let N_i and S_i denote the number of noise and signal trials given to the ith participant. Let $y_i^{(\mathrm{h})}$ and $y_i^{(\mathrm{f})}$ denote the the number of hits and false alarms, respectively. The model for $y_i^{(\mathrm{h})}$ and $y_i^{(\mathrm{f})}$ is given by

$$y_i^{(\mathrm{h})} \mid h_i \overset{\text{ind}}{\sim} \text{Binomial}\left[S_i, \Phi\left(h_i\right)\right]$$

and

$$y_i^{(\mathrm{f})} \mid f_i \overset{\text{ind}}{\sim} \text{Binomial}\left[N_i, \Phi\left(f_i\right)\right].$$

It is assumed that each participant's parameters are drawn from parent distributions. These parent distributions serve as the first stage of the hierarchical prior and are given by

$$h_i \overset{\text{ind}}{\sim} \text{Normal}\left(\mu_{\mathrm{h}}, \sigma_{\mathrm{h}}^2\right) \tag{36}$$

and

$$f_i \overset{\text{ind}}{\sim} \text{Normal}\left(\mu_{\mathrm{f}}, \sigma_{\mathrm{f}}^2\right). \tag{37}$$

Second-stage priors are placed on parent distribution parameters (μ_k, σ_k^2) as

$$\mu_k \sim \text{Normal}\left(u_k, v_k\right),\ k = \mathrm{h}, \mathrm{f}, \tag{38}$$

and

$$\sigma_k^2 \sim \text{Inverse Gamma}\left(a_k, b_k\right),\ k = \mathrm{h}, \mathrm{f}. \tag{39}$$

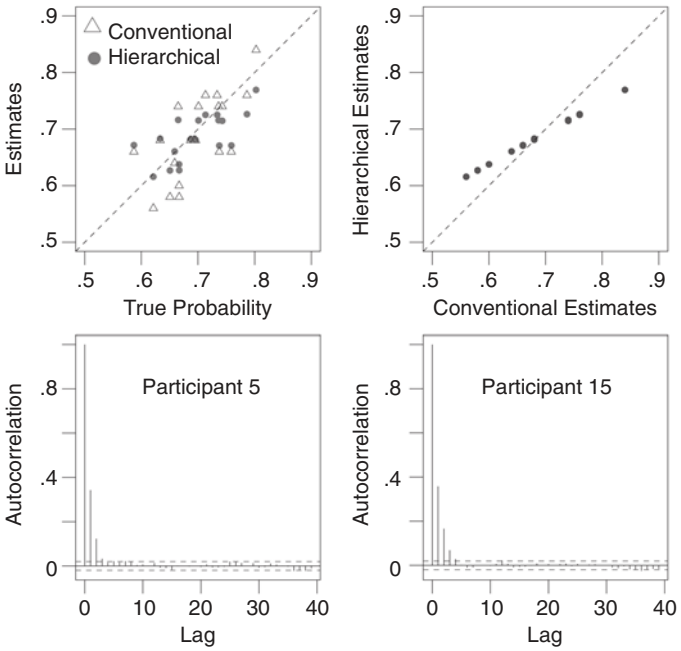

Figure 16. Modeling the probability of success across several participants. (Top left) Probability estimates as functions of true probability; the triangles represent estimator $\hat{p}_0$, and the circles represent posterior means of p_i from the hierarchical model. (Top right) Hierarchical estimates as a function of $\hat{p}_0$. (Bottom) Autocorrelation functions of the posterior probability estimates for 2 participants from the hierarchical model. The degree of autocorrelation is typical of the other participants.

Analysis takes place in two steps. The first is to obtain posterior samples from h_i and f_i for each participant. These are probit-transformed parameters that may be estimated following the exact procedure in the previous section. Separate runs are made for signal trials (producing estimates of h_i) and for noise trials (producing estimates of f_i). These posterior samples are denoted $[h_i]$ and $[f_i]$, respectively. The second step is to use these posterior samples to estimate posterior distributions of individual d_i' and c_i. This is easily accomplished. For all samples,

$$\left[d_i'\right]_m = \left[h_i\right]_m - \left[f\right]_m \tag{40}$$

and

$$\left[c_i\right]_m = -\left[f_i\right]_m . \tag{41}$$

Hypothetical data and parameter estimates are shown in Table 3. Individual variability in the data was implemented by sampling each participant's d' and c from a log-normal and from a normal distribution, respectively. True d' values are shown, and they range from 1.23 to 2.34. From these true values of d' and c, true individual hit and false alarm probabilities were computed, and individual hit and false alarm frequencies were sampled from a binomial. Hit and false alarm frequencies were produced in this manner for 20 hypothetical individuals, each observing 50 signal and 50 noise trials.

A conventional sensitivity estimate for each individual was obtained using the formula

$$\Phi^{-1}(\text{Hit rate}) - \Phi^{-1}(\text{False alarm rate}).$$

The resulting estimates are shown in the last column of Table 3, labeled *Conventional*. The hierarchical signal detection model was analyzed with diffuse prior parameters: $a_h = a_f = b_h = b_f = .01$, $u_h = u_f = 0$, and $v_h = v_f = 1{,}000$. The chain was run for 10,000 iterations, with the first 500 serving as burn-in. This choice of burn-in period is very conservative. Autocorrelation for sensitivity for 2 select participants is displayed in the bottom panels of Figure 17; this degree of autocorrelation was typical for all parameters. Autocorrelation was about the same as in the previous application, and the resulting estimates are shown in the next-to-last column of Table 3.

The top left panel of Figure 17 shows a plot of estimated d' as a function of true d'. Estimates from the hierarchical model are closer to the diagonal than are the conventional estimates, indicating more accurate estimation. Table 3 also provides efficiency information; the hierarchical model is 33% more efficient than the conventional estimates (this amount is typical of repeated runs of the same simulation). The reason for this gain is evident in the top right panel of Figure 17. The extreme estimates in the conventional method, which most likely

Table 3
Data and Estimates of Sensitivity

Participant	True d'	Hits*	False Alarms*	Hierarchical	Conventional
1	1.238	35	7	1.608	1.605
2	1.239	38	17	1.307	1.119
3	1.325	38	11	1.554	1.478
4	1.330	33	16	1.146	0.880
5	1.342	40	19	1.324	1.147
6	1.405	39	8	1.730	1.767
7	1.424	37	12	1.466	1.350
8	1.460	27	5	1.417	1.382
9	1.559	35	9	1.507	1.440
10	1.584	37	9	1.599	1.559
11	1.591	33	8	1.486	1.407
12	1.624	40	7	1.817	1.922
13	1.634	35	11	1.427	1.297
14	1.782	43	13	1.687	1.724
15	1.894	33	3	1.749	1.967
16	1.962	45	12	1.839	1.988
17	1.967	45	4	2.235	2.687
18	2.124	39	6	1.826	1.947
19	2.270	45	5	2.172	2.563
20	2.337	46	6	2.186	2.580
RMSE				.183	.275

*Data are numbers of signal responses in 50 trials.

reflect sample noise, are shrunk to more moderate estimates in the hierarchical model.

SIGNAL DETECTION WITH ITEM VARIABILITY

In the introduction, we stressed that Bayesian hierarchical modeling may provide a solution for the statistical problems associated with unmodeled variability in nonlinear contexts. In all of the previous examples, the gains from hierarchical modeling were in better efficiency. Conventional analysis, although not as accurate as Bayesian analysis, certainly yielded consistent if not unbiased estimates. In this section, item variability is added to signal detection. Conventional analyses, in which scores are aggregated over items, result in inconsistent estimates characterized by asymptotic bias. In this section, we show that in contrast to conventional analysis, Bayesian hierarchical models provide consistent and accurate estimation.

Our route to the model is a bit circuitous. Our first attempt, based on a straightforward extension of the previous techniques, fails because of an excessive amount of autocorrelation. The solution to this failure is to use a powerful and recently developed sampling scheme, *blocked sampling* (Roberts & Sahu, 1997). First, we will proceed with the straightforward but problematic extension. Then we will introduce blocked sampling and show how it leads to tractable analysis.

Straightforward Extension

It is straightforward to specify models that include item effects. The previous notation is here expanded by indexing hits and false alarms by both participants and items. Let $y_{ij}^{(\mathrm{h})}$ and $y_{ij}^{(\mathrm{f})}$ denote the response of the ith person to the jth item. Values of $y_{ij}^{(\mathrm{h})}$ and $y_{ij}^{(\mathrm{f})}$ are dichotomous, indicating whether the response was *old* or *new*. Let $p_{ij}^{(\mathrm{h})}$ and $p_{ij}^{(\mathrm{f})}$ be true probabilities that $y_{ij}^{(\mathrm{h})} = 1$ and $y_{ij}^{(\mathrm{f})} = 1$, respectively; that is, these probabilities are true hit and false alarm rates for a given participant-by-item combination, respectively. Let h_{ij} and f_{ij} be the probit transforms $[h_{ij} = \Phi^{-1}(p_{ij}^{(\mathrm{h})}), f_{ij} = \Phi^{-1}(p_{ij}^{(\mathrm{f})})]$. The model is given by

$$y_{ij}^{(\mathrm{h})} \overset{\text{ind}}{\sim} \text{Bernoulli}\left[\Phi\left(h_{ij}\right)\right] \tag{42}$$

and

$$y_{ij}^{(\mathrm{f})} \overset{\text{ind}}{\sim} \text{Bernoulli}\left[\Phi\left(f_{ij}\right)\right]. \tag{43}$$

Linear models are placed on h_{ij} and f_{ij}:

$$h_{ij} = \mu_{\mathrm{h}} + \alpha_i + \gamma_j \tag{44}$$

and

$$f_{ij} = \mu_{\mathrm{f}} + \beta_i + \delta_j. \tag{45}$$

Parameters μ_{h} and μ_{f} correspond to the grand means of the hit and false alarm rates, respectively. Parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ denote participant effects on hits and false alarms, respectively; parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ denote item effects on hits and false alarms, respectively. These participant and item effects are assumed to be zero-centered random effects. Grand mean signal detection parameters, denoted $d'_{\cdot\cdot}$ and $c_{\cdot\cdot}$, are given by

$$d'_{\cdot\cdot} = \mu_{\mathrm{h}} - \mu_{\mathrm{f}} \tag{46}$$

and

$$c_{\cdot\cdot} = -\mu_{\mathrm{f}}. \tag{47}$$

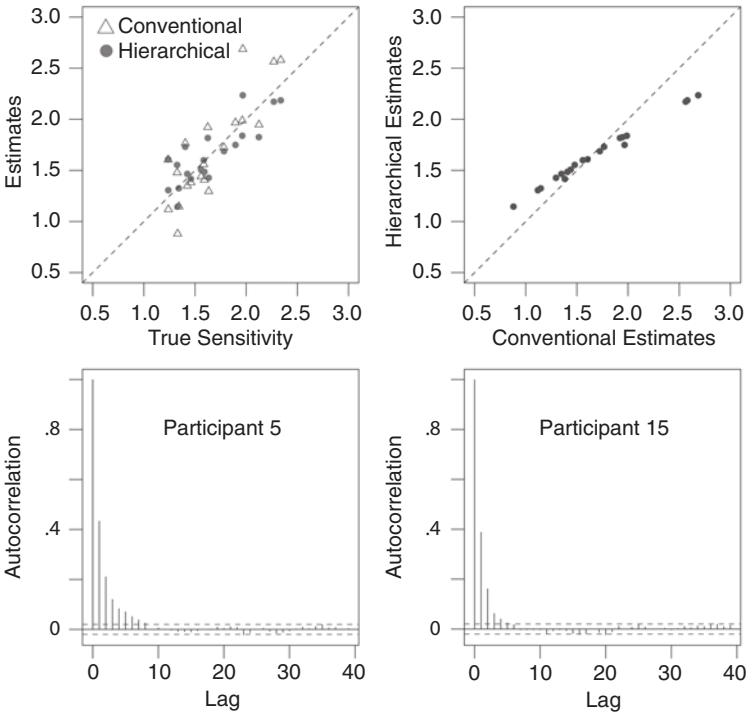

Figure 17. Modeling signal detection across several participants. (Top left) Parameter estimates from the conventional approach and from the hierarchical model as functions of true sensitivity. (Top right) Hierarchical estimates as a function of the conventional ones. (Bottom) Autocorrelation functions for posterior samples of d_i' for 2 participants.

Participant-specific signal detection parameters, denoted $d'_{i\cdot}$ and $c_{i\cdot}$, are given by

$$d'_{i\cdot} = \mu_{\mathrm{h}} - \mu_{\mathrm{f}} + \alpha_i - \beta_i \tag{48}$$

and

$$c_{i\cdot} = -\mu_{\mathrm{f}} - \beta_i. \tag{49}$$

Item-specific signal detection parameters, denoted $d'_{\cdot j}$ and $c_{\cdot j}$, are given by

$$d'_{\cdot j} = \mu_{\mathrm{h}} - \mu_{\mathrm{f}} + \gamma_j - \delta_j \tag{50}$$

and

$$c_{\cdot j} = -\mu_{\mathrm{f}} - \delta_j. \tag{51}$$

Priors on $(\mu_{\mathrm{h}}, \mu_{\mathrm{f}})$ are independent normals with mean $\mu_{0,k}$ and variance $\sigma^2_{0,k}$, where k indicates hits or false alarms. In the implementation, we set $\mu_{0,k} = 0$ and $\sigma^2_{0,k} = 500$, a sufficiently large number. The priors on $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta})$ are hierarchical. The first stages (parent distributions) are normals centered around 0:

$$\alpha_i \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma^2_\alpha\right), \tag{52}$$

$$\beta_i \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma^2_\beta\right), \tag{53}$$

$$\gamma_j \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma^2_\gamma\right), \tag{54}$$

and

$$\delta_j \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma^2_\delta\right). \tag{55}$$

Second-stage priors on all four variances are inverse-gamma distributions. In application, the parameters of the inverse gamma were set to ($a = .01$, $b = .01$). The model was evaluated with the same simulation from the introduction (Equations 8–11). Twenty hypothetical participants observed 50 signal and 50 noise trials. Data in the simulation were generated in accordance with the mirror effect principle: Item and participant increases in sensitivity corresponded to both increases in hit probabilities and decreases in false alarm probabilities. Although the data were generated with a mirror effect, this

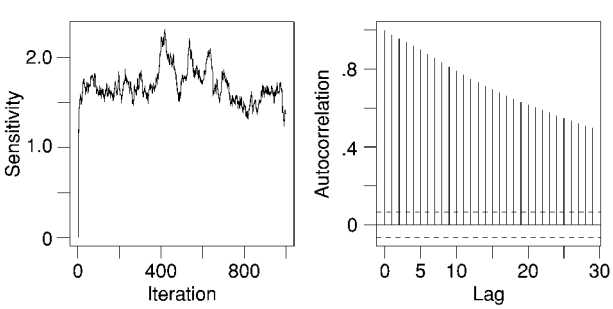

Figure 18. Autocorrelation of overall sensitivity $(\mu_{\mathrm{h}} - \mu_{\mathrm{f}})$ in a zero-centered random-effects model. (Left) Values of the chain. (Right) Autocorrelation function. This degree of autocorrelation is problematic.

effect is not assumed in model analysis: Participant and item effects on hits and false alarms are estimated independently.

Figure 18 shows plots of the samples of overall sensitivity ($d'_{\cdot\cdot}$). The chain has a much larger degree of autocorrelation than was previously observed, and this autocorrelation greatly complicates analysis. If a short chain is used, the resulting sample will not reflect the true posterior. Two strategies can mitigate autocorrelation. The first is to run long chains and *thin* them. For this application, the Gibbs sampling is sufficiently quick that the chain can be run for millions of cycles in the course of a day. With exceptionally long chains, convergence occurs. In such cases, researchers do not use all iterations, but instead skip a fixed number—for example, discarding all but every 50th observation. The correlation between sample m and sample $m + 50$ is greatly attenuated. The choice of the multiple, 50 in the example above, is not completely arbitrary, because it should reflect the amount of autocorrelation. Chains with larger amounts of autocorrelation should implicate larger multiples for thinning.

Although the brute-force method of running long chains and thinning is practical in this context, it may be impractical in others. For instance, we have encountered a greater degree of autocorrelation in Weibull hierarchical models (Lu, 2004; Rouder et al., 2005); in fact, we observed correlations spanning tens of thousands of cycles. In the Weibull application, sampling is far slower, and running chains of hundreds of millions of replicates is simply not feasible. Thus, we present blocked sampling (Roberts & Sahu, 1997) as a means of mitigating autocorrelation in Gibbs sampling.

Blocked Sampling

Up to now, the Gibbs sampling we have presented may be termed *componentwise*: For each iteration, parameters have been updated one at a time and in turn. This type of sampling is advocated because it is straightforward to implement and sufficient for many applications. It is well known, however, that componentwise sampling leads to autocorrelation in models with zero-centered random effects.

The autocorrelation problem comes about when the conditional posterior distributions are highly determined by conditioned-upon parameters and not by the data. In the hierarchical signal detection model, the sample of μ_h is determined largely by the sum of $\sum_i \alpha_i$ and $\sum_j \gamma_j$. Likewise, the sample of each α_i is determined largely by the values of μ_h. On iteration m, the value of $[\mu_h]_{m-1}$ influences each value of $[\alpha_i]_m$, and the sum of these $[\alpha_i]_m$ values has a large influence on μ_h. By transitivity, the value of $[\mu_h]_{m-1}$ influences the value of $[\mu_h]_m$; that is, they are correlated. This type of correlation is also true of μ_f, and subsequently for the overall estimates of sensitivity.

In blocked sampling, sets of parameters are drawn jointly, in one step, rather than componentwise. For the model on old-item trials (hits), the parameters (μ_h, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$) are sampled together from a multivariate distribution. Let $\boldsymbol{\theta}_h$ denote the vector of parameters: $\boldsymbol{\theta}_h = (\mu_h, \boldsymbol{\alpha}, \boldsymbol{\gamma})$. The derivation of conditional posterior distributions follows by multiplying the likelihood by the prior:

$$f\left(\boldsymbol{\theta}_h \mid \sigma_\alpha^2, \sigma_\gamma^2; \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \boldsymbol{\theta}_h, \sigma_\alpha^2, \sigma_\gamma^2\right) \times f\left(\boldsymbol{\theta}_h \mid \sigma_\alpha^2, \sigma_\gamma^2\right).$$

Each component of $\boldsymbol{\theta}$ has a normal prior. Hence, the prior $f(\boldsymbol{\theta}_h \mid \sigma_\alpha^2, \sigma_\gamma^2)$ is a multivariate normal. When the Albert and Chib (1995) alternative is used (Appendix B), the likelihood term becomes that of the latent data $\mathbf{w}$, which is distributed as a multivariate normal. The resulting conditional posterior is also a multivariate normal, but one with nonzero covariance terms (Lu, 2004). The other needed conditional posteriors are those for the variance terms (σ_α^2, σ_γ^2). These remain the same as before; they are distributed as inverse-gamma distributions. In order to run Gibbs sampling in the blocked format, it is necessary to be able to sample from a multivariate normal with an arbitrary covariance matrix, a feature that is built in to some statistical packages. An alternative is to use a Cholesky decomposition to sample a multivariate normal from a collection of univariate ones (Ashby, 1992). Conditional posteriors for new-item parameters (μ_f, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, σ_β^2, σ_δ^2) are derived and sampled analogously. The R code for this blocked Gibbs sampler may be found at www.missouri.edu/~pcl.

To test blocked sampling in this application, data were generated in the same manner as before (Equations 8–11, 20 participants observing 50 items, with item and participant increases in d' resulting in increased hit rate probabilities and decreased false alarm rate probabilities). Figure 19 shows the results. The two top panels show dramatically improved convergence over componentwise sampling (cf. Figure 18). Successive samples of sensitivity are virtually uncorrelated. The bottom left panel shows parameter estimation from both the conventional method (aggregating hit and false alarm events across items) and the hierarchical model. The conventional method has a consistent downward bias of .27 in this sample. As was mentioned before, this bias is asymptotic and remains even in the limit of increasingly large numbers of items. The hierarchical estimates are much closer to the diagonal and have no detectable overall bias. The hierarchical estimates are 40% more efficient than their conventional counterparts. The bottom right panel shows a comparison of the hierarchical estimates with the conventional ones. It is clear that the hierarchical model produces higher valued estimates, especially for smaller true values of d'.

There is, however, a noticeable but small misfit of the hierarchical model—a tendency to overestimate sensitivity for participants with smaller true values and to underestimate it for participants with larger true values. This type of misfit may be termed *over-shrinkage*. We suspect over-shrinkage comes about from the choice of priors. We discuss potential modifications of the priors in the next section. The good news is that over-shrinkage decreases as the number of items is increased. All in all, the hierarchical model presented here is a dramatic im-

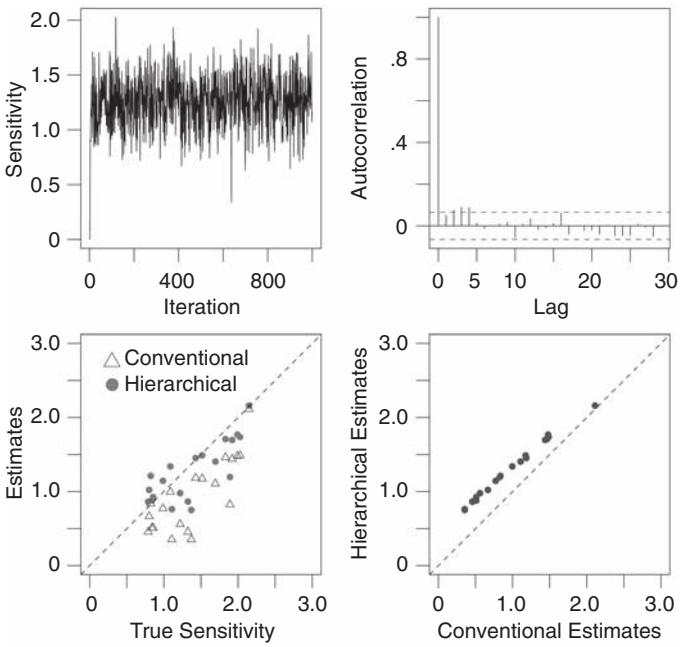

Figure 19. Analysis of a signal detection paradigm with participant and item variability. The top row shows that blocked sampling mitigates autocorrelation. The bottom row shows conventional and hierarchical estimates of participants' sensitivity as functions of the true sensitivity (left) and hierarchical estimates as a function of the conventional estimates (right). True random effects were generated with a mirror effect.

provement on the conventional practice of aggregating hits and false alarms across items.

The hierarchical model can also be used to explore the nature of participant and item effects. The data were generated by mirror effect principles: Increases in hit rates corresponded to decreases in false alarm rates. Although this principle was not assumed in analysis, the resulting estimates of participant and item effects should show a mirror effect. The left panel in Figure 20 shows the relationship between participant effects. Each point is from a different participant. The *y*-axis value of the point is α_i, the participant's hit rate effect; the *x*-axis value is β_i, the participant's false alarm rate effect. The negative correlation is expected: Participants with higher hit rates have lower false alarm rates. The right panel presents the same effects for items (γ_j vs. δ_j). As was expected, items with higher hit rates have lower false alarm rates.

To demonstrate the model's ability to disentangle different types of item and participant effects, we performed a second simulation. In this case, participant effects re-

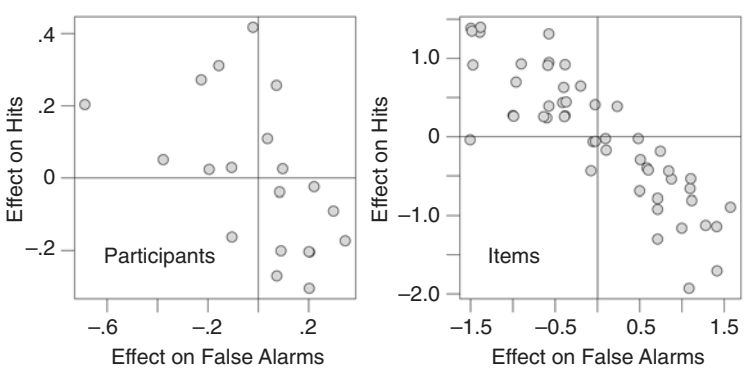

Figure 20. Relationship between estimated random effects when true values were generated with a mirror effect. (Left) Estimated participant effects on hit rate as a function of their effects on false alarm rate reveal a negative correlation. (Right) Estimated item effects also reveal a negative correlation.

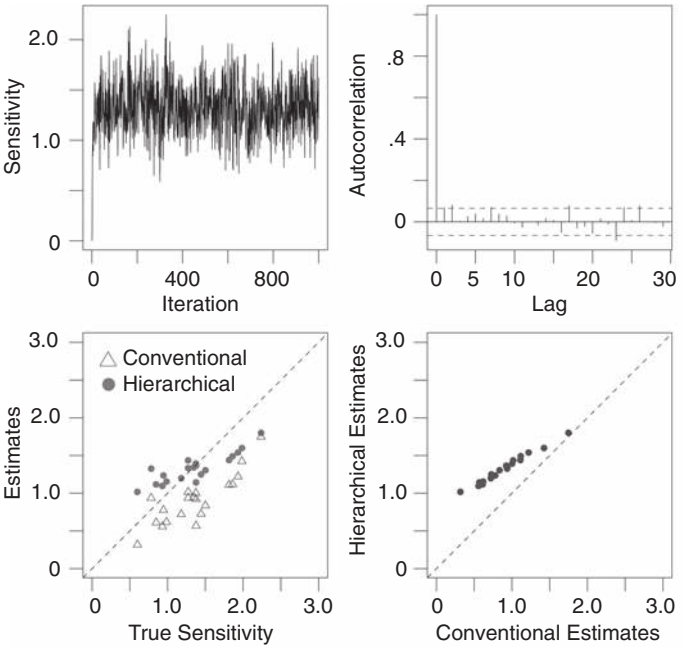

Figure 21. Analysis of the second simulation of a signal detection paradigm with participant and item variability. The top row shows little autocorrelation. The bottom row shows conventional and hierarchical estimates of participants' sensitivity as functions of the true sensitivity (left) and hierarchical estimates as a function of the conventional estimates (right). True participant random effects were generated with a mirror effect, and true item random effects were generated with a shift in baseline familiarity.

flected the mirror effect as before. Item effects, however, reflected baseline familiarity. Some items were assumed to be familiar, which resulted in increased hits and false alarms. Others lacked familiarity, which resulted in decreased hits and false alarms. This baseline-familiarity effect was implemented by setting the effect of an item on hits equal to that on false alarms ($\gamma_j = \delta_j$). Overall sensitivity results of the simulation are shown in Figure 21. These are highly similar to the previous simulation in that (1) there is little autocorrelation; (2) the hierarchical estimates are much better than their conventional counterparts; and (3) there is a tendency toward over-shrinkage. Figure 22 shows the relationships among the random effects. As expected, estimates of participant effects are negatively correlated, reflecting a mirror effect. Estimates of item effects are positively correlated, reflecting a

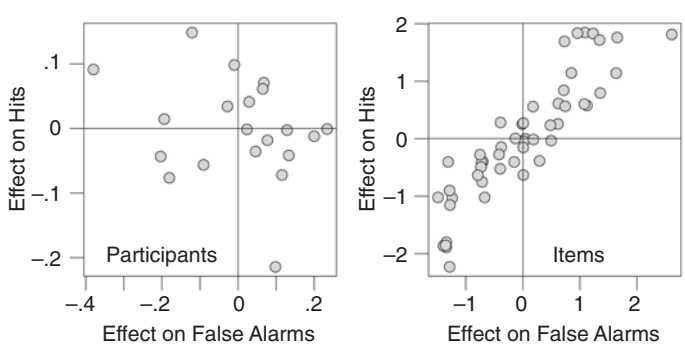

Figure 22. Relationship between estimated random effects when true participant effects were generated with a mirror effect and true item effects were generated with a shift in baseline familiarity. (Left) Estimated participant effects on hit rate as a function of their effects on false alarm rate reveal a negative correlation. (Right) Estimated item effects reveal a positive correlation.

baseline-familiarity effect. In sum, the hierarchical model offers detailed and relatively accurate analysis of participant and item effects in signal detection.

FUTURE DEVELOPMENT OF A SIGNAL DETECTION MODEL

Expanded Priors

The simulations reveal that there is some over-shrinkage in the hierarchical model (see Figures 19 and 21). We speculate that the reason this comes about is that the prior assumes independence between α and β and between γ and δ. This independence may be violated by negative correlation from mirror effects or by positive correlation from criterion effects or baseline-familiarity effects. The prior is neutral in that it does not favor either of these effects, but it is informative in that it is not concordant with either. A less informative alternative would not assume independence. We seek a prior in which mirror effects and baseline-familiarity effects, as well as a lack thereof, are all equally likely.

A prior that allows for correlations is given as follows:

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \underset{\sim}{\text{ind}} \text{Bivariate Normal}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(\mathbf{i})}\right)$$

and

$$\begin{pmatrix} \gamma_j \\ \delta_j \end{pmatrix} \underset{\sim}{\text{ind}} \text{Bivariate Normal}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(\mathbf{j})}\right)$$

The covariance matrices, $\boldsymbol{\Sigma}^{(\mathbf{i})}$ and $\boldsymbol{\Sigma}^{(\mathbf{j})}$, allow for arbitrary correlations of participant and item effects, respectively, on hit and false alarm rates. The prior on the covariance matrix elements must be specified. A suitable choice is that the prior on the off-diagonal elements be diffuse, and thus able to account for any degree and direction of correlation. One possible prior for covariance matrices is the Wishart distribution (Wishart, 1928), which is a conjugate form. Lu (2004) developed a model of correlated participant and item effects in a related application, Jacoby's (1991) process dissociation model. He provided conditional posteriors as well as an MCMC implementation. This approach looks promising, but significant development and analysis in this context are still needed.

Fortunately, the present independent-prior model is sufficient for asymptotically consistent analysis. The advantage of the correlated prior discussed above would be twofold: First, there may be some increase in the accuracy of sensitivity estimates. The effect would emerge for extreme participants, for whom there is a tendency in the independent-prior model to over-shrink estimates. The second benefit may be increased ability to assess mirror and baseline-familiarity effects. The present independent-prior model does not give sufficient credence to either of these effects; hence, its estimates of true correlations are biased toward 0. A correlated prior would lead to more power in detecting mirror and baseline-familiarity effects, should they be present.

Unequal Variance Signal Detection Model

In the present model, the variances of old- and new-item familiarity distributions are assumed to be equal. This assumption is fairly standard in many studies. Researchers who use signal detection in memory do so because it is a quick and principled psychometric model to measure sensitivity without a detailed commitment to the underlying processing. Our models are presented with this spirit; they allow researchers to assess sensitivity without undue influence from item variability.

There is some evidence that the equal-variance assumption is not appropriate. The experimental approach for testing the equal-variance assumption is to explicitly manipulate criteria and to trace out the receiver-operating characteristic (Macmillan & Creelman, 1991). Ratcliff, Sheu, and Gronlund (1992) followed this approach in a recognition memory experiment and found that the standard deviation of familiarity from the old-item distribution was .8 that of the new-item distribution. This result, taken at face value, provides motivation for future development. It seems feasible to construct a model in which the variance of the old-item familiarity distribution is a free parameter. This may be accomplished by adding a variance parameter to the probit transform of hit rate. Within the context of such a model, then, the variance may be simultaneously estimated along with overall sensitivity, participant effects, and item effects.

GENERAL DISCUSSION

This article presents an introduction to Bayesian hierarchical models. The main goals of this article have been to (1) highlight the dangers of aggregating data in nonlinear contexts, (2) demonstrate that Bayesian hierarchical models are a feasible approach to modeling participant and item variability simultaneously, and (3) implement a Bayesian hierarchical signal detection model for recognition memory paradigms. Our concluding discussion focuses on a few select issues in Bayesian modeling: subjectivity of prior distributions, model selection, pitfalls in Bayesian analysis, and the contribution of Bayesian techniques to the debate about the usefulness of significance tests.

Subjectivity of Prior Distributions

Bayesian analysis depends on prior distributions, so it is prudent to wonder whether these distributions add too much subjectivity into analysis. We believe that priors can be chosen that enhance analysis without being overly subjective. In many situations, it is possible to choose priors that result in posteriors that exactly match frequentist sampling distributions. For example, Bayesian estimation of the probability of success in a binomial distribution is the proportion of successes if the prior is a beta with parameters $a = b = 0$. A second example is

from the normal distribution. The posterior of the population mean is t distributed with the appropriate degrees of freedom when the prior on the mean is flat and the prior on variance is $1/\sigma^2$.

Although researchers may choose noninformative priors, it may be prudent in some situations to choose vaguely informative priors. There are two general reasons to consider an informative prior: First, vaguely informative priors are often more convenient than noninformative ones, and second, they may enhance estimation in those cases in which preexperimental knowledge is well established. We will examine these reasons in turn.

There are two ways in which vaguely informative priors are more convenient than noninformative ones. First, the use of vaguely informative conjugate priors greatly facilitates the derivation of posterior conditional distributions. This, in turn, allows for rapid sampling of conditionals in the Gibbs sampler in estimating marginal posterior distributions. Second, the vaguely informative priors here were proper, and this propriety guaranteed the propriety of the posteriors. The disadvantage of informative priors is the possibility that the choice of the prior will unduly influence the posterior. In the applications we presented, this did not occur for the chosen parameters of the prior. Increasing the variance of the priors above the chosen values has minuscule effects on estimates for reasonably sized data sets.

Although we highlight the convenience of vaguely informative, proper priors, researchers may also choose nonimformative priors. Noninformative priors, however, are often improper. The consequences of this choice are that (1) the researcher must prove the propriety of the posteriors, and (2) MCMC may need to be implemented with the somewhat more complicated Metropolis–Hastings algorithm rather than with Gibbs sampling. Neither of these activities are prohibitive, but they do require additional skill and effort. For the models presented here, vaguely informative conjugate priors were sufficient for valid analysis.

The second reason to use informative priors is that in some applications, researchers have a reasonable expectation of the range of data. In the signal detection example, we placed a normal prior on d' that was centered at 0 and had a very wide variance. More statistical power could have been obtained by using a prior with a majority of its mass above 0 and below 6, but this increase in power would have been slight. In estimating a response probability in a two-choice psychophysics experiment, it is advantageous to place a prior with more mass above .5 than below it. The hierarchical priors advanced in this article can be viewed as a form of prior information. This is information that people are not arbitrarily dissimilar—in other words, that true parameters come from orderly parent distributions (an alternative view, however, is presented by Lee & Webb, 2005). Judicious use of informative priors can enhance estimation.

We argue that the subjectivity in choosing priors is no greater than other elements of subjectivity in research. On the contrary, it may even be less so. Researchers routinely make seemingly arbitrary choices during the course of design and analysis. For example, researchers using response times typically discard those outside an arbitrary window as being too extreme to reflect the process of interest. Participants themselves may be discarded for excessive errors or for not following directions. The determination of a response-time window, an error-prone participant, or a participant not following directions imposes a fair degree of subjectivity on analysis. Within this context, the subjectivity of Bayesian priors seems benign.

Bayesian analysis may be less subjective because it can limit the need for other subjective practices. Priors in effect serve as a filter for cleaning data. In the present example, hierarchical priors mitigate the need to censor extremely poor-performing participants. The estimates from these participants are adjusted upward because of the influence of the prior (see Figure 15). Likewise, in our hierarchical response time models (Rouder et al., 2005; Rouder et al., 2003), there is no need to truncate long response times. The impact of these extreme observations is mediated by the prior. Hierarchical priors may be less arbitrary than other selection practices because the researcher does not have to specify criterial values for selection.

Although the use of priors may be advantageous, it does not come without risks—the main one being that the choice of a prior could unduly influence the outcome. The straightforward method of assessing this risk is to perform analyses repeatedly with a few different priors. The posterior will always be affected somewhat by the choice of prior. If this effect is marginal, then the resulting inference can be accepted with greater confidence. This need for assessing the effects of the prior on the posterior is similar in spirit to the safeguards required in frequentist analysis. For example, when truncating data, the researcher is obligated to explore the effects of different points of truncation in order to ensure that this fairly arbitrary decision only marginally affects the results.

Model Selection

We have not covered model testing or model selection in this article. One of the negative consequences of using hierarchical models is that individual parameter estimates are no longer independent. The value of an estimate for any participant depends, in part, on the values of others. Accordingly, it is invalid to analyze these parameters with ANOVA or ordinary least-squares regression to test hypotheses about groups or experimental manipulations.

One proper method of model selection (which includes hypothesis testing) is to compute and compare Bayes factors. This approach has been discussed extensively in the statistical literature (see Kass & Raftery, 1995; Meng & Wong, 1996) and has been imported to the psycho-

logical literature (see, e.g., Pitt, Myung, & Zhang, 2003). When using a Bayes factor approach, one computes the odds of one hypothesis being true relative to another hypothesis. Unlike estimation, computing Bayes factors is complicated, and a discussion is beyond the scope of this article. We anticipate that Bayes factors will play an increasing role in inference with nonlinear models.

Bayesian Pitfalls

There are a few special pitfalls of Bayesian analysis that deserve mention. First, there is a great temptation to use improper priors wherever possible. The main problem associated with improper priors is that without analytic proof, there is no guarantee that the resulting posteriors will be proper. To complicate the situation, MCMC sampling can be done, unwittingly, from improper posterior distributions, even though the analysis is meaningless. Researchers choosing improper priors should provide evidence that the resulting posterior is proper. The propriety of the posterior is not an issue if researchers choose proper priors, but the use of proper priors is not foolproof, either. In some situations, the variance of the posterior is directly related to the variance of the prior without constraint from the data (Hobert & Casella, 1996). This unacceptable situation is an example of undue influence of the prior. When this happens, the model under consideration is most likely not appropriate for analysis. Another pitfall is that in many hierarchical situations, MCMC integration converges slowly (see Figure 18). In these cases, researchers who do not check for convergence may fail to estimate true posterior distributions. In sum, although Bayesian approaches are conceptually simple and are easily implemented in high-level packages, researchers must approach both the issues of choosing a prior and checking the ensuing analysis with thoughtfulness and care.

Bayesian Analysis and Significance Tests

Over the last 40 years, researchers have offered various critiques of null-hypothesis significance testing (see, e.g., Cohen, 1994; Cumming & Finch, 2001; Hunter, 1997; Rozeboom, 1960; Smithson, 2001; Steiger & Fouladi, 1997). Some authors offer Bayesian inference as an alternative on the basis of its philosophical underpinnings (Lee & Wagenmakers, 2005; Pruzek, 1997; Rindskopf, 1997). Our rationale for advocating Bayesian analysis is different: We adopt the Bayesian approach out of practicality. Simply put, we know how to analyze these nonlinear hierarchical models in the Bayesian framework alone. Should there be tremendous advances in frequentist nonlinear hierarchical models that make their implementation more practical than Bayesian ones, we would gladly consider these advances.

Unlike those engaged in the significance test debate, we view both Bayesian and classical methods as having sufficient theoretical grounding. These methods are tools whose applicability depends on the situation. In many experimental situations, researchers interested in the mean difference between conditions or groups can safely draw inferences using classical tools such as *t* tests, ANOVA, and ordinary least-squares regression. Those worrying about robustness, power, and control of Type I error can increase the validity of analysis by replication. In many simple cases, Bayesian analysis is numerically equivalent to classical analyses. We are not advocating that researchers discard useful and convenient tools such as null-hypothesis significance tests within ANOVA and regression models; we are advocating that they consider modeling multiple sources of variability in analysis, especially in nonlinear contexts. In sum, the usefulness of the Bayesian approach is its tractability in these more complicated settings.

REFERENCES

Ahrens, J. H., & Dieter, U. (1974). Computer methods for sampling from gamma, beta, Poisson and binomial distributions. *Computing*, **12**, 223-246.

Ahrens, J. H., & Dieter, U. (1982). Generating gamma variates by a modified rejection technique. *Communications of the Association for Computing Machinery*, **25**, 47-54.

Albert, J., & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, **82**, 747-759.

Ashby, F. G. (1992). Multivariate probability distributions. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 1-34). Hillsdale, NJ: Erlbaum.

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.

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 & Language*, **81**, 55-65.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370-418.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, **49**, 997-1003.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational & Psychological Measurement*, **61**, 532-574.

Curran, T. C., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 531-547.

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.

Einstein, G. O., McDaniel, M. A., & Lackey, S. (1989). Bizarre imagery, interference, and distinctiveness. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 137-146.

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

Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', and min F'. *Journal of Verbal Learning & Verbal Behavior*, **15**, 135-142.

Gelfand, A. E., & 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 ed.). London: Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457-511.

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

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (Vol. 4, pp. 169-194). Oxford: Oxford University Press, Clarendon Press.

Gilks, W. R., Richardson, S. E., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. London: Chapman & Hall.

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, **25**, 425-431.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, **100**, 546-567.

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

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science*, **273**, 1699-1702.

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, & 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 & Review*, **7**, 185-207.

Hirshman, E., Whelley, M. M., & Palij, M. (1989). An investigation of paradoxical memory effects. *Journal of Memory & Language*, **28**, 594-609.

Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, **91**, 1461-1473.

Hohle, R. H. (1965). Inferred components of reaction time as a function of foreperiod duration. *Journal of Experimental Psychology*, **69**, 382-386.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, **8**, 3-7.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, **30**, 513-541.

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

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

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

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

Lee, M. D., & Webb, M. R. (2005). *Modeling individual differences in cognition*. Manuscript submitted for publication.

Lu, J. (2004). *Bayesian hierarchical models for process dissociation framework in memory research*. Unpublished manuscript.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103-189). New York: Wiley.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

Massaro, D. W., & Oden, G. C. (1979). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, **88**, 375-407.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

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

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

Pitt, M. A., Myung, I.-J., & Zhang, S. (2003). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491.

Pra Baldi, A., de Beni, R., Cornoldi, C., & Cavedon, A. (1985). Some conditions of the occurrence of the bizarreness effect in recall. *British Journal of Psychology*, **76**, 427-436.

Pruzek, R. M. (1997). An introduction to Bayesian inference and its applications. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (pp. 221-257). Mahwah, NJ: Erlbaum.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory & Language*, **41**, 416-426.

Raftery, A. E., & Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, **7**, 493-497.

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

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for decisions between two choices. *Psychological Science*, **9**, 347-356.

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

Riefer, D. M., & Rouder, J. N. (1992). A multinomial modeling analysis of the mnemonic benefits of bizarre imagery. *Memory & Cognition*, **20**, 601-611.

Rindskopf, R. M. (1997). Testing "small," not null, hypotheses: Classical and Bayesian approaches. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (pp. 221-257). Mahwah, NJ: Erlbaum.

Roberts, G. O., & Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B*, **59**, 291-317.

Rouder, J. N. (2000). Assessing the roles of change discrimination and luminance integration: Evidence for a hybrid race model of perceptual decision making in luminance discrimination. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 359-378.

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

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, **68**, 589-606.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, **57**, 416-428.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational & Psychological Measurement*, **61**, 605-632.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (pp. 221-257). Mahwah, NJ: Erlbaum.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (3rd ed.). New York: Springer.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, **22**, 1701-1728.

WICKELGREN, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology*, **5**, 102-122.

WISHART, J. (1928). A generalized product moment distribution in samples from normal multivariate population. *Biometrika*, **20**, 32-52.

WOLLEN, K. A., & COX, S. D. (1981). Sentence cueing and the effect of bizarre imagery. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 386-392.

NOTES

1. $\mathrm{Be}(a, b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$,

where Γ is the generalized factorial function; for example, $\Gamma(n) = (n-1)!$.

2. True probabilities for individuals were obtained by sampling a beta distribution with parameters $a = 35$ and $b = 15$.

APPENDIX A

This appendix describes the Albert and Chib (1995) method for estimating a probit-transformed probability. Let x_j be 1 if the participant is successful on the jth trial, and 0 otherwise. For each trial, a latent variable, w_j, is defined as follows:

$$x_j = \begin{cases} 1, & w_j \geq 0, \\ 0, & w_j < 0. \end{cases} \tag{A1}$$

Latent variables **w** are never known. All that is known is their sign. If a trial j was successful, it was positive; otherwise, it was nonpositive. Although **w** is not known, its construction is nonetheless essential.

Without any loss of generality, the random variable w_j is assumed to be a normal with a variance of 1.0. For Equation A1 to hold, the mean of these normals needs be z, where $z = \Phi^{-1}(p)$.[A1] Therefore,

$$w_j \overset{\text{ind}}{\sim} \text{Normal}(z, 1).$$

This construction of w_j is shown in Figure A1a. The parameters of interest are z and $\mathbf{w} = (w_1, \ldots, w_J)$; the data are $(x_1, \ldots, x_J)$. Note that if the latent variables **w** were known, the problem of estimating z would be simple: Parameter z is simply the population mean of **w** and could be estimated accordingly. The fact that **w** is latent will not prove to be a stumbling block.

The goal is to derive the marginal posterior of z. To do so, conditional posterior distributions are derived and then integrated using Gibbs sampling. The desired conditionals are $z \mid \mathbf{w}$ and $w_j \mid z; x_j, j = 1, \ldots, J$. We start with the former. Parameter z is the population mean of **w**; hence, this case is that of estimating a population mean from normally distributed observations (**w**). The prior on z is a normal with parameters μ_0 and σ_0^2, so the previous derivation is applicable. A direct application of Equations 23 and 24 yields that the conditional posterior $z \mid \mathbf{w}$ is a normal with parameters

$$\mu' = \left(N + \frac{1}{\sigma_0^2} \right)^{-1} \left(N\bar{w} + \frac{\mu_0}{\sigma_0^2} \right) \tag{A2}$$

and

$$\sigma^{2'} = \left(N + \frac{1}{\sigma_0^2} \right)^{-1}. \tag{A3}$$

The conditional $w_j \mid z; x_j$ is only slightly more complex. If x_j is unknown, latent variable w_j is normally distributed and centered at z. When w_j is conditioned on x_j, the sign of w_j is known. If $x_j = 0$, then by Equation A1 w_j must be negative. Hence, the conditional $w_j \mid z; x_j$ is a normal density that is truncated above at 0 (Figure A1b). Likewise, if $x_j = 1$, then w_j must be positive, and the conditional is a normal truncated below at 0 (Figure A1c). The conditional posterior, therefore, is a truncated normal, truncated either above or below depending on whether or not the trial was answered successfully.

Once conditionals are specified, analysis proceeds with Gibbs sampling. Sampling from a normal is straightforward; sampling from a truncated normal is not too difficult, and an algorithm for such sampling is coded in Appendix C.

(*Continued on next page*)

APPENDIX A (Continued)

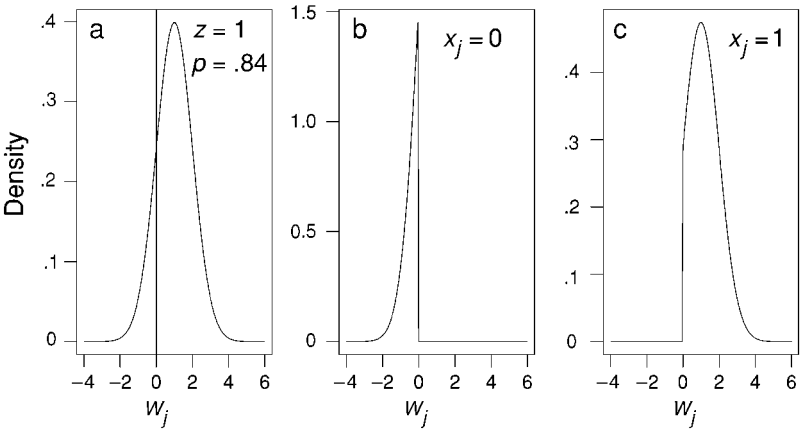

Figure A1. The Albert and Chib (1995) construction of a trial-specific latent variable for probit transforms. (a) Construction of the unconditional latent variable w_j for $p = .84$. (b) Conditional distribution of w_j on a failure on trial j. In this case, w_j must be negative. (c) Conditional distribution of w_j on a success on trial j. In this case, w_j must be positive.

NOTE TO APPENDIX A

A1. Let μ_w be the mean of **w**. Then, by construction,

$$\Pr(w_j < 0) = \Pr(x_j = 0) = 1 - p.$$

Random variable w_j is a normal with variance of 1.0. The term $\Pr(w_j < 0)$ is its cumulative distribution function evaluated at 0. Therefore,

$$\Pr(w_j < 0) = \Phi[(0 - \mu_w)/1] = \Phi(-\mu_w).$$

Combining this equality with the one above yields

$$\Phi(-\mu_w) = 1 - p,$$

and by symmetry of the normal,

$$\mu_w = \Phi^{-1}(p) = z.$$

APPENDIX B

Gibbs sampling for the hierarchical model is based on the Albert and Chib (1995) method, as follows: Let x_{ij} be the indicator for the ith person on the jth trial: $x_{ij} = 1$ if the trial is successful, and 0 otherwise. Latent variable w_{ij} is constructed as

$$x_{ij} = \begin{cases} 1, & w_{ij} \geq 0, \\ 0, & w_{ij} < 0. \end{cases} \tag{B1}$$

Conditional posterior $w_{ij} \mid z_i; x_{ij}$ is a truncated normal, as was discussed in Appendix A. Conditional posterior $z_i \mid \mathbf{w}, \mu, \sigma^2$ is a normal with mean and variance given in Equations A2 and A3, respectively (with the substitution of μ for μ_0 and σ^2 for σ_0^2). Conditional posterior densities of $\mu \mid \sigma^2, \mathbf{z}$ and $\sigma^2 \mid \mu, \mathbf{z}$ are given in Equations 25 and 28, respectively.

APPENDIX C

This appendix provides code in R for the hierarchical analysis of several individuals' probabilities. R is a freely available, easy-to-install, open-source statistical package based on SPlus. It runs on Windows, Macintosh, and UNIX platforms and can be downloaded from www.R-project.org.

```
#functions to sample from a truncated normal
#-------------------------------------------
#generates n samples of a normal (b,1) truncated below zero
rtnpos=function(n,b)
{
u=runif(n)
b+qnorm(1-u*pnorm(b))
}
#generates n samples of a normal (b,1) truncated above zero
rtnneg=function(n,b)
{
u=runif(n)
b+qnorm(u*pnorm(-b))
}
#-----------------------------
#generate true values and data
#-----------------------------
I=20 #Number of participants
J=50 #Number of Trials per participant
#generate each individual's true p
p=rbeta(I,35,15)
#result, use scan() to enter
#0.6932272 0.6956900 0.6330869 0.6504598 0.7008421 0.6589233 0.7337305
#0.6666958 0.5867875 0.7132572 0.7863823 0.6869082 0.6665865 0.7426910
#0.6649086 0.6212024 0.7375715 0.8025261 0.7587382 0.7363251
#generate each individual's observed numbers of successes
y=rbinom(I,J,p)
#result, use scan() to enter
#34 34 34 29 37 32 38 29 33 38 38 34 30 37 37 28 33 42 33 37
#---------------------------------------
#Analysis Set Up
M=10000 #total number of MCMC iterations
myrange=101:M #portion kept, burn-in of 100 iterations discarded
#needed vectors and matrices
z=matrix(ncol=I,nrow=M)
#matrix of each person's z at each iteration
#note z is probit of p e.g. z=qnorm(p)
sigma2=1:M
mu=1:M
#prior parameters
sigma0=100 #this parameter is a variance: sigma_0^2
mu0=0
a=.1
b=.1
#initial values
sigma2[1]=1
mu[1]=1
z[1,]=rep(1,I)
#shape for inverse gamma calculated outside of loop for speed
shape=(I)/2+a
#-----------------------------
#Main MCMC loop
#-----------------------------
for (m in 2:M) #iteration
{
for (i in 1:I) #participant
{
w=c(rtnpos(y[i],z[m-1,i]),rtnneg(J-y[i],z[m-1,i])) #samples of w|y,z
prec=J+1/sigma2[m-1]
mean.z=(sum(w)+mu[m-1]/sigma2[m-1])/prec
```

APPENDIX C (Continued)

```
z[m,i]=rnorm(1,mean.z,sqrt(1/prec))     #sample of z|w,mu,sigma2
}
prec=I/sigma2[m-1]+1/sigma0
mean.mu=(sum(z[m,])/sigma2[m-1]+mu0/sigma0)/prec
mu[m]=rnorm(1,mean.mu,sqrt(1/prec))     #sample of mu|z
rate=sum((z[m,]-mu[m])^2)/2+b
sigma2[m]=1/rgamma(1,shape=shape,scale=1/rate)  #sample of sigma2|z
}
#---------------------------------------
#Gather estimates of p for each individual
allp.est=pnorm(z[myrange,]) # MCMC chains of p
p.est.h=apply(allp.est,2,mean) #posterior means
```

(Manuscript received May 25, 2004;
revision accepted for publication January 12, 2005.)