PSYCHOMETRIKA
2007
DOI: 10.1007/s11336-005-1350-6

SIGNAL DETECTION MODELS WITH RANDOM PARTICIPANT AND ITEM EFFECTS

JEFFREY N. ROUDER
UNIVERSITY OF MISSOURI-COLUMBIA

JUN LU
AMERICAN UNIVERSITY

DONGCHU SUN, PAUL SPECKMAN, RICHARD MOREY, AND MOSHE NAVEH-BENJAMIN
UNIVERSITY OF MISSOURI-COLUMBIA

The theory of signal detection is convenient for measuring mnemonic ability in recognition memory paradigms. In these paradigms, randomly selected participants are asked to study randomly selected items. In practice, researchers aggregate data across items or participants or both. The signal detection model is nonlinear; consequently, analysis with aggregated data is not consistent. In fact, mnemonic ability is underestimated, even in the large-sample limit. We present two hierarchical Bayesian models that simultaneously account for participant and item variability. We show how these models provide for accurate estimation of participants' mnemonic ability as well as the memorability of items. The model is benchmarked with a simulation study and applied to a novel data set.

Key words: recognition memory, theory of signal detection, Bayesian models, hierarchical models, MCMC methods.

The theory of signal detection is a dominant psychometric model in perceptual and cognitive psychology. The focus of this paper is the application of signal detection to the measurement of mnemonic ability. The problem is complex because in typical designs, researchers sample both participants and items. Consequently, the effects of participants and items should be modeled as random effects. In conventional analysis in the memory literature, however, researchers aggregate scores across items. Aggregation implicitly treats items as fixed effects (Clark, 1973). Because the signal detection model is nonlinear, this misspecification leads to inconsistent estimation—estimates of mnemonic ability are asymptotically downward biased (Rouder & Lu, 2005; Wickelgren, 1968). To provide for accurate estimation, we propose hierarchical Bayesian models with two sets of random effects: one set for participants and another set for items.

The Theory of Signal Detection

The theory of signal detection was first proposed to study audition (Tanner & Birdsall, 1958). The goal in this early application was to measure a listener's ability to perceive pure tones embedded in noise. Since then, the theory of signal detection has become a dominant measurement

This research is supported by NSF grants SES-0095919 and SES-0351523, NIH grant R01-MH071418, a University of Missouri Research Leave grant and fellowships from the Spanish Ministry of Education and the University of Leuven, Belgium.

Requests for reprints should be sent to Jeffrey N. Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211, USA. E-mail: rouderj@missouri.edu

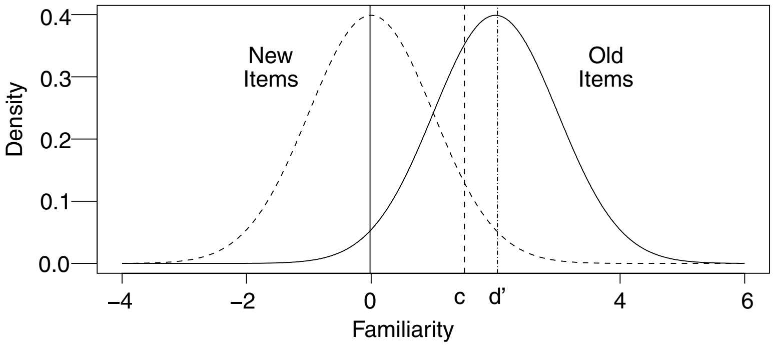

FIGURE 1.
The signal detection model. Solid and dashed curves denote densities of familiarity distributions for old and new items, respectively. Solid, dashed, and dashed–dotted vertical lines denote 0, criterion c, and sensitivity d', respectively. If familiarity is greater than criterion c, then an old-item response is produced; otherwise, a new-item response is produced.

model of participants' ability in many cognitive and perceptual domains. In particular, it has become a mainstay in the study of memory.

The signal detection model is often applied in a *recognition-memory paradigm*. The task consists of a study and a test phase. At study, the participant is presented a list of items and instructed to remember them for a later test. Then, after a suitable delay, the participant is tested. At test, items are presented sequentially. Some of these items were previously studied while others were not. The participant's task is to judge each item as previously studied (denoted an *old* item) or not (denoted a *new* item).

Figure 1 depicts the signal detection model. According to the model, when a participant is presented with an item at test, he or she assesses its familiarity. This familiarity varies from trial to trial because the participant is assumed to be a noisy system. The distribution of familiarity depends on whether the item was studied or not. New-item trials give rise to familiarities centered around zero; old-item trials, which have greater familiarity, give rise to distributions centered around parameter d' with $d' \geq 0$. Distributions of familiarity are assumed to be normal with unit variances. To make a decision, the participant sets a criterion c on familiarity. If the familiarity is greater than c, then the participant indicates that the item is old, otherwise he or she indicates that the item is new. Parameter d' serves as a measure of mnemonic ability and is often referred to as *sensitivity*. The assumption of equal variance is adopted for expository convenience and is easily generalized as discussed in the General Discussion and Appendix.

There is a second version of signal detection in which participants evaluate likelihood ratios of familiarity instead of familiarity itself. Let x denote the familiarity on a trial and let f_S and f_N denote the pdfs of familiarity for old and new items, respectively. Then, in this second version, participants compute the likelihood ratio g:

$$g(x) = f_S(x)/f_N(x).$$

Participants set a criterion δ on g and produce an old-item response if $g > \delta$ and a new-item response otherwise. For the case in which familiarity is distributed as a normal with constant variance, the model with criterion on familiarity (c) and on likelihood ratio (δ) are formally equivalent. There is a one-to-one monotonically increasing mapping from c to δ. The psychological underpinnings of the versions differ. In the likelihood-ratio version, participants must have at least implicit knowledge of the distributions of familiarity for studied and unstudied items. For the first version in which familiarity itself is directly assessed, this knowledge is not necessary. The validity of the two versions is assessed by studying how criteria change across different stimulus conditions. As the first version is more standard in recognition memory experiments, we develop it below, but return to the issue subsequently.

In recognition-memory experiments, a *hit* event is an old-item response to a studied item; a *false alarm* event is an old-item response to a new item. Probabilities of hits and false alarms, denoted by H and F, respectively, are given by

$$H = 1 - \Phi(c - d') = \Phi(d' - c), \tag{1}$$

$$F = 1 - \Phi(c) = \Phi(-c), \tag{2}$$

where Φ denotes the standard normal cdf. It is convenient to parametrize this model in terms of probit transforms of hit and false alarm probabilities. Let $h = \Phi^{-1}(H)$ and $f = \Phi^{-1}(F)$. Then,

$$h = d' - c,$$

$$f = -c.$$

There are two complementary events: a new-item response to a studied item (termed a *miss* event) and a new-item response to a new item (termed a *correct rejection* event). The probability of these events are complements of hit and false alarm probabilities, respectively. Therefore, consideration of hit and false alarms alone is sufficient for analysis.

The data in the recognition memory paradigm are nonidentically distributed dichotomous events: participants judge items as either old or new. The conventional approach is to aggregate responses across items, participants, or both. The proportion of studied items judged old is called the *hit rate*; the proportion of new items judged old is called the *false alarm rate*. Probit transforms of hit and false alarm rates serve as estimates of h and f, respectively.[1] Estimates of d' and c are given by

$$\hat{d'} = \hat{h} - \hat{f}, \tag{3}$$

$$\hat{c} = -\hat{f}. \tag{4}$$

When events are aggregated across items, it is implicitly assumed that each item has the same memorability and induces the same criterion. Likewise, when events are aggregated across participants, it is implicitly assumed that people have the same sensitivity and criterion. These implicit assumptions are assuredly too strict. People will vary from each other in their sensitivity (d') and in their criteria (c). Items will certainly vary from each other in sensitivity—some items may be more memorable than others. Likewise, there may be item effects on criteria. Some items may seem more familiar than others whether studied or not. This additional baseline familiarity will shift both the new-item and old-item distributions. This shift in distributions raises hits when the item is old and false alarms when it is new. It is equivalent to a criterion shift.

The Deleterious Effects of Unmodeled Variability

Unfortunately, aggregation may lead to an asymptotic underestimation of sensitivity in the signal-detection model (e.g., Rouder & Lu, 2005). We performed a small simulation to assess the effects of aggregation as follows: Let d'_{ij} and c_{ij} denote the sensitivity and criterion of the ith person responding to the jth item, respectively. Each d'_{ij} was the sum of a participant effect and an item effect, $d'_{ij} = a_i + b_j$, with a_i sampled from a log-normal distribution and b_j

[1] On occasion, participants produce no misses or no false alarms. Probit transforms are not finite in these cases. There are several reasonable alternatives proposed in the literature (Hautus & Lee, 1998; McMillan & Creelman, 1991; Snodgrass & Corwin, 1985). The most common alternative is to add half a count to hit, miss, false-alarm, and correct-rejection frequencies. We adopt this approach in analysis of experimental data.

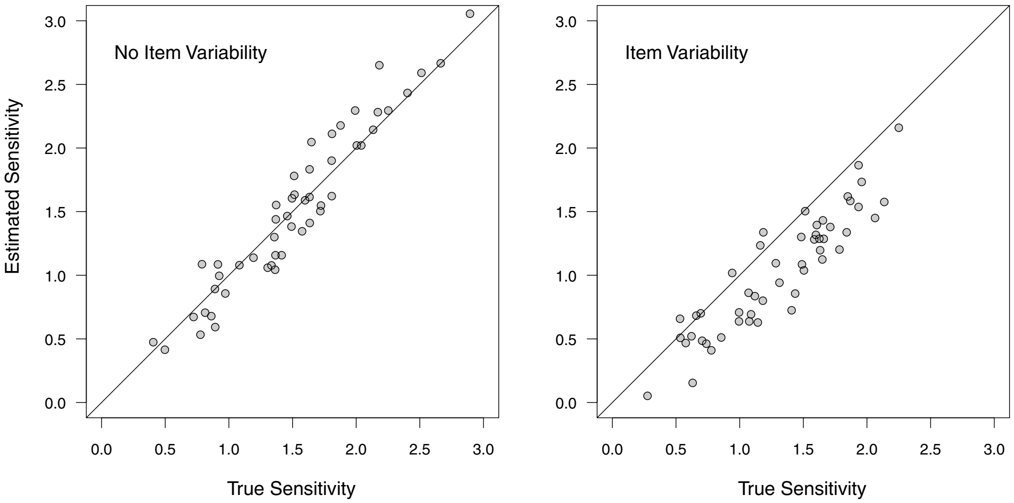

FIGURE 2.
The effect of unmodeled item variability on the estimation of sensitivity (d'). The left panel shows the results with no item variability; the right panel shows the same with a high degree of item variability. Figure reprinted from Rouder and Lu (2005). Permission pending.

sampled from a zero-centered normal distribution. Criterion c_{ij} was assumed to be unbiased for each participant-item combination, i.e., $c_{ij} = d'_{ij}/2$. The motivation for this choice comes from an empirically observed phenomenon, the mirror effect, which is discussed subsequently. From these sensitivities and criteria, true participant-by-item hit and false-alarm probabilities were computed by (1) and (2). Simulated data were generated from these probabilities—there were 50 hypothetical participants observing 50 old-item trials and 50 new-item trials. Event frequencies were aggregated across items for each participant, and an individualized d' was estimated. The left-hand panel of Figure 2 shows these estimates against the true values for the case in which there was no variability in b_j (i.e., $b_j = 0, \forall j$). In this case, the implicit independence-and-identically-distributed (iid) assumption about items is met and aggregation is valid. Not surprisingly, sensitivity estimates show no systematic bias. The right-hand panel shows the case in which there is variability in b_j. This variability violates the iid assumption implicit in aggregation. This violation has a deleterious effect—sensitivity estimates are too low. Most troubling, this bias is asymptotic.

Participant and item variability should be treated as more than just a nuisance. There are many participant variables that influence mnemonic ability including age, vocabulary, and intelligence. Studying how individual differences affect memory has been a productive line in theory building (e.g., Kane, Hambrick, Tuholski, Wilhelm, Payne, & Engle, 2004; Salthouse, 1996). Likewise, there are item variables that affect memory. One example is word frequency. Word frequency is literally the number of times a word appears in a corpus of printed text (Kucera & Francis, 1967). High-frequency words, such as *dog* are used often; low-frequency words, such as *lynx*, are rare. Low-frequency words are better recognized than high-frequency words (Glanzer, Adams, Iverson, & Kim, 1993; Gillund & Shiffrin, 1984). This fact is somewhat surprising as people have more experience with high-frequency words. The low-frequency advantage provides constraints on mnemonic theories—they cannot simply be about fluency with material.[2] In sum, participant and item effects themselves are suitable targets of measurement for theory construction.

[2] Current theories of recognition memory implicitly assume that the noise in mnemonic systems is more similar to high-frequency items than to low-frequency items. Hence, against this noise, low-frequency items are more distinct than high-frequency items (e.g., Shiffrin & Steyvers, 1997).

Two Hierarchical Models with Random Effects

To provide for accurate estimation of participant and item effects, we propose two hierarchical signal detection models. These models posit random participant and item effects. Instead of placing models on sensitivity and criterion, it is more convenient to place them on probit transforms of hit and false alarm probabilities. Let h_{ij} and f_{ij} be the hit and false alarm probits, respectively, for the ith participant tested on the jth item. We model h_{ij} and f_{ij} as the sum of participant and item effects:

$$h_{ij} = \mu^{(h)} + \alpha_i^{(h)} + \beta_j^{(h)},$$

$$f_{ij} = \mu^{(f)} + \alpha_i^{(f)} + \beta_j^{(f)}.$$

Parameters $\mu^{(h)}$ and $\mu^{(f)}$ are overall means; parameters $\alpha_i^{(h)}$ and $\alpha_i^{(f)}$ are participant effects; and parameters $\beta_j^{(h)}$ and $\beta_j^{(f)}$ are item effects. These participant and item effects are modeled with zero-centered bivariate normal distributions:

$$\begin{pmatrix} \alpha_i^{(h)} \\ \alpha_i^{(f)} \end{pmatrix} \overset{\text{iid}}{\sim} N_2(\mathbf{0}, \mathbf{\Sigma}_\alpha), \qquad i = 1, \ldots, I, \tag{5}$$

$$\mathbf{\Sigma}_\alpha = \begin{pmatrix} \sigma^2_{\alpha,h} & \rho_\alpha \sigma_{\alpha,h}\sigma_{\alpha,f} \\ \rho_\alpha \sigma_{\alpha,h}\sigma_{\alpha,f} & \sigma^2_{\alpha,f} \end{pmatrix},$$

$$\begin{pmatrix} \beta_j^{(h)} \\ \beta_j^{(f)} \end{pmatrix} \overset{\text{iid}}{\sim} N_2(\mathbf{0}, \mathbf{\Sigma}_\beta), \qquad j = 1, \ldots, J, \tag{6}$$

$$\mathbf{\Sigma}_\beta = \begin{pmatrix} \sigma^2_{\beta,h} & \rho_\beta \sigma_{\beta,h}\sigma_{\beta,f} \\ \rho_\beta \sigma_{\beta,h}\sigma_{\beta,f} & \sigma^2_{\beta,f} \end{pmatrix}.$$

The covariance structures on the bivariate normals allow for arbitrary variance and correlation of participant effects on hit and false alarm rates; the same applies for item effects. We call this model the *correlated random effects signal detection model* and refer to it as the *correlated model*.

The correlated model is highly similar to our hierarchical process dissociation memory model (Lu, Speckman, Sun, & Rouder, submitted; Rouder, Lu, Morey, Sun, & Speckman, submitted). The goal of the process dissociation procedure is to separate conscious recollection and automatic activation processes in memory recall (Jacoby, 1991). Analogous to the current development, we modeled probit transforms of conscious recollection and automatic activation probabilities as the sum of a grand mean, participant effects, and item effects. Participant effects across conscious recollection and automatic activation processes were assumed to arise from a bivariate normal with possible correlation. Likewise, item effects across both processes were similarly modeled as arising from a bivariate normal. Hence, development of mixed linear models is broadly applicable in several mnemonic measurement models.

For signal detection we also consider the case without correlation; i.e., the submodel of (5) and (6) with $\rho_\alpha = \rho_\beta = 0$. For $i = 1, \ldots, I$ and $j = 1, \ldots, J$:

$$\alpha_i^{(h)} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2_{\alpha,h}),$$

$$\alpha_i^{(f)} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2_{\alpha,f}),$$

$$\beta_j^{(h)} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2_{\beta,h}),$$

$$\beta_j^{(f)} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2_{\beta,f}).$$

We call this model the *independent random effects signal detection model* and refer to it as the *independent model*. We first proposed the independent model in Rouder and Lu (2005).

In both the correlated and independent models, participant-by-item sensitivity and criterion are given by $d'_{ij} = h_{ij} - f_{ij}$ and $c_{ij} = -f_{ij}$, respectively. Sensitivity and criterion for the ith participant are given by $\mu^{(h)} + \alpha_i^{(h)} - \mu^{(f)} - \alpha_i^{(f)}$ and $-\mu^{(f)} - \alpha_i^{(f)}$, respectively. Sensitivity and criterion for the jth item are given by analogous expressions with $\beta_j^{(h)}$ and $\beta_j^{(f)}$. Overall sensitivity and criterion are given by $\mu^{(h)} - \mu^{(f)}$ and $-\mu^{(f)}$, respectively.

Plausibility of Correlated Random Effects

In the preceding section we proposed two models: one in which random effects were assumed to be independent and another in which they are allowed to be correlated. This correlation has important theoretical interpretations and is used to assess whether participants set criteria on familiarity itself or on likelihood ratios. The most studied correlation in recognition memory is the mirror effect. The mirror effect refers to a negative correlation between hit and false alarm rates (Glanzer et al., 1993). Perhaps the most dramatic display comes from Singer, Gagnon, and Richard (2002). These researchers asked participants to recognize sentences. Some of the sentences were studied and then a delay was introduced. After the delay, more sentences were studied and then there was a recognition memory test. Not surprisingly, old sentences studied after the delay produced more hits than those studied before the delay. More surprisingly, new sentences similar to those studied after the delay produced fewer false alarms to new sentences similar to those studied before the delay. This change in false alarms implies that the criterion on familiarity shifted as a function of delay, even for items not studied. Current theoretical explanations of the mirror effect invoke the likelihood ratio version of the signal detection theory (e.g., Glanzer et al., 1993; McClelland & Chappell, 1994; Shiffrin & Steyvers, 1997). In these theories, participants compute a separate likelihood ratio for each class of stimuli, implying that participants operate with a large degree of knowledge about the mnemonic properties of different classes.

Mirror effects are not universally observed. Some variables affect the criterion and induce a positive correlation between hits and false alarms. Depressed elderly people, for example, tend to have exceptionally conservative criteria emphasizing both correct rejection and misses (Miller & Lewis, 1977). Other variables affect either hit or false-alarm rates, exclusively. Stretch and Wixted (1998), for example, asked participants to study some words one time (weak items) and others three times (strong items). To make this manipulation clear to participants, strong items were colored blue while weak items were colored red. At test, strong and weak targets appeared in their respective colors. New items were also colored; half were red and half were blue. Hit rate varied as a function of color, with strong (blue) words having a higher hit rate than weak (red) ones. False alarm rates, however, did not vary with color. Thus, the effect of increasing the strength of a word through repetition affects only the hit rate and not the false-alarm rate. These effects, in contrast to mirror effects, imply a constant criterion on familiarity itself rather than on the likelihood ratio.

Mapping out the correlation among hits and false alarms effects across conditions, items, and participants is germane for theory construction. The hierarchical models provide for the principled measurement of these correlations. At first glance, it would appear that the appropriate model is the correlated one, which can account for any degree of correlation including no correlation. Yet, this reasoning is not complete. The independent model is consistent, and this implies that as the sample size is increased, true correlation in random effects will be evident in their estimates. For example, the correlation between participant random effects may be explored by examining scatter plots of $\hat{\alpha}_i^{(h)}$ versus $\hat{\alpha}_i^{(f)}$. The difference between the independent model and the correlated is a difference in priors. The prior for the correlated model allows for correlation among random effects; the prior for the independent model does not. With sufficient data, these priors are less influential. The advantages of using the correlated model is that it is possible to perform formal inference on the correlations among random effects. There are, however, disadvantages to the correlated model as the priors we employ are necessarily informative. We discuss these issues next.

Prior Distributions of Parameters

We analyzed the models in the Bayesian framework and used Markov chain Monte Carlo (MCMC) sampling (Gelfand & Smith, 1990) to estimate posterior distributions of parameters. Priors are needed for grand means ($\mu^{(h)}$, $\mu^{(f)}$) and the variance (and covariance) of the random effects. For the independent model, specifying priors is fairly straightforward. In Bayesian analysis, flat distributions[3] are commonly used as priors on grand mean parameters, and inverse gamma distributions are commonly used as priors on variance parameters. The density of the inverse gamma is given by

$$f(\sigma^2 \mid a, b) = \frac{b^a}{\Gamma(a)(\sigma^2)^{a+1}}\, e^{-\frac{b}{\sigma^2}}, \quad a, b, \sigma^2 > 0.$$

As the parameters a and b approach 0, the prior approaches $1/\sigma^2$, which is the Jeffreys prior for the variance (Jeffreys, 1961) in the simpler problem of estimating the variance of a normal distribution. We used values of $a = b = .01$ to approximate the Jeffreys prior. One must be careful, however, in using priors that are improper or nearly so like the flat prior or the Jeffreys prior. In some cases, improper priors lead to improper posteriors, which invalidates MCMC sampling (Hobert & Casella, 1996). When nearly improper priors are used, one should determine if the Bayes inference is influenced by prior values.

We chose the inverse Wishart distribution as priors for the covariance matrices ($\mathbf{\Sigma}_\alpha$, $\mathbf{\Sigma}_\beta$) in the correlated model. Choice of priors for covariance matrices is an active area of research and the inverse Wishart is convenient. This distribution is semiconjugate in the sense that the full conditional posteriors of $\mathbf{\Sigma}_\alpha$ and $\mathbf{\Sigma}_\beta$ are also inverse Wishart. The two-dimensional inverse Wishart density on a covariance matrix $\boldsymbol{S}$ with m degrees of freedom and scale matrix $\mathbf{\Omega}$ has density

$$f(\boldsymbol{S} \mid m, \mathbf{\Omega}) = \frac{|\mathbf{\Omega}|^{m/2}|\boldsymbol{S}|^{-(m+3)/2}}{2^m\Gamma_2(m/2)} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{\Omega}\boldsymbol{S}^{-1})\right), \tag{7}$$

where $\boldsymbol{S}$ is a positive definite matrix and Γ_p is the multivariate gamma function given by $\Gamma_p(a) = \pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma(a - \frac{1}{2}(j-1))$ (see, e.g., Gelman, Carlin, Stern, & Rubin, 2004, p. 574). In the bivariate case, the mean of the distribution only exists for $m \geq 4$, in which case it is $\mathbf{\Omega}/(m-3)$. Parameter m must be at least 2 for the prior to be proper.

Parameters m and $\mathbf{\Omega}$ must be selected beforehand. The choice $m = 2$ is natural as it is least informative. The choice of $\mathbf{\Omega}$, however, is more complicated. Because it is plausible that the correlation between hits and false alarms in the data may be negative, positive, or null, there should be little information about the direction and magnitude of the correlation coefficient. This lack of information is represented by assigning off-diagonal elements of $\mathbf{\Omega}$ to zero. Likewise, there is no a-priori information about the relative size of variability of random effects for hits and

[3] Flat distributions may be approximated by normal distributions with large variance.

TABLE 1.
Effect of ω on marginal distribution of variance with inverse Wishart prior (7) and (8).

	Estimated quantiles						
ω	.01	.1	.3	.5	.7	.9	.99
.1	.015	.037	.093	.22	.67	6.3	630
1	.15	.37	.93	2.2	‘6.7	63	6300
10	1.5	3.7	9.3	22	67	630.	63,000

Note: Quantiles estimated from 500,000 samples.

false alarms. This lack of information is represented by assigning diagonal terms of $\boldsymbol{\Omega}$ equal to each other. With these two restrictions,

$$\boldsymbol{\Omega} = \begin{pmatrix} \omega & 0 \\ 0 & \omega \end{pmatrix}.$$

The advantage of this prior is that it is fairly diffuse and properly skewed (see Table 1). The disadvantage, however, is that this prior is informative. The value of ω serves as a scale factor on the variance. To explore this prior, we simulated the inverse Wishart and listed estimated quantiles as a function of ω (see Table 1). The larger ω, the more mass is concentrated on larger values of variances. This scaling property means that the researcher must make an assumption about the size of the variance for random effects. Because it may be difficult to develop prior knowledge for these variances, we proceed by repeating the analysis for a few values of ω that span a wide range. In the simulations and application, we performed the analysis with three different values of ω ($\omega = .1, 1, 10$) to assess the effects of different choices. The choice of $\omega = 1$ corresponds to the Wishart prior studied by Browne and Draper (2000).

The situation is far more sanguine with regards to the prior on correlation coefficient—the choice of ω does not affect the marginal prior. Figure 3 shows this prior for $m = 2$. The marginal prior has increased mass at the extreme values of $\rho = -1$ and $\rho = 1$. This type of prior is similar in form to the noninformative Haldane prior on a probability parameter.

Method of Analysis

Derivation of closed-form expressions for the marginal or joint posterior distributions is intractable. Instead, posterior quantities are estimated by MCMC, specifically via Gibbs sampling (Gelfand & Smith, 1990). We discuss how to sample from the conditional posteriors necessary for Gibbs sampling.

In practice, it is often useful to treat some groups of parameters as blocks to reduce autocorrelation in the Markov chain (Roberts & Sahu, 1997). Previously, we have found that taking the grand mean and the random effects together as a block greatly speeds convergence and reduces the autocorrelation in MCMC (Rouder & Lu, 2005; Lu, Sun, Speckman, & Sun, submitted). Let $\boldsymbol{\mu} = (\mu^{(h)}, \mu^{(f)})^t$, $\boldsymbol{\alpha} = (\alpha_1^{(h)}, \alpha_1^{(f)}, \ldots, \alpha_I^{(h)}, \alpha_I^{(f)})^t$, and $\boldsymbol{\beta} = (\beta_1^{(h)}, \beta_1^{(f)}, \ldots, \beta_I^{(h)}, \beta_J^{(f)})^t$; where I and J are the number of participants and items, respectively. With this notation, the vector of additive components is given by $\boldsymbol{\lambda} = (\boldsymbol{\mu}^t, \boldsymbol{\alpha}^t, \boldsymbol{\beta}^t)^t$.

To simplify Gibbs sampling, we follow the data-augmentation method of Albert and Chib (1995) (see Rouder & Lu, 2005, for a tutorial review). The method is based on positing a latent

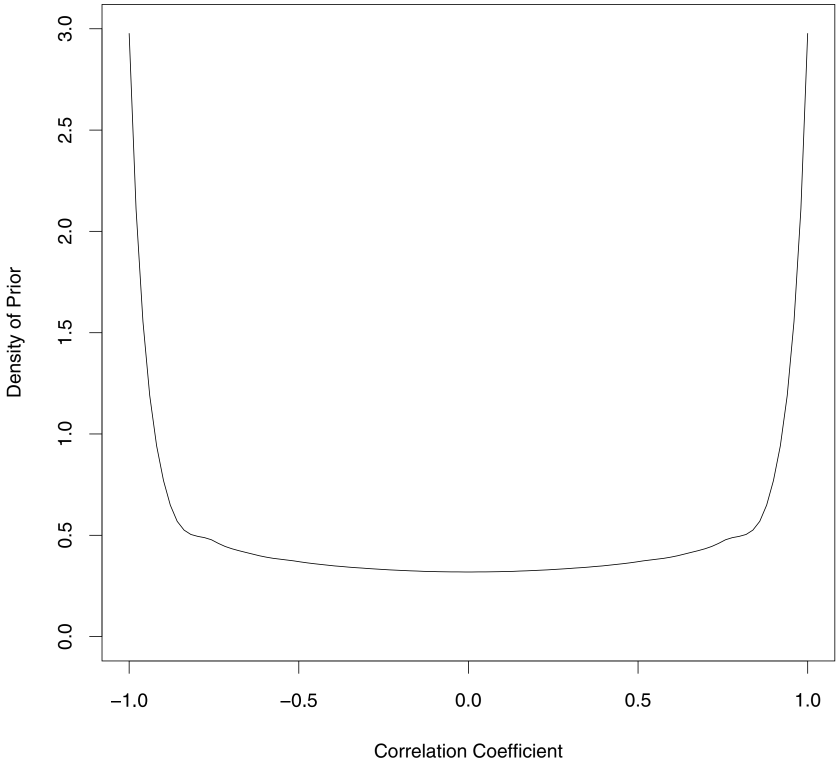

FIGURE 3.
Marginal prior on correlation coefficient. This prior does not depend on the choice of ω and has increased mass for extreme values.

vector of normally distributed observations. Let $\boldsymbol{X}$ denote the design matrix that relates the block of parameters ($\boldsymbol{\lambda}$) to observations. Let $\boldsymbol{y}$ denote the vector of observations. Note that

$$\Pr(y_k = 1) = \Phi([\boldsymbol{X\lambda}]_k),$$

where $k = 1, \dots, IJ$ indexes the row in the design matrix. For the ith participant observing the jth item, k is defined as $k = (i-1)J + j$. For our case, there are IJ latent observations:

$$w_k \overset{\text{indep}}{\sim} \text{Normal}([\boldsymbol{X\lambda}]_k, 1).$$

With this construction, it is fairly simple to show that $\Pr(y_k = 1) = \Pr(w_k > 0)$. It is simpler to sample from the full conditional distributions of the parameters conditioned on $\boldsymbol{w}$ than on $\boldsymbol{y}$. We do so below.

Independent Model

It is sufficient to derive the following full conditionals for analysis: $(\boldsymbol{\lambda} \,|\sigma^2_{\alpha,h}, \sigma^2_{\alpha,f}, \sigma^2_{\beta,h}, \sigma^2_{\beta,f}, \boldsymbol{w})$, $(\boldsymbol{\sigma}^2_{m,n} \mid \boldsymbol{\lambda})$ for $m = \alpha, \beta$ and $n = h, f$, and $(\boldsymbol{w} \mid \boldsymbol{\lambda}; \boldsymbol{y})$. These conditionals are specified below in Facts A, B, and C. Proofs are straightforward.

- *Fact* A. Let $\boldsymbol{B}_\alpha = \text{diag}(\sigma^{-2}_{\alpha,h}, \sigma^{-2}_{\alpha,f})$ be the 2×2 diagonal matrix with diagonal elements $(\sigma^{-2}_{\alpha,h}, \sigma^{-2}_{\alpha,f})$ and, similarly let, $\boldsymbol{B}_\beta = \text{diag}(\sigma^{-2}_{\beta,h}, \sigma^{-2}_{\beta,f})$. Finally, let $\boldsymbol{B}_I = \text{diag}(0, 0, \boldsymbol{B}_\alpha, \dots, \boldsymbol{B}_\alpha, \boldsymbol{B}_\beta, \dots, \boldsymbol{B}_\beta)$ be the block diagonal precision matrix of $\boldsymbol{\lambda}$ with $\boldsymbol{B}_\alpha$ repeated I times and $\boldsymbol{B}_\beta$ repeated J times. With this notation, the full conditional distribution of $\boldsymbol{\lambda}$ is

$$\boldsymbol{\lambda} \mid \sigma^2_{\alpha,h}, \sigma^2_{\alpha,f}, \sigma^2_{\beta,h}, \sigma^2_{\beta,f}, \boldsymbol{w} \sim N_q(\boldsymbol{V}_I \boldsymbol{X}^t \boldsymbol{w}, \boldsymbol{V}_I), \tag{8}$$

where $q = 2(1 + I + J)$ and $\boldsymbol{V}_I = (\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{B}_I)^{-1}$.

- *Fact* B. The full conditional distribution of $\boldsymbol{\sigma}^2_{\alpha,h}$ is

$$\boldsymbol{\sigma}^2_{\alpha,h} \mid \boldsymbol{\lambda} \sim \text{Inverse Gamma}\left(I/2 + a, \sum_i (\alpha_i^{(h)})^2/2 + b\right), \tag{9}$$

where a and b are parameters of the prior. The conditional posteriors for the other three variances are expressed analogously.

- *Fact* C. The full conditional distributions of $w_1, \dots, w_{IJ}$ are

$$w_k \mid \boldsymbol{\lambda}; \boldsymbol{y} \overset{\text{indep}}{\sim} \begin{cases} N^-([\boldsymbol{X\lambda}]_k, 1), & y_k = 0, \\ N^+([\boldsymbol{X\lambda}]_k, 1), & y_k = 1, \end{cases} \tag{10}$$

where N^- and N^+ denote normal distributions truncated at 0 from above and below, respectively.

Gibbs sampling was implemented in the **R** statistical language (R Foundation, `www.r-project.org`). To sample these conditionals, it is necessary to sample from a multivariate normal, an inverse gamma, and a truncated univariate normal. Sampling from a multivariate normal is provided in the MASS package. Sampling from an inverse gamma is accomplished by taking the reciprocal of samples from a gamma distribution. Sampling a truncated normal may be accomplished by the inversion method (Devroye, 1986): to sample from the $Y \sim N(\mu, \sigma^2)$ distribution truncated at (a, b), sample a uniform variate U and take

$$Y = \sigma\Phi^{-1}\left(U\left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right] + \Phi\left(\frac{a-\mu}{\sigma}\right)\right) + \mu.$$

In this expression, one may take $a = -\infty$ or $b = \infty$.

Correlated Model

It is sufficient to derive the following full conditional distributions for analysis: $(\boldsymbol{\lambda} \mid \boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta, \boldsymbol{w})$, $(\boldsymbol{\Sigma}_\alpha \mid \boldsymbol{\lambda})$, $(\boldsymbol{\Sigma}_\beta \mid \boldsymbol{\lambda})$ and $(\boldsymbol{w} \mid \boldsymbol{\lambda}; \boldsymbol{y})$. The last of these conditionals has already been provided as Fact C. The first two are presented below as Facts D and E.

- *Fact* D. Let $\boldsymbol{B}_C$ be the block diagonal matrix denoting the precision of $\boldsymbol{\lambda}$ given by

$$\boldsymbol{B}_C = \text{diag}(0, 0, \boldsymbol{\Sigma}_\alpha^{-1}, \dots, \boldsymbol{\Sigma}_\alpha^{-1}, \boldsymbol{\Sigma}_\beta^{-1}, \dots, \boldsymbol{\Sigma}_\beta^{-1}),$$

where $\boldsymbol{\Sigma}_\alpha^{-1}$ is repeated I times and $\boldsymbol{\Sigma}_\beta^{-1}$ is repeated J times. With this notation, the full conditional distribution of $\boldsymbol{\lambda}$ is

$$\boldsymbol{\lambda} \mid \boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta, \boldsymbol{w} \sim N_q(\boldsymbol{V}_C \boldsymbol{X}^t \boldsymbol{w}, \boldsymbol{V}_C), \tag{11}$$

where $q = 2(1 + I + J)$ and $\boldsymbol{V}_C = (\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{B}_C)^{-1}$.

- *Fact* E. Let $\boldsymbol{\alpha}_i = (\alpha_i^{(h)}, \alpha_i^{(f)})^t$, and let $\boldsymbol{S}_\alpha = \sum_i \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^t$ be the 2×2 matrix formed by summing the outer products. The full conditional distribution of $\boldsymbol{\Sigma}_\alpha$ is

$$\boldsymbol{\Sigma}_\alpha \mid \boldsymbol{\lambda} \sim \text{Inverse Wishart}(I + m, \boldsymbol{\Omega} + \boldsymbol{S}_\alpha). \tag{12}$$

The condition posterior for $\boldsymbol{\Sigma}_\beta$ is expressed analogously.

This model was also implemented in **R**, Sampling from the inverse Wishart distribution was provided by the MCMC package. Source code for both models may be found at `web.missouri.edu/~umcaspsychpcl/code`.

Simulation Studies

To get a reasonable idea of the performance of the two models, we simulated data with three basic relationships: strong positive correlation ($\rho = .9$), strong negative correlation ($\rho = -.9$) , and independence ($\rho = 0$). We chose these values as they seem to span the range of plausible values. Crossing these three with the two types of random effects (participants and items) yields nine variants. True values of grand-mean parameters were set to $\mu^{(h)} = .75$ and $\mu^{(f)} = -.75$, which is typical of data in the literature. True values of the variance of participant and item random effects were .5. The literature gives little guidance as to the choice of these variances, and it seems prudent to set these too high rather than too low. For each of the nine variants, we simulated the case of 30 hypothetical participants tested with 20 old items and 20 new items. These sample sizes are atypically small for the majority of behavioral studies. They are, however, more typical of cognitive neuroscience research on special populations (e.g., Alzheimer's patients) and research with functional imaging. Smaller sample sizes are ideal for the simulation study as they highlight the influence of the priors.

Simulations proceeded as follows. First, random participant and item effects were drawn from normals as discussed above. From these random effects, true participant-by-item values of h_{ij} and f_{ij} were calculated. These participant-by-item probabilities were used to generate Bernoulli-trial outcomes, which served as data. These data were then analyzed with the two Bayesian models and an aggregation method as discussed below. After results were tabulated, a new replicate experiment was performed. There were 400 replicate experiments for each of the nine variants.

Analysis

Conventional analysis was done by two methods. In one case, we aggregated across items to produce participant-specific sensitivities. In the second case, we aggregated across participants to produce item-specific sensitivities. Whereas the variance across items and across participants are equal, these two methods produced comparable results. Therefore, we report only the case for aggregation across participants. Conventional estimates from item aggregation may be compared against true participant-level effects, $\mu^{(h)} + \alpha_i^{(h)} - \mu^{(f)} - \alpha_i^{(f)}$.

The independent model was implemented as previously discussed. All additive effects in the Gibbs sampler were initialized with 0; all variances were initialized with 1. Chains were run for 3000 iterations with the first 500 serving as burn-in. Figure 4 provides an assessment of these choices. The upper row shows sample paths for a few parameters ($\mu^{(h)}$, $\alpha_{10}^{(f)}$, and $\sigma^2_{\beta,f}$). The lower row shows autocorrelation for samples after burn-in. Convergence is rapid for the blocked parameters and a bit slower for the variance parameters. The length of the chains and choice of the burn-in period seem appropriate for estimating posterior means, which is the objective in these simulation studies. The simulated data analyzed in Figure 4 had true positive correlation. The displayed convergence is typical for the other relationships between random effects as well.

The correlated model was implemented in a similar fashion. The main difference was for each simulated data set; there were separate analysis for each of three values of ω. The values of burn-in and chain length were the same as with the independent model. Convergence in this model was fairly quick for all three values of ω, especially for the blocked parameters. Figure 5 shows the case when the true random effects have strong true positive correlation and $\omega = 1$.

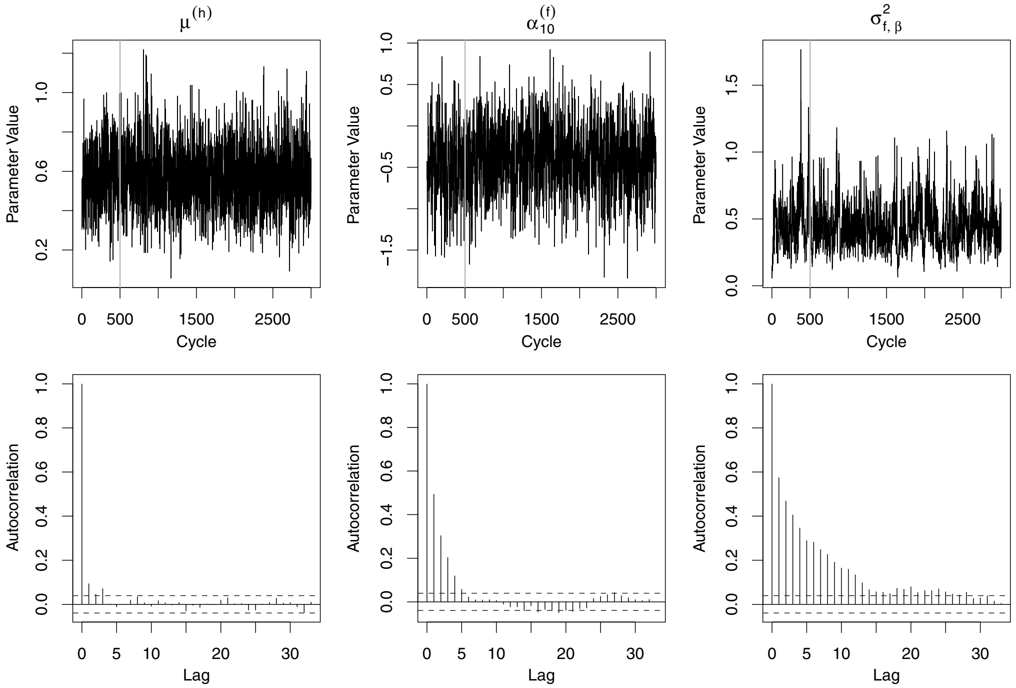

FIGURE 4.
Chain convergence in the independent model with strong positive correlations among both participant and item random effects. The top row shows MCMC values for selected parameters. The vertical line denotes the end of the burn-in period. The bottom row shows autocorrelations of select parameters after burn-in. The speed of convergence is typical for other parameter and other relationships between random effects.

The displayed convergence is typical for the other relationships between random effects and the other values of ω as well. The two right columns show chains for parameters $\sigma^2_{f,\beta}$ and ρ_α. For each cycle, MCMC values of the precision matrices were inverted to yield values of covariance matrices, from which variances and correlations were computed.

Results

Estimation of Sensitivity

One objective is to compare these models with conventional aggregation. To that end, we examined the accuracy of each method's estimates of individual-level d'. For each individual, hierarchical estimates were calculated from posterior means as $\hat{\mu}^{(h)} + \hat{\alpha}_i^{(h)} - \hat{\mu}^{(f)} - \hat{\alpha}_i^{(f)}$. Accuracy of these estimates was assessed by root-mean-square error (RMSE), where the mean is over participants and replicate experiments. These RMSE values are plotted in Figure 6. The five lines correspond to the five different estimation methods (correlated model with three different values of ω, independent model, and aggregation). The values are plotted for each of the nine relationships. There are a number of noteworthy trends: the poor performance of the conventional method; the intermediate performance of the independent model; and the dependency of the correlated model on the choice of ω. Insight into these trends is provided by contour plots of residuals (Figure 7). Each contour plot is derived from a scatter plot of residual sensitivity as a function of true sensitivity. The scatter plots themselves contain 12,000 points (30 participants

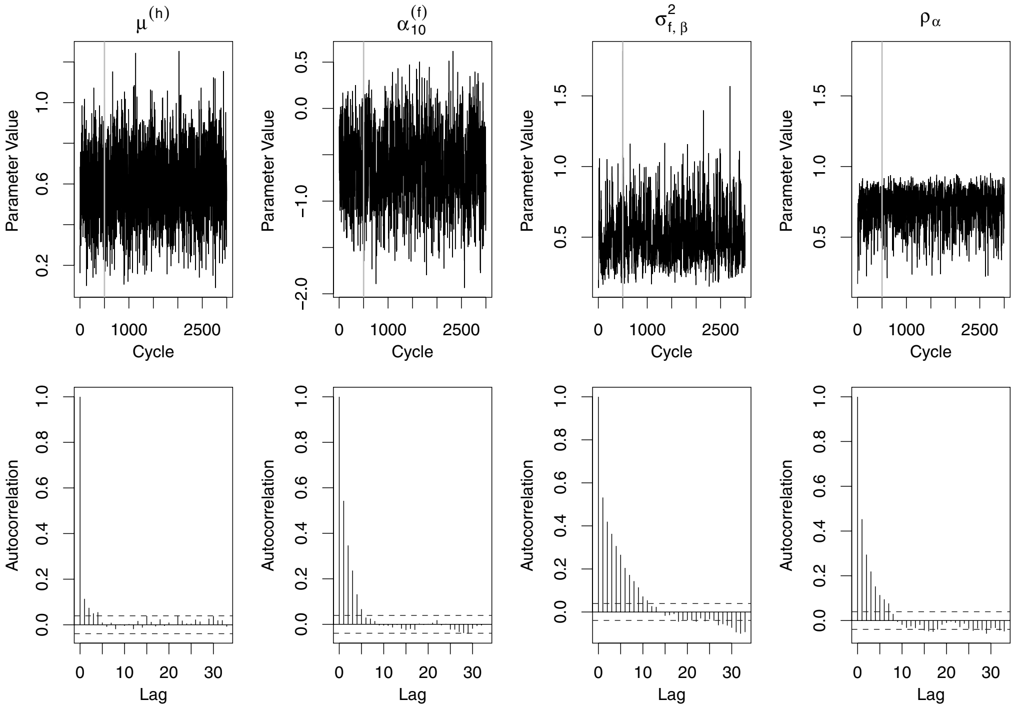

FIGURE 5.
Chain convergence in the correlated model ($\omega = 1$) with strong positive correlations among both participant and item random effects. The top row shows MCMC values for selected parameters. The vertical line denotes the end of the burn-in period. The bottom row shows autocorrelations of select parameters after burn-in. The speed of convergence is typical for other parameters, other values of ω, and other relationships between random effects.

by 400 replications), which is a large number for inspection. The contours are bivariate kernel density estimates of the scatter plot points. Six contour plots are shown. Those on the left, middle, and right are from aggregation, the independent model, and the correlated model ($\omega = 1$), respectively. Those on the top and bottom rows are for the (0, 0) and ($-.9$, $-.9$) relationships.

From the contour plots, it is evident that the poor RMSE performance of the aggregation method is due to the underestimation bias (left column). Underestimation increases linearly with sensitivity: estimated values were about 77% of true values. Contour plots for the two hierarchical models are shown. The independent model (middle column) has improved performance relative to aggregation. The residuals, however, tend toward the diagonal, which indicate an over-shrinkage bias. When the participants' true sensitivity is low in value, the model overestimates sensitivity. Likewise, when the true value is high in value, the model underestimates sensitivity. Although not shown, further analysis reveals that over-shrinkage in sensitivity is the result of over-shrinkage in both $\alpha^{(h)}$ and $\alpha^{(f)}$. Fortunately, over-shrinkage bias is not asymptotic; it reduces with increasing sample size. Therefore, as the sample size increases, the accuracy advantage of the hierarchical model increases over aggregation.

The results with the correlated model depend on the choice of ω (see Figure 6). In the case that $\omega = 10$, the correlated model performed relatively poorly, whereas for $\omega = 1$ and $\omega = .1$, the correlated model performed well. This result is not too surprising. The value of ω scales the prior on the random-effects variance. The true value of random effect variance was set to .5, which is the marginal median of the inverse Wishart prior for $\omega \approx .23$. Not surprising, choices of $\omega = .1$ and $\omega = 1$ give relatively good results whereas the prior with $\omega = 10$ places too much mass away

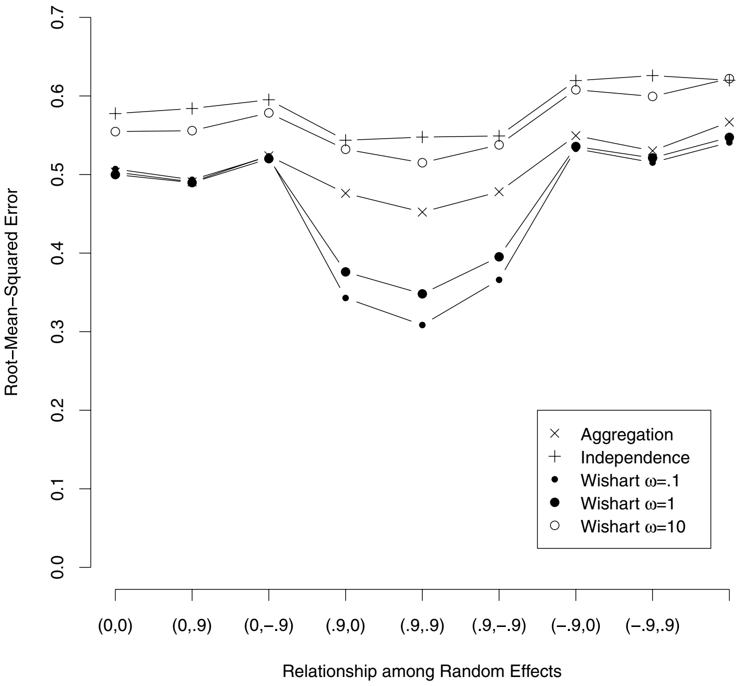

FIGURE 6.
Estimation accuracy as a function of relationship between random effects for five methods (aggregation, independent model, and correlated model with $\omega = (.1, 1, 10)$). The nine relationships are indicated by an ordered pair (i, j) where i denotes the correlation among participants, $i = (0, .9, -.9)$, and j denotes the correlation among items, $j = (0, .9, -.9)$.

from the true value. Figure 7, right panel, shows that over-shrinkage is present for the correlated model when the random effects are independent, but less so when they are negatively correlated.

In Figure 6 the largest difference for the methods occurs when participant effects are positively correlated. In this case, not only does the correlated model with appropriate ω outperform the independent model, the amount of gain decreases with ω. The reason for this behavior, however, is somewhat artifactual. Participant's sensitivity is given by $d'_i = (\mu^{(h)} - \mu^{(f)}) + (\alpha_i^{(h)} - \alpha_i^{(f)})$. Strong positive correlation implies that $\alpha_i^{(h)}$ nearly equals $\alpha_i^{(f)}$, which, in turn, implies that d'_i nearly equals $\mu^{(h)} - \mu^{(f)}$. Hence there is very little variability in participants' true sensitivity. Priors that bias the variance of the participant random effects toward zero also reduce the variance of $\alpha_i^{(h)} - \alpha_i^{(f)}$. This downward bias in the variance of $\alpha_i^{(h)} - \alpha_i^{(f)}$ leads to better estimates of d'_i. Although priors with high values of ω lead to better estimation of d'_i in this case, they may lead to poor estimation of $\alpha_i^{(h)}$ and $\alpha_i^{(f)}$ themselves.

Variability and Correlation among Random Effects

In this section we assess the estimation of the variability and correlation among random effects. Both hierarchical models have random effect variance parameters. For the independent model these parameters are $(\sigma^2_{\alpha,h}, \sigma^2_{\alpha,f}, \sigma^2_{\beta,h}, \sigma^2_{\beta,f})$. For the correlated model, these parameters are the diagonal elements of $\mathbf{\Sigma}_\alpha$ and $\mathbf{\Sigma}_\beta$. In the simulations, the true variance of all random effects was set to .5. Figure 8, left panel, shows the average posterior mean of participant variance parameters. These estimates are not ideal. Those from the independent model overestimate the true variance value by about 10%; the estimates for the correlated model are affected by the choice of ω, and in a reasonable way. Whereas ω denotes prior scale on variance, it is expected

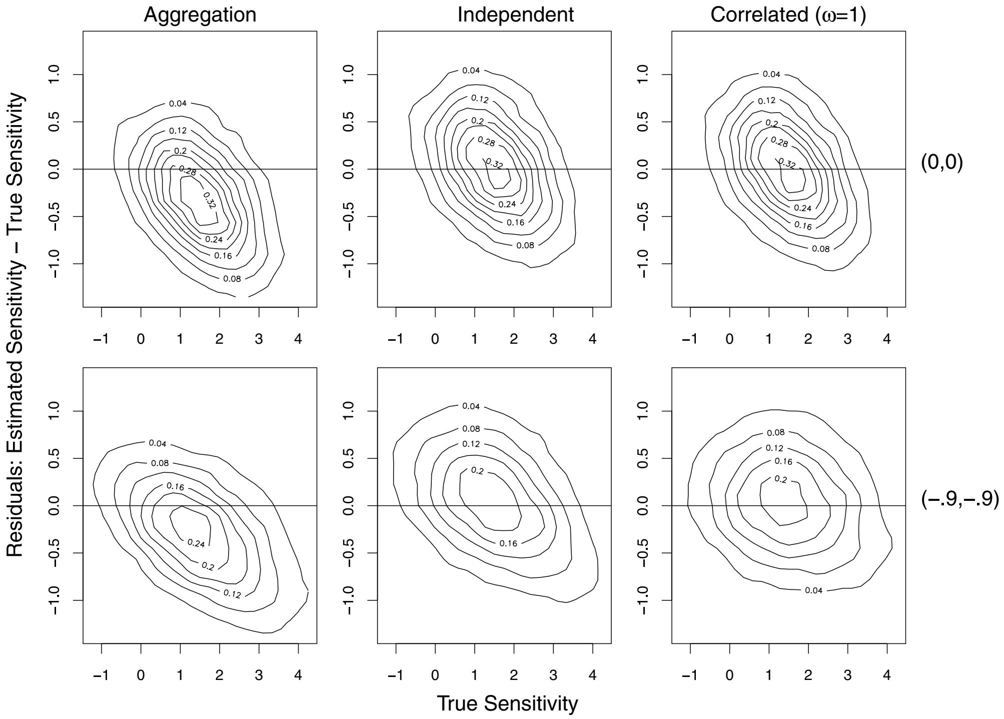

FIGURE 7.
Residuals for three methods and two relationships. Left, middle, and right columns show residuals from aggregation, the independent model, and the correlated model ($\omega = 1$), respectively. Top and bottom rows show residuals for relationships $(0, 0)$ and $(-.9, -.9)$, respectively. Contour plots are derived from scatter plots of residuals (estimated sensitivity–true sensitivity) as a function of true sensitivity.

that large values of ω lead to an overestimation of variance. Figure 8 shows estimates for two relationships and these are typical of the other relationships as well.

The right panel shows estimates for correlation. Population-level estimates of correlation are only provided by the correlated model. Overall, all methods revealed very low correlations for the $(0, 0)$ relationship and more extreme correlations in the appropriate directions for the $(.9, -.9)$ relationship. Estimated correlation also varied with ω. Estimates of correlation were biased toward zero for larger values of ω and closer to true values for smaller values.

Analysis of a Data Set

We compared the hierarchical Bayesian model and aggregation estimates in a reasonably sized recognition memory experiment. The main goal of the experiment was to assess the effect of attention on memory retrieval. Participants were asked to divide their attention between the recognition memory task and a secondary task. We manipulated the degree of attention on each task through instructions. For some items, participants were instructed to devote all their attention to the memory task; for other items, attention was to be divided either 80% to 20%; 50% to 50%, or 20% to 80% across the memory and secondary tasks. Our concern is establishing whether there is an effect of the attention instructions on recognition memory and, if so, establishing whether the effect is on hits or false alarms.

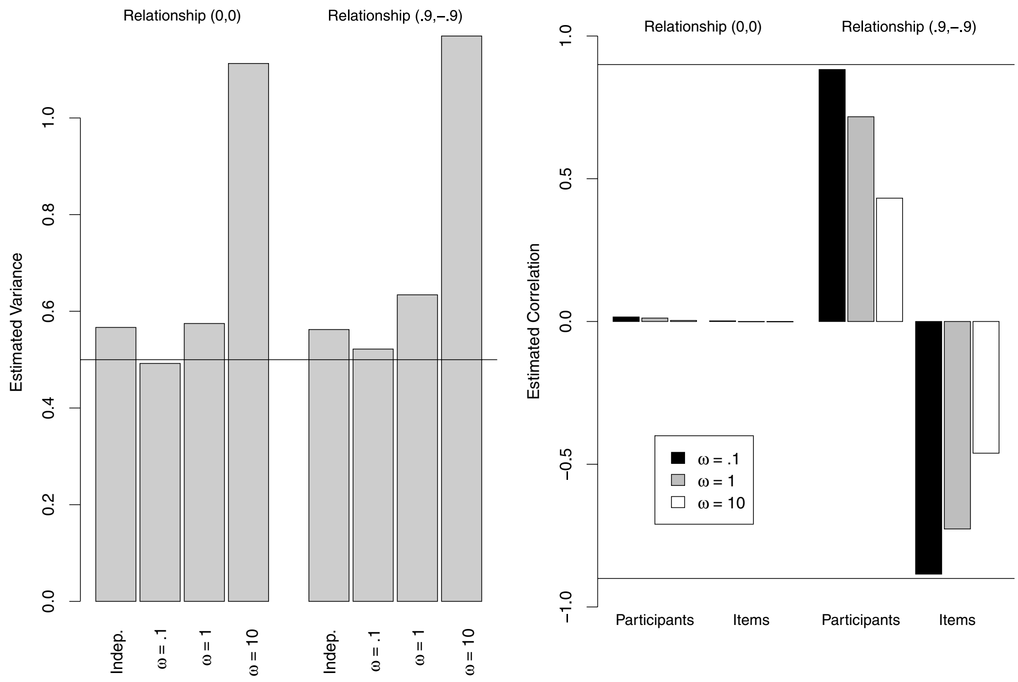

FIGURE 8.
Estimated variances (left) and correlations (right) of random effects in the hierarchical models. See the text for details.

Details of the experiment are as follows: Thirty-four participants were tested on 104 old and 104 new items. The attention instruction was manipulated within subjects but was blocked within the list; i.e., participants first studied a set and were tested at one attention level, then they studied a second set and were tested at a second attention level, and so on. There was one additional independent manipulation: the type of secondary task. Half of the participants indicated the location of an asterisk on the screen as a secondary task. The asterisk was presented at one of three locations and participants had to respond by depressing one of three keys below their right hand. As soon as the response to the asterisk was made, the asterisk shifted to a new location prompting a new response. The secondary task was easy and straightforward and participants were able to perform it simultaneously with the memory recognition task. In the recognition task, participants indicated whether target items were old or new by depressing one of two keys below their left hand. The other half of the participants performed a tone-identification secondary task. One of three tones (high pitched, medium pitched, or low pitched) was presented and participants depressed a corresponding key with their right hand while indicating their memory recognition response with their left hand. Crossing the two independent manipulations (four instructions × two secondary tasks) yields eight distinct conditions. We expanded the hierarchical Bayesian model to account for these eight conditions:

$$h_{ijk} = \mu_k^{(h)} + \alpha_i^{(h)} + \beta_j^{(h)},$$

$$f_{ijk} = \mu_k^{(f)} + \alpha_i^{(f)} + \beta_j^{(f)},$$

where $i = 1, \dots, 34$ indexes participants, $j = 1, \dots, 208$ indexes items, and $k = 1, \dots, 8$ indexes conditions.

Figure 9 shows the effect of condition on sensitivity d'. The darkest bars are from the correlated model ($\omega = 1$, run length of 10,000 iterations, burn-in of 1000 iterations). Two con-

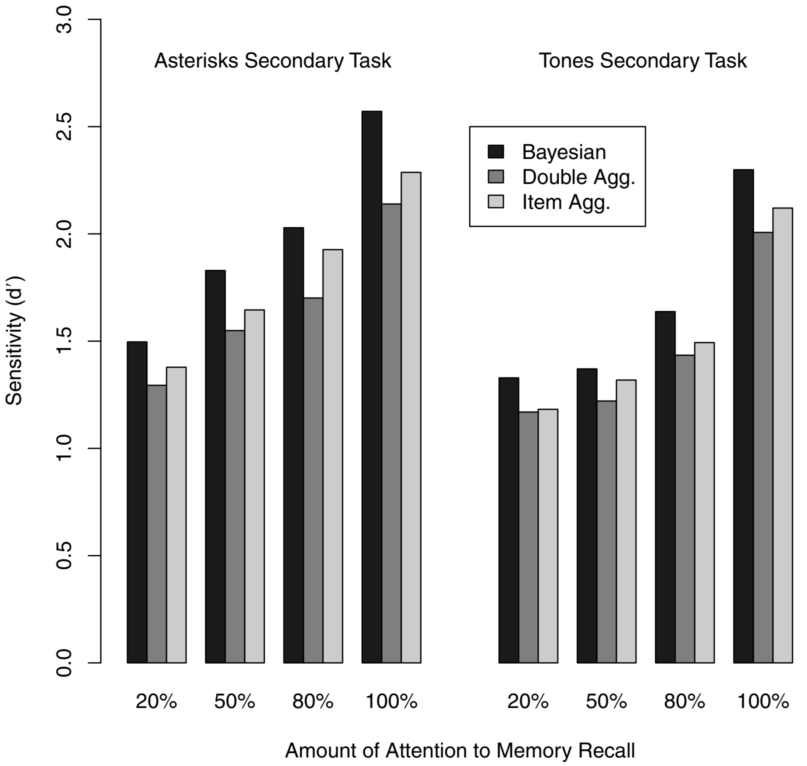

FIGURE 9.
Estimates of condition sensitivity from the correlated model ($\omega = 1$) and two aggregation methods.

ventional aggregation measures are also shown. The grey bars show a *double aggregation* measure in which grand aggregate hit and false alarm rates were computed for each condition. Sensitivity was estimated for each condition from these aggregates according to (3). The light bars show an *item aggregation* measure in which aggregate hit and false alarm rates were computed for each participant-by-condition pairing. Sensitivity was estimated for each of these pairings according to (3) after half a count was added to all event frequencies (see Footnote 1). These sensitivities were averaged across participants to produce condition-specific sensitivity estimates. The aggregated estimates are between 10% and 20% less in value than the comparable Bayesian estimates. The decrease reflects asymptotic bias from aggregation. The different estimation techniques all reveal a main effect for the divided-attention manipulation.

Figure 10 shows condition effects on hits ($\mu^{(h)}$) as a function of that on false alarms ($\mu^{(f)}$). The open and closed points are from asterisk and tone secondary tasks, respectively. It is evident from the figure that the mnemonic increase with attention is attributable to both an increase in hit rate and a decrease in false alarm rate; i.e., a mirror effect. From a psychological perspective, the results indicate that participants adjust their criterion based on the amount of attention deployed. As attention is increased, sensitivity (d') increases, and so does the criterion.

A second set of theoretical questions concerns correlations of participant and item effects. Figure 11, top-left, shows the individuals' effects on hits ($\alpha^{(h)}$) as a function of that on false alarms ($\alpha^{(f)}$). The sample correlation for these points is $-.37$, indicating the possibility of a participant-based mirror effect. To provide a test of this possibility, we plotted the posterior of the correlation coefficient from $\boldsymbol{\Sigma}_\alpha$ as estimated from the MCMC outputs (bottom-left). The dotted vertical lines indicate the 95% credible interval on the posterior. Because the value of zero is within

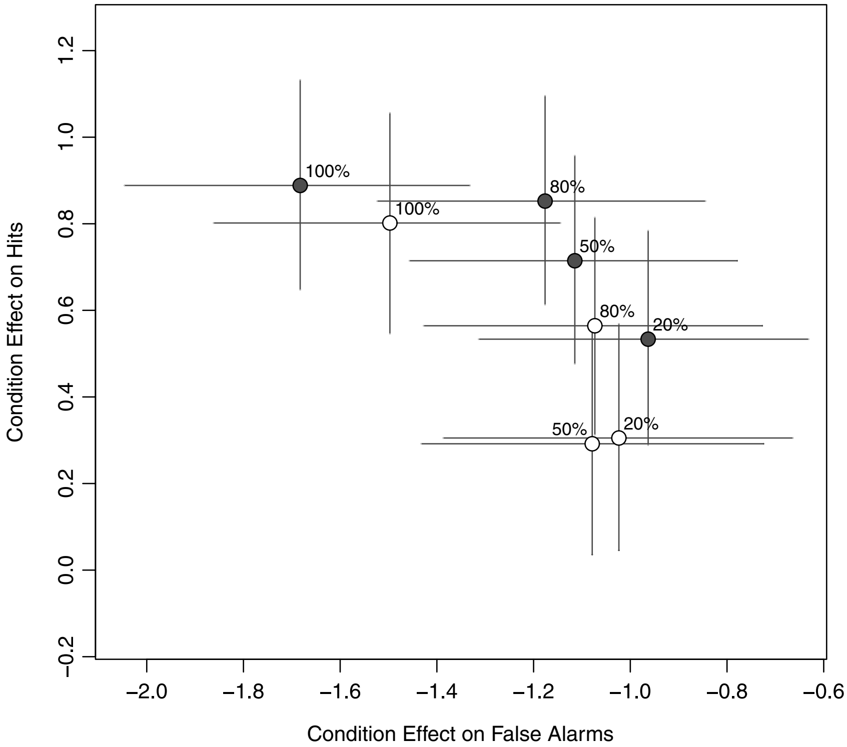

FIGURE 10.
Effect of condition on hit ($\mu^{(h)}$) as a function of effect on false alarm ($\mu^{(f)}$). Error bars denote 95% credible intervals. Filled and open circles denote effects of instruction in the asterisk and tone secondary tasks, respectively. There is a mirror effect across the attention levels.

this interval, the results do not provide sufficient evidence to generalize this participant-based mirror effect to the population. The right column of Figure 11 provides analogous information about item effects. The sample correlation is $-.22$; the 95% credible interval on the posterior of the correlation coefficient from $\mathbf{\Sigma}_\beta$ includes the value of zero indicating insufficient evidence to generalize this correlation.

We have also analyzed the data with $\omega = .1$ and $\omega = 10$. There were no qualitative differences with these choices from the reported results in Figures 9 to 11.

General Discussion

In this paper we have presented two Bayesian hierarchical models for the analysis of recognition memory data with the theory of signal detection. These models offer significant practical and theoretical advantages over the conventional technique of aggregation. The most salient of these is accurate estimation of participant and item random effects. Accurate estimation of participant effects is desirable because it is of use in individual-difference studies. Studying the relationship between mnemonic ability and other cognitive variables often leads to better theories of cognition and its variability across populations (e.g., Kane et al., 2004; Salthouse, 1996). Accurate estimation of item effects is desirable because it leads to better theories of memory and word processing (e.g., Gillund & Shiffrin, 1984; Spieler & Balota, 1997).

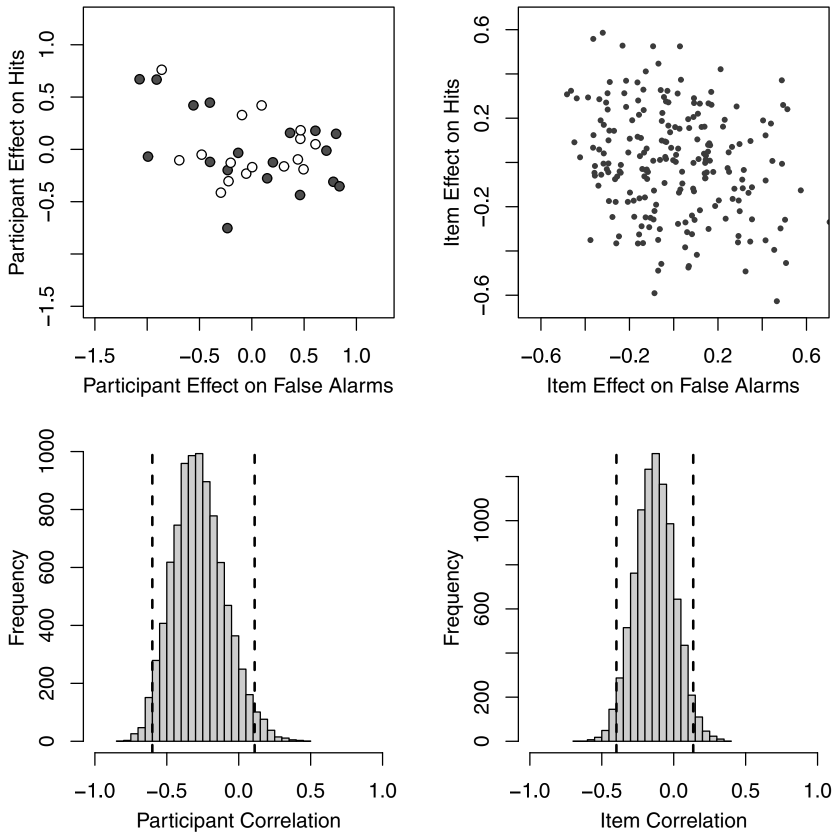

FIGURE 11.

Participant and item effects. Top-left: Scatter plot of individuals' effects on hits ($\alpha^{(h)}$) as a function of that on false alarms ($\alpha^{(f)}$). Filled and open circles are from the asterisk and tone secondary tasks, respectively. Bottom-left: Posterior distribution of correlation coefficient from $\mathbf{\Sigma}_\alpha$. The dotted vertical lines indicate the 95% credible interval on correlation. Right column: Analogous plots for item effects.

The correlated model seems superior to the independent model if one has a reasonable idea of the scale of random effects a priori. The following argument may be useful as a rough guide: In order to gain statistical power, researchers typically design experiments such that hit and false alarm rates are not too extreme. True false alarms are, hopefully, greater than .05 while true hit rates are, hopefully, less than .95. These constraints limit true d' to about 3. Within this context, it seems unreasonable that the participant and item variation would be too great. A maximal upper limit on the standard deviation of random effects may be around 2. The analysis of the our data set indicated that it is far less. Appropriate choices of ω range between $\omega = .1$ and $\omega = 1.5$.

Even though these hierarchical models offer a dramatic improvement over conventional techniques in the field, they are not fully satisfactory. The independent model is biased against correlation in random effects. The correlated model must be specified with prior information about the scale of random effects. Gross misspecification will lead to poor performance. Hopefully, advances in Bayesian hierarchical models with normally distributed priors will provide for more suitable priors (cf., Sun & Berger, 1998, 2006). Until then, the two models serve as reasonable alternatives.

The models presented here are termed *equal-variance* signal detection models because the variance of the new-item distribution is the same as that of the old-item distribution (see Figure 1).

The equal-variance model provides a convenient measure of mnemonic performance. Fortunately, it may be tested by assessing confidence, varying payoffs, or varying stimulus probabilities. In studies with these manipulations, the standard deviation for old items is estimated to be 1.25 as large as that for new items (Glanzer, Kim, Hilford, & Adams, 1999; Heatcote, 2003; Ratcliff, Sheu, & Grondlund, 1992). The hierarchical signal detection models can be generalized to reflect arbitrary variance for the old-item distribution. A straightforward approach is to introduce a scale parameter to the standard normal distribution function in (1). Hit and false alarm probabilities are

$$H = \Phi\left(\frac{d' - c}{\sigma}\right), \tag{13}$$

$$F = \Phi(-c). \tag{14}$$

The rest of the hierarchical Bayesian model can be specified analogously. Moreover, the latent variables and conditional posterior distributions need only minor modifications for analysis. We provide these modifications in the Appendix.

Appendix

For the *unequal-variance* signal detection model, H and F are given in (13) and (14), respectively, where σ is the standard deviation of the old-item distribution. There will be no changes on the prior distributions in the hierarchical model.

The following latent variables are convenient in analysis of the *unequal-variance* model. Let indicator variable $D_k = 1$ if a signal is presented, and $D_k = 0$ otherwise. The distribution of a latent variable is

$$w_k \overset{\text{indep}}{\sim} \text{Normal}([\boldsymbol{X\lambda}]_k, \sigma_k^2),$$

where

$$\sigma_k^2 = \begin{cases} \sigma^2, & D_k = 1, \\ 1, & D_k = 0. \end{cases} \tag{15}$$

The prior on σ^2 is $\sigma^2 \sim$ Inverse Gamma(a_0, b_0). Hence, for Bayes computation, the full conditional posterior distribution of σ^2 is

$$\sigma^2 \mid \boldsymbol{\lambda}, \boldsymbol{w} \sim \text{Inverse Gamma}\left(\sum_{k=1}^{IJ} D_k/2 + a_0, \frac{1}{2}\sum_{k=1}^{IJ} D_k(w_k - [\boldsymbol{X\lambda}]_k)^2 + b_0\right).$$

The conditional distributions for other parameters are derived as follows: Fact B and Fact E hold for both *equal-variance* or *unequal-variance* models without modification. Facts A, C, and D need only minor modification and are presented below as Fact A.2, Fact C.2, and Fact D.2. Proofs are straightforward.

- *Fact* A.2. Generalization of Fact A. Let $\boldsymbol{\Sigma}_w$ denote a diagonal matrix with elements σ_k^2, where σ_k^2 is defined in (15). The full conditional distribution of $\boldsymbol{\lambda}$ is

$$\boldsymbol{\lambda} \mid \boldsymbol{\Sigma}_w, \sigma_{\alpha,h}^2, \sigma_{\alpha,f}^2, \sigma_{\beta,h}^2, \sigma_{\beta,f}^2, \boldsymbol{w} \sim N_q(\boldsymbol{V}(\boldsymbol{X}^t\boldsymbol{\Sigma}_w^{-1}\boldsymbol{w}), \boldsymbol{V}),$$

where $q = 2(1 + I + J)$ and $\boldsymbol{V} = (\boldsymbol{X}^t\boldsymbol{\Sigma}_w^{-1}\boldsymbol{X} + \boldsymbol{B})^{-1}$.

- *Fact* C.2. This is a generalization for Fact C. The distribution of $(\boldsymbol{w} \mid \boldsymbol{\lambda}; \boldsymbol{y})$ may be given in terms of components $w_1, \dots, w_{IJ}$:

$$w_k \mid \boldsymbol{\lambda}; \mathbf{y} \overset{\text{indep}}{\sim} \begin{cases} N^-([\boldsymbol{X\lambda}]_k, \sigma_k^2), \ y_k = 0, \\ N^+([\boldsymbol{X\lambda}]_k, \sigma_k^2), \ y_k = 1, \end{cases}$$

  where N^- and N^+ denote normal distributions truncated at 0 from above and below, respectively, and σ_k^2 is defined in (15).

- *Fact* D.2. With the notation in Fact D, the full conditional distribution of $\boldsymbol{\lambda}$ is

$$\boldsymbol{\lambda} \mid \boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_\alpha^{-1}, \boldsymbol{\Sigma}_\beta^{-1}, \boldsymbol{w} \sim N_q(\boldsymbol{V X}^t \boldsymbol{\Sigma}_w^{-1} \boldsymbol{w}, \boldsymbol{V}_c),$$

  where $q = 2(1 + I + J)$ and $\boldsymbol{V}_c = (\boldsymbol{X}^t \boldsymbol{\Sigma}_w^{-1} \boldsymbol{X} + \boldsymbol{B}_c)^{-1}$.

References

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

Browne, W.J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computer Statistics*, *15*, 391–420.

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

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer-Verlag.

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

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

Gillund, G., & Shiffrin, R. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 97–123.

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

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

Hautus, M.J., & Lee, A.L. (1998). The dispersions of estimates of sensitivity obtained from four psychophysical procedures: Implications for experimental design. *Perception & Psychophysics*, *60*, 638–649.

Heathcote, A. (2003). Item recognition memory and the receiver-operating characteristic. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 1210–1230.

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.

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

Jeffreys, H. (1982). *Theory of probability.* New York: Oxford University Press.

Kane, M.J., Hambrick, D.Z., Tuholski, S.W., Wilhelm, O., Payne, T.W., & Engle, R.E. (2004). The generality of working-memory capacity: A latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*, 189–217.

Kucera, H., & Francis, W.N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Lu, J., Speckman, P.L., Sun, D., & Rouder, J. (submitted). *An objective Bayesian approach for assessing conscious and unconscious processes in human memory.*

Macmillan, N.A., & Creelman, C.D. (2001). *Detection theory: A user's guide.* New York: Cambridge University Press.

McClelland, J.L., & Chappell, M. (1994). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *101*, 103–128.

Miller, E., & Lewis, P. (1977). Recognition memory in elderly patients with depression and dementia: A signal detection analysis. *Journal of Abnormal Psychology*, *86*, 84–86.

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

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

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

Rouder, J.N., Lu, J., Morey, R.D., Sun, D., & Speckman, P.L. (submitted). *A hierarchical process dissociation model.*

Salthouse, T.A. (1996). The processing speed theory of adult age differences in cognition. *Psychological Review*, *103*, 403–428.

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

Singer, M., Gagnon, N., & Richard, E. (2002). Strategies of text retrieval: A criterion shift account. *Canadian Journal of Experimental Psychology*, *56*, 41–57.

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

Spieler, D.H., & Balota, D.A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, *8*, 411–416.

Stretch, V., & Wixted, J.T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 1379–1396.

Sun, D., & Berger, J. (1998). Reference priors under partial information. *Biometrika*, *85*, 55–71.

Sun, D., & Berger, J.O. (2006). Objective priors for the multivariate normal model. *Bayesian Statistics*, *8*.

Tanner, J.W.P., & Birdsall, T.G. (1958). Definition of *d'* and *n* as psychophysical measures. *Journal of the Acoustical Society of America*, *30*, 922–928.

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.

Manuscript received 8 JUL 2005
Final version received 10 JUL 2006
Published Online Date: 19 MAR 2007