

Separating Mnemonic Process From Participant and Item Effects in the Assessment of ROC Asymmetries

Michael S. Pratte and Jeffrey N. Rouder
University of Missouri

Richard D. Morey
University of Groningen

One of the most influential findings in the study of recognition memory is that receiver operating characteristic (ROC) curves are asymmetric about the negative diagonal. This result has led to the rejection of the equal-variance signal detection model of recognition memory and has provided motivation for more complex models, such as the unequal-variance signal detection and dual-process models. Here, the authors test the possibility that previous demonstrations of ROC asymmetry do not reflect mnemonic process but rather reflect distortions due to averaging data over items. Application of a hierarchical unequal-variance signal detection model reveals that asymmetries are in fact a real phenomenon and do not reflect distortions from averaging data.

Keywords: recognition memory, ROC curves, ROC asymmetry, signal detection, hierarchical models

Although the field of memory is evolving rapidly, it remains characterized by several long-standing debates. One of these is whether memory is served by one or several distinct systems or processes (Schacter & Tulving, 1994). The recognition memory paradigm, in which participants decide if targets were previously studied or not, has provided key findings and fueled vigorous exchanges (e.g., Wixted, 2007; Yonelinas & Parks, 2007). In recognition memory experiments, receiver operating characteristic (ROC) curves are often used to summarize results. One of the most influential findings for current theory development is that ROC curves are asymmetric around the negative diagonal. An asymmetric ROC curve typical of recognition memory data is shown as the dashed line in Figure 1B. A symmetric ROC, in contrast, is shown as the solid line. The asymmetry in the dashed line has been replicated repeatedly (see Glanzer, Kim, Hilford, & Adams, 1999; Yonelinas & Parks, 2007) and has served as a first-order phenomenon to be explained by mnemonic theory (e.g., Ratcliff, Sheu, & Gronlund, 1992).

When considering the role of asymmetry in mnemonic theory, it is convenient to start with the theory of signal detection (Green & Swets, 1966/1974) for recognition memory (Kintsch, 1967). At test, participants assess the mnemonic strength of an item. If the strength is above criterion, the participant judges the item as studied; otherwise, the participant judges the item as new. In the simplest version of the model, strengths are distributed as equal-variance normals, and the effect of study is to shift the distribution.

This simple model, however, predicts symmetric ROCs. There are a number of modifications that predict asymmetries, the most common of which is to discard the equal-variance assumption. The resulting model is called the *unequal variance signal detection model* (UVSD). In UVSD, study affects both the mean (d') and the variance (σ^2) of the normally distributed strengths. The parameter σ has a one-to-one relationship with the degree of ROC asymmetry, with larger values of σ resulting in a larger degree of asymmetry.

ROC asymmetries have inspired a plethora of theoretical accounts. The dual-process model (Yonelinas, 1994), for example, posits that recognition memory is mediated by either familiarity or recollection. Familiarity is modeled as an equal-variance signal detection process; recollection is modeled as an all-or-none discrete process. The degree of asymmetry is related to the amount of recollection. Other theoretical accounts of asymmetry include mixture models (e.g., DeCarlo, 2002) and global memory models, such as REM (Shiffrin & Steyvers, 1997) and TODAM (Murdock, 1993). In sum, almost all researchers view the asymmetries in ROCs as reflecting characteristics of the mnemonic system. We provide and assess an alternative account that these asymmetries reflect a flaw in current practices of data analysis rather than a characteristic of mnemonic processing.

Separating Process From Participant and Item Effects

In recognition memory experiments, each participant provides a single response to each item. This response is not sufficient for computing the response proportions necessary to construct ROCs. Hence, researchers average data across participants, items, or both. To understand the effects of this averaging, one needs to distinguish between latent mnemonic processes, on one hand, and what we term *surface variables* on the other. Surface variables refers to effects that are under much experimental control, such as the selection of participants, items, and number of items between study and test (*lag*). Latent mnemonic processes, roughly speaking, refers to commonalities in the structure of encoding, mainte-

Michael S. Pratte and Jeffrey N. Rouder, Cognition and Neuroscience, University of Missouri; Richard D. Morey, Psychometrics and Statistical Techniques, University of Groningen, Groningen, the Netherlands.

This research was supported by National Science Foundation Grant SES-0720229 and by National Institute of Mental Health Grant R01-MH071418. We thank Kathleen Wenzlick, Patrick Lockwood, and Robert Paterson for help in data collection.

Correspondence concerning this article should be addressed to Michael S. Pratte, 210 McAlester Hall, University of Missouri, Columbia, MO 65211. E-mail: prattems@gmail.com

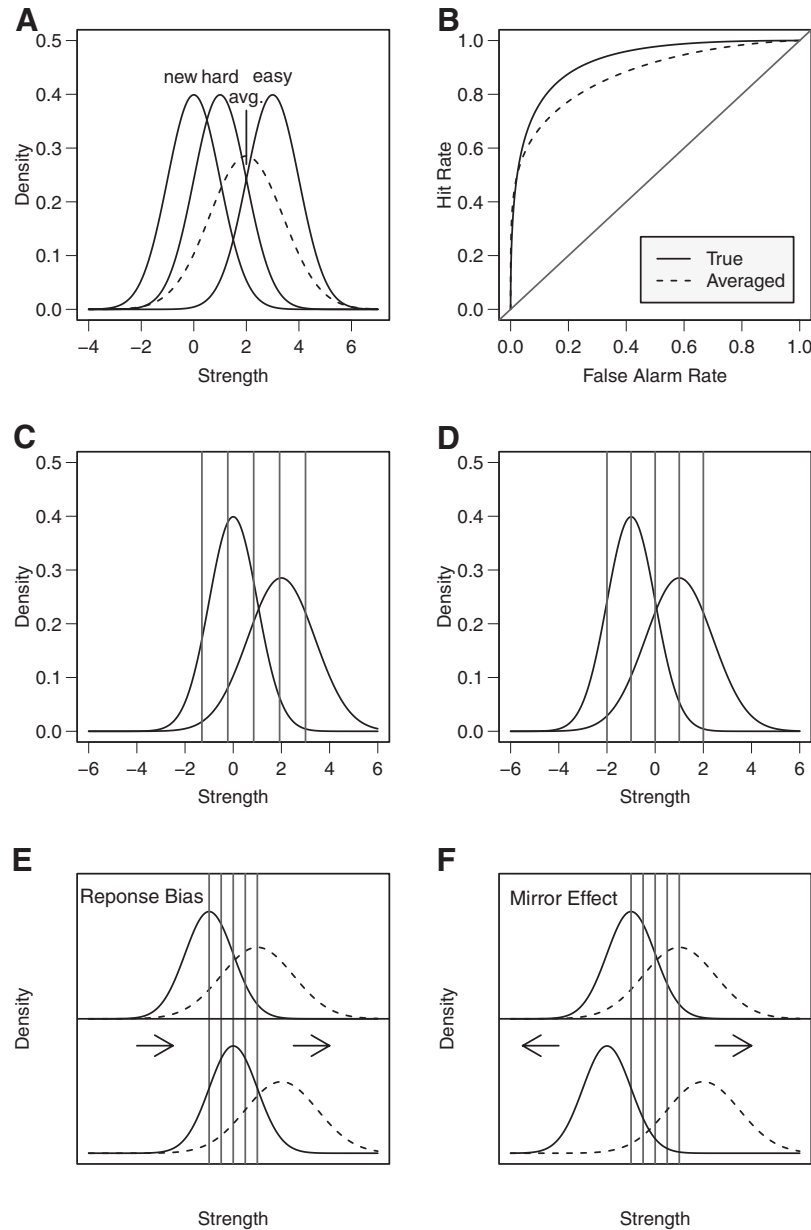


Figure 1. Signal detection models and ROC curves. A: The solid lines denote an equal-variance signal detection model for the example with hard items (low d') and easy items (high d'). When rates are averaged, the result is the dotted-line distribution. B: The solid curve is a symmetric ROC curve from an equal-variance signal detection model with $d' = 2$. The dashed ROC curve is an asymmetric curve that results from averaging across easy and hard items. Averaging introduces an asymmetry in the curve. C: The traditional UVSD parameterization in which the mean and variance of the new-item distribution are fixed to 0.0 and 1.0, respectively. D: The new parameterization in which the middle criterion is fixed to 0.0 and the variance of the new-item distribution is fixed to 1.0. In this parameterization, the means of both the new-item ($d^{(n)}$) and studied-item distributions ($d^{(s)}$) are free. E: Bias effects are a concurrent increase in both distributions. F: Mirror effects are a concurrent decrease in the new-item distribution and an increase in the studied-item distribution.

nance, retrieval, and the evaluation of information. For instance, the theory of signal detection posits that all participants evaluate the strength of the test item relative to criteria. Processing may differ across surface variables (e.g., participants may vary in their sensitivity and criteria), but, nonetheless, there is a commonality of

evaluation of strength. Most cognitive theories are about latent mnemonic process. For example, in Yonelinas' dual-process model, recollection and familiarity are latent processes. The key distinction between surface and latent variables is that surface variation is mapped onto particular elements in the design (a

specific person, item, or lag) and therefore can be parceled out. Latent variation, including trial-to-trial variability from, say, attentional fluctuation or criterial variation, cannot be mapped to such elements and is considered part of the latent mnemonic process.

Separating surface variables from latent process is difficult in recognition memory experiments. Each participant is tested only once on each item and only as either new or studied. The result is that every response in a recognition memory test arises from a unique combination of a participant, an item, a study condition (new vs. studied), and a lag if the item was studied. Moreover, only half of the potential data are observed, as an item cannot be both new and studied for a given person. If data are averaged over people or items, participant or item variability is confounded with mnemonic process. In some cases, this averaging has been shown to have serious ramifications in interpretation of memory experiments (Curran & Hintzman, 1995; Hintzman, 1980; Hintzman & Hartry, 1990). In the next section we demonstrate how averaging data over surface variables may distort the estimation of latent process parameters in ROC analysis (see also Morey, Pratte, & Rouder, 2008; Rouder & Lu, 2005).

The Consequences of Averaging

We provide an example of how item variability distorts measurement of ROC asymmetry. Though the example is of item variability, it extends to other surface variables. We start with an equal-variance signal detection model. Assume that half of the items are easy ($d' = 3$) and half are hard ($d' = 1$; see solid lines in Figure 1A). From this model, true hit and false alarm rates are computed for each item across several criteria (ranging from -1 to 3). These rates are then averaged over the easy and hard items. We then construct the resulting z ROC curve, which provides an estimate of parameter σ in the UVSD model. If averaging has no ill effects this analysis should yield $\sigma = 1$, as the rates for each item were generated from equal-variance models.

The true ROC curve for an equal-variance signal detection model with $d' = 2$ (the average of 1.0 and 3.0) is shown as the solid line in Figure 1B. The ROC curve generated from hit and false alarm rates averaged over easy and hard items is shown as the dashed line in Figure 1B. The resulting ROC is asymmetric even though the data come from an equal-variance signal detection model. This deviation from symmetry is consistent with a UVSD model in which $\sigma = 1.39$ or a dual-process model in which the probability of recollection is 24%. These values are distortions, the cause of which is shown in Figure 1A. If data are averaged over easy and hard items (solid lines), the resulting distribution is too wide (dashed line) because of confounding item variability.

The example shows that it is critical to separate mnemonic process from surface variability when interpreting ROC asymmetry. If asymmetry simply reflects unaccounted surface variability, then all models that treat it as a characteristic of the mnemonic system (including UVSD, dual-process, and mixture accounts) are misspecified. Unfortunately, this separation has not previously occurred. Some researchers have performed individual-item analyses while averaging over participants; others have performed individual-participant analyses while averaging over items (e.g., Heathcote, 2003; Heathcote, Raymond, & Dunn, 2006). The former analyses are subject to distortions from participant vari-

ability; the latter are subject to distortions from item variability. The key point is that surface variables need to be accounted for simultaneously if psychological process is to be assessed accurately.

Hierarchical Models of ROC Asymmetry

Our main goal is to assess the source of ROC asymmetry—it may reflect either surface variability or a characteristic of the mnemonic system. Clearly, we cannot draw ROC curves for each participant-by-item combination, as these combinations are unreplicated. Instead, we take a hierarchical modeling approach in which both latent psychological process and surface variables are explicitly modeled. The model developed herein (and discussed in greater detail in Morey, Pratte, and Rouder, 2008) is a hierarchical extension of UVSD.

We use UVSD as a psychometric model for measuring ROC asymmetry. The key parameter is σ ; if ROCs are symmetric, then $\sigma = 1.0$. As asymmetry in ROCs increases, σ increases above 1.0. Our use of UVSD is motivated by a number of factors: First, UVSD was the initial, and remains the most straightforward, generalization of signal detection theory made to account for asymmetry. Second, UVSD has repeatedly been shown to provide a good fit to averaged data (e.g., Glanzer et al., 1999; Heathcote, 2003; Slotnick & Dodson, 2005; Wixted, 2007) and thus provides a logical starting point for fitting nonaveraged data. Third, providing a hierarchical extension of UVSD is both feasible and convenient, as shown by Morey, Pratte, and Rouder (2008) and Rouder et al. (2007). In theory, it is possible that other models, such as Yonelinas' dual-process model and DeCarlo's mixture model, may be extended hierarchically to measure asymmetry. In the former, the estimate of recollection serves as the index of asymmetry; in the latter, the estimate of mixing probability serves as the index. Hierarchical extensions of these other models have not been developed to date.

Model Development

When constructing a hierarchical version of UVSD, it is convenient to adopt an alternative parameterization. Figure 1C shows the traditional parameterization in which the mean and variance of the new-item distribution are fixed to 0.0 and 1.0, respectively. Figure 1D shows the alternative parameterization. The means of the new- and studied-item distributions are free parameters denoted by $d^{(n)}$ and $d^{(s)}$, respectively. The middle criterion is fixed at 0.0 to make the model identifiable. As in the conventional parameterization, the variance of the new-item distribution is fixed at 1.0. This parameterization does not change the model substantively. It is, however, convenient for modeling participant and item effects.

To account for participant and item effects, we allow separate mean strengths for each participant-by-item combination for both the new- and studied-item distributions. Let $d_{ij}^{(n)}$ denote the mean of the distribution for the i th participant tested on the j th item when this j th item is new, $i = 1, \dots, I, j = 1, \dots, J$. Let $d_{ijk}^{(s)}$ denote the same when the j th item is studied. We include the subscript k to indicate the level of lag ($k = 1, \dots, K$) between study and test.

Without participant-by-item replicates, it is not possible to estimate all of these means without restrictions. We therefore assume an additive model for surface effects:

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

$$d_{ijk}^{(s)} = \mu^{(s)} + \alpha_i^{(s)} + \beta_j^{(s)} + \gamma(L_k - L_0),$$

where $\mu^{(n)}$ and $\mu^{(s)}$ are grand means; $\alpha_i^{(n)}$ and $\alpha_i^{(s)}$ are participant effects; $\beta_j^{(n)}$ and $\beta_j^{(s)}$ are item effects; γ is the slope of the lag effect; L_k is the k th lag; and L_0 is the mean lag across the experiment such that the sum of $(L_k - L_0)$ is zero. Participant and item effects are treated as random effects and are modeled as zero-centered normal distributions:

$$\alpha_i^{(n)} \stackrel{iid}{\sim} \text{Normal}(0, \sigma_{\alpha,n}^2),$$

$$\alpha_i^{(s)} \stackrel{iid}{\sim} \text{Normal}(0, \sigma_{\alpha,s}^2),$$

$$\beta_j^{(n)} \stackrel{iid}{\sim} \text{Normal}(0, \sigma_{\beta,n}^2),$$

$$\beta_j^{(s)} \stackrel{iid}{\sim} \text{Normal}(0, \sigma_{\beta,s}^2).$$

The variances of these distributions are estimated from the data, and they provide measures of the magnitude of participant and item variability. We have used this additive structure to account for participant and item effects in two-alternative forced-choice recognition memory (Rouder et al., 2007), stem completion (Rouder, Lu, Morey, Sun, & Speckman, 2008), subliminal perception (Morey, Rouder, & Speckman, 2008), and lexical decision (Rouder, Tuerlinckx, Speckman, Lu, & Gomez, 2008). The additive model is similar to that used in repeated-measures analysis of variance in which interactions between participants and items are treated as errors. Criteria are assumed to vary across participants but not items. These criteria parameters reflect participants' relative preference for certain responses. A single value of σ is estimated for all participants and items.

In the reparameterized UVSD model, sensitivity for each participant-by-item-by-lag combination (denoted d'_{ijk}) is the distance between the distributions ($d_{ijk}^{(s)} - d_{ijk}^{(n)}$). Note that sensitivity is measured in units of the new-item standard deviation. Biases are correlated shifts in the distributions. For example, shifting both distributions to the right (see Figure 1E) results in an increased hit and false alarm rate (i.e., a bias to respond "studied"). Participant or item variation in bias will result in a positive correlation between $\alpha_i^{(n)}$ and $\alpha_j^{(s)}$, or between $\beta_j^{(n)}$ and $\beta_j^{(s)}$, respectively. Alternatively, a negative correlation in participant or item effects reflects a mirror effect (Glanzer & Adams, 1990), in which increases in d' are increases in the mean of the studied-item distribution and decreases in the mean of the new-item distribution (see Figure 1F). Although the main motivation for the hierarchical model is to assess asymmetry, hierarchical modeling has many other advantages, such as providing insight about participant and item variability, which are considered in the General Discussion.

Our goal is to localize the source of ROC asymmetry. To do so, we construct a set of submodels that represent theoretically important restrictions on the above hierarchical model (see Figure 2). The top-left model, denoted M_1 , is the full model presented above. The label "PIL σ " indicates that participants (P), items (I), and lags (L) have variable effects and that σ is not constrained. The top-right model, denoted M_2 , is the same as M_1 but with the restriction that $\sigma = 1$. Models M_3 and M_4 posit no effect of lag. Models M_5 and M_6 do not allow item or participant variability, respectively. Models M_7 and M_8 do not allow for any surface variation. We refer

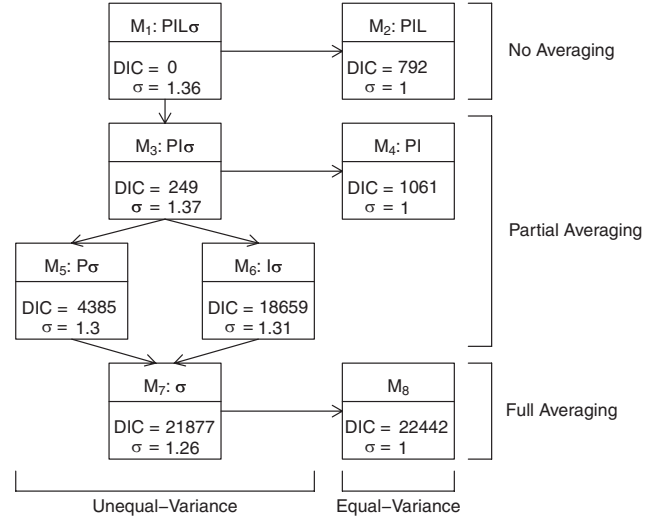


Figure 2. Relationships among the models. The labels for each model denote what effects were included. For example, " M_1 : PIL σ " indicates that M_1 includes effects of participants (P), items (I), lags (L), and free variance (σ). All other models place one or more restrictions on model M_1 . Deviance information criterion (DIC) values and estimates $\hat{\sigma}$ are from the fits to Experiment 1. All DIC values are reported in reference to that of Model M_1 ; larger DIC values denote a worse fit than that of Model M_1 .

to M_3 through M_6 as *partial averaging* models, as they are equivalent to averaging data over at least one surface variable. We refer to M_7 and M_8 as *full averaging* models, as they are equivalent to averaging over people, items, and lags.

Psychometric Properties

The hierarchical UVSD model is used here as a psychometric measurement tool. We show here that, for this purpose, the model is robust. That is, it may be used to measure ROC asymmetry even if UVSD or the constituent hierarchical assumptions fail.

1. How well does the model measure σ ? We addressed this question by simulating data from the equal-variance signal detection model M_2 (with true values from estimates in the subsequently reported experiment) and fitting the unequal-variance model M_1 . Over 100 such simulations, the mean estimate of σ was 1.04, and 95% of the estimates fell between 0.998 and 1.08. Although these simulations show that the estimate of σ has a slight positive bias, it can distinguish between symmetric ROCs ($\sigma = 1$) and those typically observed ($\sigma = 1.28$).

2. What happens if the effects of people and items are not distributed as normals? Morey, Pratte, and Rouder (2008) demonstrated that this assumption is not important. To show this fact, they simulated data in which participant and item effects were distributed as exponentials and found very good parameter recovery by fitting the UVSD model assuming normally distributed effects, especially with regard to σ .

3. What happens if the additivity assumption fails? The effects of items, people, and lags are assumed to be additive (i.e., no interactions among these factors exist). This explicit treatment of surface variation is far more realistic than the implicit assumption made when data are averaged that there are no item, partici-

pant, or lag effects. Nonetheless, potential violations of additivity may affect the estimates of σ . We assess the effects of violations in additivity after discussing the experiment, as the results guide our choices in simulation.

4. What happens if the underlying UVSD model is wrong?

To show that the model accurately measures asymmetry in general, we generated data from Yonelinas' (1994) dual-process model with participant and item effects on familiarity and one recollection parameter (R) across all participants and items. We ran 50 simulations in which the true value of R ranged from 0 to .6 in increments of .012. Estimates of σ were obtained by fitting model M_1 to these data. The corresponding estimates of σ are given approximately by $\hat{\sigma} = 1.02 + 2.8R$ (this linear relationship accounted for 97% of the variance). Thus, σ serves as a reasonable index of ROC asymmetry even when this asymmetry arises from recollection rather than strength.

Analysis

The hierarchical UVSD model is analyzed with conventional Bayesian hierarchical methods (Gelfand & Smith, 1990; Gelman, Carlin, Stern, & Rubin, 2004; Rouder & Lu, 2005). Bayesian analysis yields an estimate of the distribution of a parameter, termed the *posterior distribution*. The mean of this distribution, termed the *posterior mean*, serves as a point estimate. The region with 95% of the distribution mass is termed the *95% credible interval* (CI_{95}) and is analogous to a 95% confidence interval. To compare the models in Figure 2, we use the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). DIC is similar to the Akaike information criterion (AIC); however, it is better suited for comparing hierarchical models in which the number of parameters does not provide a good measure of a model's parsimony (some parameters are added to gain constraint rather than flexibility). DIC, like AIC, is measured on a log scale. Differences of 10 or more in DIC are interpreted as very strong evidence for the model with the smaller DIC.

In Bayesian models, priors are needed for free parameters. In M_1 , priors are needed for $\mu^{(s)}$, $\mu^{(n)}$, γ , σ^2 , and variances for random effects. We have experimented with a number of priors and find that diffuse, nearly noninformative priors are appropriate for this application. Morey, Pratte, and Rouder (2008) have provided extensive details concerning the priors and the method of analysis.¹ The R library HBMEM (available from CRAN or pcl.missouri.edu) provides functions for fitting the models presented herein.

Experiment

We ran a large-scale confidence-rating recognition memory experiment in which 97 participants were tested on 480 items. Although a 480-item test list is large, Glanzer et al. (1999) used the same length and obtained a typical sensitivity ($\hat{d}' = .94$) and variance ($\hat{\sigma} = 1.28$). If previous estimates of $\sigma > 1.0$ are confounded with participant or item variability, we expect the estimate of σ from the hierarchical model (M_1) to be closer to 1.0. Alternatively, if recognition memory ROC curves are truly asymmetric, the hierarchical model estimate of σ should be greater than 1.0.

Method

Participants

Ninety-seven University of Missouri students participated in Experiment 1 in return for credit toward a course requirement.

Stimuli

The word pool for the experiment consisted of 480 words from the MRC Psycholinguistic Database (Coltheart, 1981). Words were between four and nine letters in length and had a Kucera–Francis frequency of occurrence between 1 and 200 (Kucera & Francis, 1967). Study lists were constructed by randomly sampling 240 of these words, and there was a separate study list per participant. Test lists were composed of all 480 words. Presentation order was randomized at study and at test and across all participants. A second word pool was used for practice items. These practice items did not overlap with the main word pool.

Procedure

Participants began with a practice session in which they studied five items and made confidence ratings to a subsequent 10-item test list of words. Following practice, participants were presented the study list. Each word was displayed in the center of the screen for 1,850 ms, followed by a 250-ms blank period before the next word was presented. Participants were required to read each out loud to ensure that they attended to each word; compliance was monitored. Following study, participants completed the test phase. Each item was presented on the screen, and participants rated their confidence using the ratings “sure new,” “believe new,” “guess new,” “guess studied,” “believe studied,” and “sure studied.”

Model Analysis

Measuring Variance Components

We fit all of the models in Figure 2. The differences in DIC for each model compared with that for model M_1 are shown in the figure. As can be seen, all submodels are inferior to the full model, M_1 . This result indicates not only that participants, items, and lags contribute substantial variability but that σ is assuredly greater than 1.0.

Figure 3A shows the estimates of overall sensitivity d' from each model. The models with fixed σ (M_2 , M_4 , M_8) lead to lower estimates of sensitivity d' than do those with free σ (M_1 , M_3 , M_5 , M_6 , M_7). This pattern is expected because d' is defined with reference to the new-item distribution. More interesting is the effect of averaging. First, the more averaging, the lower the estimate of d' . This result is also expected; Rouder and Lu (2005)

¹ The model developed in Morey, Pratte, and Rouder (2008) was parameterized slightly differently than the one used here. Here, we fix the middle criterion to zero and the variance of the new-item distribution to 1.0. Morey, Pratte, and Rouder fixed the lowest and highest criteria to 0 and 1, respectively, and the variances of both the new- and studied-item distributions were free parameters. The conclusions drawn from the experiment do not depend on the parameterization.

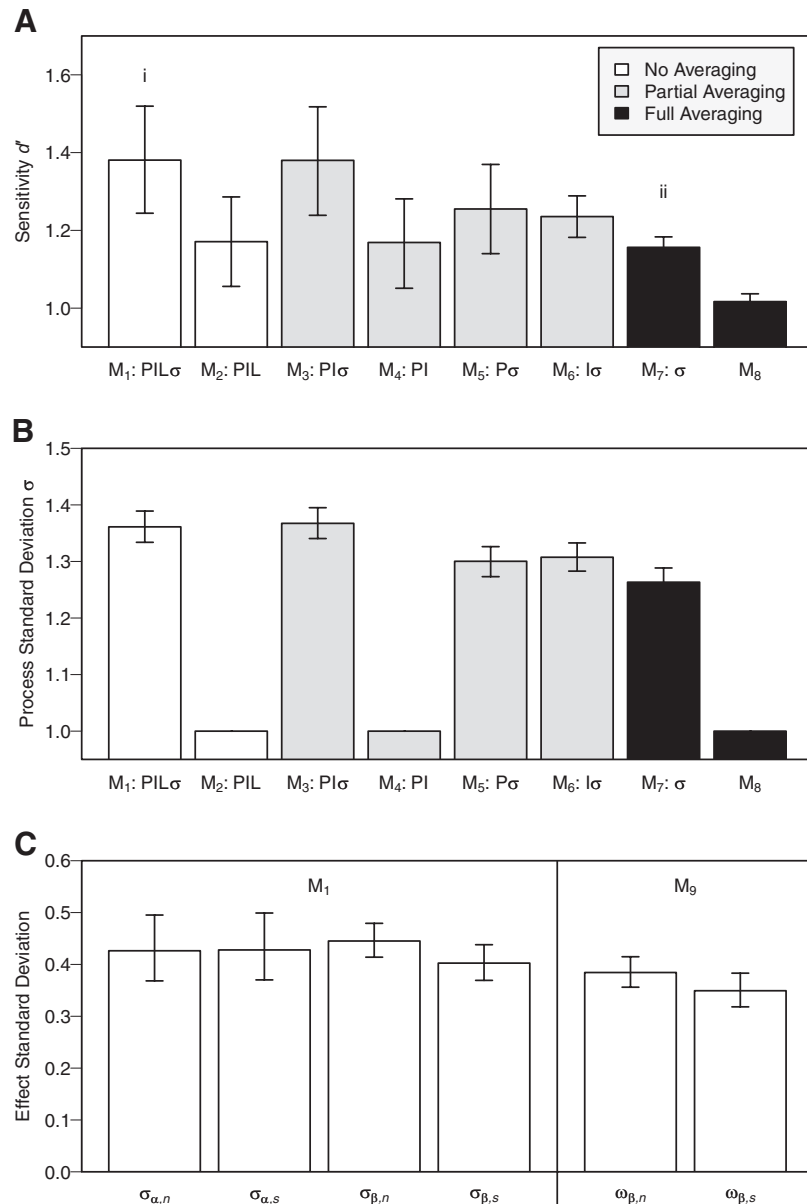


Figure 3. Model analyses. A: Estimates of d' from the eight models. B: Estimates of σ from the five unequal-variance models. Darker colors in A and B denote greater levels of averaging. C: Estimates of standard deviations of participant effects ($\sigma_{\alpha,n}$, $\sigma_{\alpha,s}$) and item effects ($\sigma_{\beta,n}$, $\sigma_{\beta,s}$) from model M_1 (left) and item effects from model M_9 , which do not include word-frequency effects (right). Error bars show 95% credible intervals.

noted that averaging artifactually reduces sensitivity estimates (see also Wickelgren, 1968).

Second, and more important, averaging leads to an underestimation of sampling error. For M_1 , the estimate of d' has a credible interval with a width of .28 (see error bar i). For M_7 , in contrast, the credible interval width is .05 (see error bar ii). The former interval is accurate, but the latter is wrong (for coverage estimates, see Morey, Pratte, and Rouder, 2008). The reason for this underestimation in confidence intervals with averaging is that systematic variation across items and participants includes correlations in performance; this inclusion, in turn, reduces the effective sample size (see Clark, 1973; Rouder & Lu,

2005). Here, the reduction in standard error is severe and leads researchers to have far too much confidence in a distorted estimate.

Figure 3B shows the estimates of σ . First, all models that have free parameter σ indicate an estimate well above 1.0, validating the large DIC increase for equal variance models in Figure 2. Second, averaging has only minimal effects, with increased averaging associated with marginally smaller estimates.

In the previous example with hard and easy items, we speculated that averaging artifactually increases estimates of σ . This speculation is clearly wrong. Why did the hierarchical model not provide a smaller estimate of σ than the averaging methods? The answer lies in

a comparison of participant and item variability in mnemonic strength across the new- and studied-item conditions. Estimates of this variability for participants ($\sigma_{\alpha,s}$, $\sigma_{\alpha,n}$) and items ($\sigma_{\beta,s}$, $\sigma_{\beta,n}$) are shown in Figure 3C (first four bars). Here we see that although there is substantial systematic item and participant variability, it is similar in size across the new- and studied-item conditions. In the example in Figures 1A and 1B, in contrast, there was sizable variability for studied items and none for new ones. Averaging data from the studied-item condition conflates estimates of process variability in the studied condition with studied-item variability. Averaging data from the new-item condition conflates estimates of process variability in the new condition with new-item variability. If both of these item-variability sources are the same size, the degree of inflation is about the same in both conditions. Because the estimate of σ reflects the ratio of studied-item strength variance to new-item strength variance, averaging the data from both conditions inflates both variances about equally, and their ratio, σ , is approximately preserved. This pattern cannot be predicted a priori and could not have been discovered without the hierarchical model analysis.

In the preceding section on psychometric properties, we worried about violations of the additivity of surface variation. The observed pattern, in which there are equal amounts of item and participant variability across conditions, provides guidance for assessment. The worst-case scenario is that violations of additivity occur only for studied items and not at all for new items, as this pattern of variability leads to an artifactual overestimation of σ . To assess the implications, we simulated data from the equal-variance model M_2 with the addition of an interaction term to the studied-item distribution mean only. This interaction was of the same magnitude as the item effects themselves. Over 100 such simulations, the mean estimate of σ from model M_1 was 1.21, and 95% of estimates fell within 1.16 and 1.26. This estimate is certainly too high and reflects the fact that interactions enter only for studied items. Even so, it is far lower than and cannot account for our finding of $\sigma = 1.36$. Thus, even in this worst-case and highly implausible scenario, violations of additivity cannot account for the degree of asymmetry present in ROCs.

The Structure of Lag, Participant, and Item Effects

Although model M_1 was designed to measure ROC asymmetry, it is also useful for assessing lag, participant, and item effects. We consider first the effect of lag on the mean of the studied-item distribution. The DIC value of 249 for model M_3 in Figure 2 indicates that lag does indeed have an effect. The estimated slope of the lag effect, the amount of change in $d^{(s)}$ per one-item of lag, is $\hat{\gamma} = -.001$. Although this number is small, the effect of lag across all 717 levels on d' is .717, which is quite large given that the average sensitivity is $d' = 1.38$. Even though there is a lag effect, this effect does not have an appreciable influence on the estimate of σ (compare models M_3 and M_1 in Figure 3B).

Participants and items also had significant effects on the new- and studied-item distribution means. In model M_1 , \hat{d}' values for participants ranged from 0.12 to 2.44, and those for items ranged from 0.21 to 3.60. Figure 4A shows a scatter plot in which the participant effects in the studied condition are plotted as a function of those in the new condition. Figure 4B shows the same for items. For participants, there is a positive relationship, $r = .30$, $t(95) = 3.11$; that is, people with higher strength in the new condition have

higher strength in the studied condition.² As discussed previously, this positive correlation indicates that people primarily vary in overall response bias (see Figure 1E), in addition to differences in sensitivity d' . In contrast, items exhibit a mirror effect: Items that have higher strength in the new condition tend to have lower strength in the studied condition, $r = -.22$, $t(478) = 4.82$.

The ability to estimate item effects without distortion from averaging over participants also allows for an examination of the effects of item characteristics. Here, we highlight the effect of word frequency on mnemonic strength. Item effects for the new- and studied-item conditions are plotted as a function of word frequency in Figures 5A and 5B, respectively. As can be seen, there is a substantial mirror effect: Increases in word frequency are associated with increases in baseline strength ($\beta_j^{(n)}$) and with decreases in studied strength ($\beta_j^{(s)}$).

The finding of strong word-frequency effects raises two additional questions: (a) Are there item effects in this data set above those from word frequency? and (b) If there are such extra-word-frequency item effects, do they also exhibit a mirror-effect pattern? To answer these questions, we constructed a model, M_9 , that included word frequency as a covariate on item effects:

$$\beta_j^{(n)} \sim \text{Normal}(\theta^{(n)}[f_j - f_0], \omega_{\beta,n}^2),$$

$$\beta_j^{(s)} \sim \text{Normal}(\theta^{(s)}[f_j - f_0], \omega_{\beta,s}^2),$$

where $\theta^{(n)}$ and $\theta^{(s)}$ are slopes of the word frequency effect; f_j is the logarithm of the Kucera–Francis word frequency for the j th item; and f_0 is the mean of these log-word frequencies, such that the sum of these terms is zero. These parameters are estimated at $\theta^{(n)} = .15$ ($CI_{95} = [.14, .16]$) and $\theta^{(s)} = -.14$ ($CI_{95} = [-.15, -.12]$), which confirms the existence of a substantial frequency-based mirror effect.

The systematic item variability not accounted for by word frequency is given by $\omega_{\beta,n}^2$ and $\omega_{\beta,s}^2$. These variances are useful for answering the first question about extra-word-frequency item effects. Estimates of these variances are shown as the two rightmost bars in Figure 3C. The credible intervals are well above zero, indicating substantial extra-word-frequency item effects. In fact, only 25% of the systematic item-level variance may be attributed to word frequency; the remaining 75% is from other factors.

The second question about whether these extra-word-frequency item effects exhibit a mirror effect may be assessed by plotting item effects from M_9 (see Figure 4C). Removing word-frequency effects greatly attenuates any item-level mirror effect, $r = -.007$, $t(478) = 0.15$. Whereas there is a mirror effect for word frequency, there does not appear to be one for other item characteristics that are orthogonal to word frequency, even though these other characteristics account for a substantial proportion of item variability. We note, however, that our failure to find extra-word-frequency mirror effects may not generalize to other studies in which, for example, stimuli have a larger range on these orthogonal item characteristics (e.g., word length). The mirror

² It is not clear how to assess the statistical significance of this correlation. The associated t value is highly significant. Yet, in interpreting this value one assumes that the estimates are independent. In this case, they are not, as they are connected by a hierarchical structure. We suspect that the current priors, which assume independence, result in a bias toward attenuating correlations (see Rouder et al., 2007) and that the interpretation of the t value is, if anything, conservative.

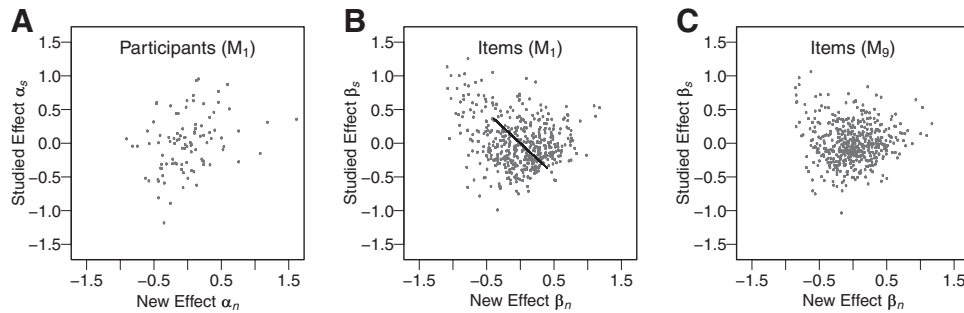


Figure 4. Participant and item effects. A: Participant effects in the studied condition plotted as a function of those in the new condition. The positive correlation is a response bias. B: Item effects in the studied condition plotted as a function of those in the new condition. The negative correlation is a mirror effect. The solid line represents the word-frequency effects obtained from model M_9 , in which word frequency is linearly regressed onto item effects. C: Item effects in the studied condition plotted as a function of those in the new condition after word frequency is partialled out. The mirror effect seen in Panel B is greatly attenuated.

effect has been interpreted as supporting a likelihood basis of decision making in which participants assess what their mnemonic strength would have been if the item was and was not studied (Glanzer, Adams, Iverson, & Kim, 1993). The current results indicate that participants may be able to make these calculations with regard to word frequency but not, perhaps, with regard to other item characteristics that are orthogonal to word frequency.

General Discussion

We analyzed ROC data with a hierarchical version of UVSD to assess the source of ROC asymmetry. The results revealed that these asymmetries are not due to averaging data but reflect the underlying mnemonic process. Although our results do not differentially support any one model of asymmetry over others, they vindicate the class of models that assume ROC asymmetry results from the structure of the mnemonic process.

Our results may seemingly license the use of averaging data, as the estimates of σ from averaged data are nearly the same as that from the full hierarchical model. We believe that this view is mistaken and that hierarchical modeling has much to offer recognition-memory researchers. Consider the following:

1. Accurate Point Estimates

Although we found a near equivalence between data-averaged and hierarchical estimates of σ in this experiment, it would be dangerous to assume that the equivalence holds generally. For example, consider a levels-of-processing experiment in which there is a common false alarm rate across levels (new items at test cannot be considered deep or shallow). Suppose items have more variability in sensitivity in the shallow-study condition than in the deep-study condition, but the true value of σ is the same for both. When averaged, this differential item variability will make it seem that there is a greater value of σ for the shallow-study than the deep-study condition. Nature cannot be relied upon to ensure equal item and participant variability across all conditions of interest to psychologists. Hence, hierarchical models serve as necessary insurance against unforeseen differences in participant and item variability across conditions.

2. Accurate Confidence Intervals on Sensitivity Estimates

Unaccounted surface variation leads to an underestimation of the confidence intervals on sensitivity and, consequently, an increase in

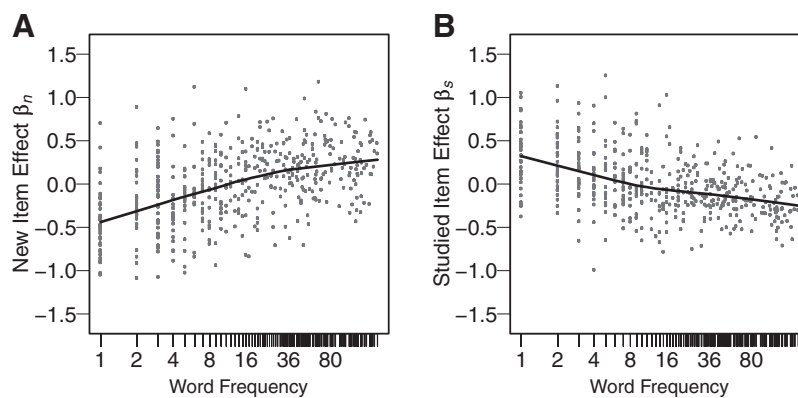


Figure 5. Word frequency effects. A and B: Item effects in the new and studied conditions, respectively, plotted as a function of word frequency on a log scale. Lines are nonparametric smooths (Cleveland, 1981).

the real Type I error rate over the nominally reported value (Clark, 1973; Rouder & Lu, 2005). This bias is asymptotic, and the magnitude of the bias depends on the magnitude of unaccounted surface variation. Hierarchical models that account for surface variation avoid these difficulties and provide for principled inference.

3. Powerful Model Comparison

Hierarchical modeling allows for more rigorous model comparisons than does data-averaged analysis. In data-averaged analysis, a model is required to fit a few averaged curves. Alternatively, hierarchical versions of the models must accurately specify not only how the memory system differs across conditions but also how it differs across people and items. This level of specification and comparison seems highly appropriate for any model that purportedly explains the mnemonic system.

4. Measurement of Item and Participant Effects

Several areas of memory research involve the study of individual differences or the effects of item covariates. Hierarchical analysis allows for the accurate characterization of participant and item covariates on the mnemonic system.

References

- 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.
- Cleveland, W. S. (1981). LOWESS: A program for smoothing scatter plots by robust locally weighted regression. *American Statistician*, *35*, 54.
- Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *33*, 497–505.
- Curran, T. C., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 531–547.
- DeCarlo, L. T. (2002). Single detection theory with finite mixture distribution: Theoretical developments with applications to recognition memory. *Psychological Review*, *109*, 710–721.
- 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, England: Chapman & Hall.
- Glanzer, M., & Adams, J. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 5–16.
- Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*, 546–567.
- Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 500–513.
- Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics*. Huntington, NY: Krieger. (Original work published 1966)
- Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1210–1230.
- Heathcote, A., Raymond, F., & Dunn, J. (2006). Recognition and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory and Language*, *55*, 495–514.
- Hintzman, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, *87*, 398–410.
- Hintzman, D. L., & Hartry, A. L. (1990). Item effects in recognition and fragment completion: Category relations vary for different subsets of words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 955–969.
- Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review*, *74*, 496–504.
- Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*, 376–388.
- Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*, 21–36.
- Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, association, and serial order information. *Psychological Review*, *100*, 183–203.
- Ratcliff, R., Sheu, C. F., & Grondlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535.
- Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.
- Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, *137*, 370–389.
- Rouder, J. N., Lu, J., Sun, D., Speckman, P. L., Morey, R. D., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621–642.
- Rouder, J. N., Tuerlinckx, F., Speckman, P. L., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, *15*, 1201–1208.
- Schacter, D., & Tulving, E. (1994). What are the memory systems of 1994? In D. Schacter & E. Tulving (Eds.), *Memory systems 1994*. Cambridge, MA: MIT Press.
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.
- Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition*, *33*, 1499–1517.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *64*, 583–639.
- Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology*, *5*, 102–122.
- Wixted, J. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*, 152–176.
- Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1341–1354.
- Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*, 800–832.

Received February 17, 2009
 Revision received July 7, 2009
 Accepted July 27, 2009 ■