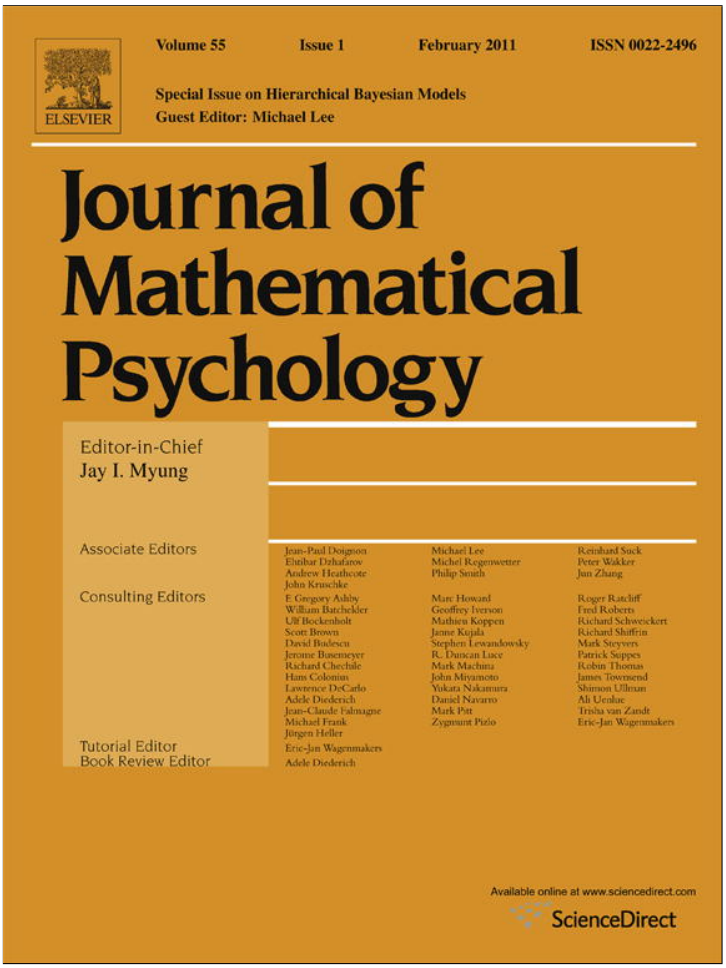

Journal of Mathematical Psychology 55 (2011) 36–46

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Hierarchical single- and dual-process models of recognition memory

Michael S. Pratte *, Jeffrey N. Rouder

University of Missouri, United States

ARTICLE INFO

Article history:
Received 6 January 2010
Received in revised form
24 May 2010
Available online 24 September 2010

Keywords:
Recognition memory
Hierarchical models
Bayesian
Signal detection
Dual process

ABSTRACT

Recognition memory is commonly modeled as either a single, continuous process within the theory of signal detection, or with two-process models such as Yonelinas' dual-process model. Previous attempts to determine which model provides a better account of the data have relied on fitting the models to data that are averaged over items. Because such averaging distorts conclusions, we develop and compare hierarchical versions of competing single and dual-process models that account for item variability. The dual-process model provides a superior account of a typical data set when models are compared with the deviance information criterion. Parameters of the dual-process model are highly correlated, however, suggesting that a single-process model may exist that can provide a better account of the data.

Memory researchers have long been interested in determining the number of *processes* that underlie memory. Examples of separating memory into distinct mnemonic processes include the division of memory into long and short term stores (James, 1890), into episodic and semantic components (Tulving & Craik, 2000), and into implicit and explicit systems (Schacter, 1990). A current debate is whether recognition memory reflects a single strength-based process or reflects distinct familiarity and recollective processes (see Wixted, 2007; Yonelinas & Parks, 2007, for reviews). The assessment of which position best describes the data has been repeated many times. Both positions have strong advocates who claim the evidence strongly favors their position.

Our goal in this paper is the same as that of previous authors — we assess the abilities' of several single and dual process models to account for data. One problem with previous attempts to fit recognition memory models is that these fits were made to averaged data. In recognition memory experiments, each participant is tested on each item only once, and the basic structure of the data is that each observation is unreplicated. In analysis, researchers almost always averaged data across either participants or items or both to construct proportions such as hit and false alarm rates. This averaging would be justified if people or items didn't differ. Unfortunately, they do, and the variation across both people and items is substantial (Pratte, Rouder, & Morey, 2010; Rouder, Lu, Morey, Sun, & Speckman, 2008). We have shown that this averaging across variable items or people is problematic in nonlinear models because it leads to asymptotic distortions in parameter estimates as well as distortions in parameter coverage (e.g., Pratte et al., 2010; Rouder & Lu, 2005; Rouder, Tuerlinckx, Speckman, Lu, & Gomez, 2008). To avoid these distortions, we develop hierarchical versions of memory models that account simultaneously for variability across participants and across items, as well as variability reflecting the mnemonic processes. These hierarchical models allow us to assess single- and dual-process models with far greater confidence than previous assessments.

A common experimental method for assessing recognition memory processing is to have participants provide unidimensional confidence ratings at test, where confidence is assessed on a univariate scale anchored by high confidence that an item was or was not studied. Data are plotted and analyzed with *receiver-operating-characteristic* (ROC) plots. These plots display the cumulative probabilities of making confidence ratings to studied items (termed *hit rates*) as a function of the cumulative probabilities of making the same ratings to new items (termed *false alarm rates*). The key findings are that ROC plots (a) are curved, and (b) exhibit a small but reliable asymmetry around the negative diagonal (see Fig. 1(B) for typically observed ROC curves). These findings motivate the following single-process and dual-process recognition models.

One common approach to modeling ROC data is the signal detection paradigm of Green and Swets (1966). Signal detection is often treated as a psychometric model that provides a separate measurement of sensitivity and bias (e.g., Macmillan & Creelman, 2005). More recently, however, memory researchers have elevated signal detection to a model of internal psychological processing, in which the stipulated representations and decision processes are considered faithful models of what people actually do (e.g., Mickes, Wixted, & Wais, 2007). The current state-of-the-art is the

* Corresponding address: 210 McAlester Hall, Columbia, MO 65211, United States.
E-mail address: prattems@gmail.com (M.S. Pratte).

0022-2496/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2010.08.007

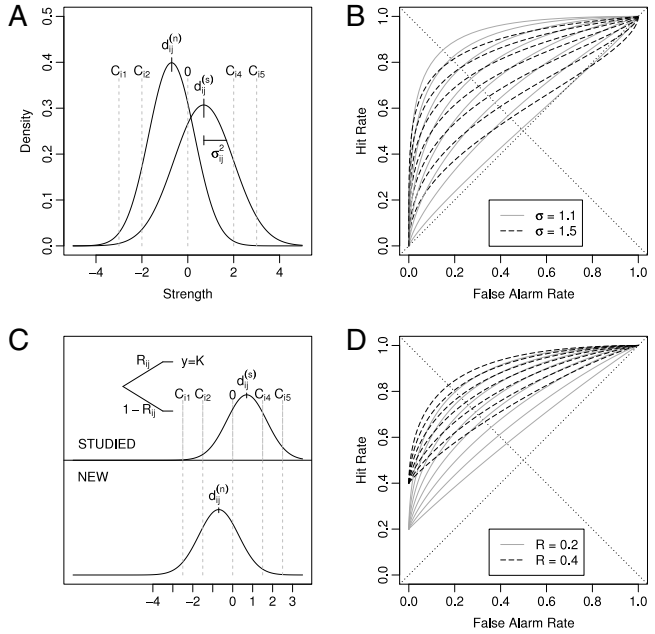

Fig. 1. Hierarchical UVSD and DPSD models. (A & B) Hierarchical unequal-variance signal detection model and resulting ROC curves. (C & D). Hierarchical dual-process signal detection model and resulting asymmetric ROC curves.

unequal-variance signal detection (UVSD) model, discussed by Egan (1975) and recently championed as a process model of recognition memory by Wixted (2007). This model posits that recognition memory judgments are based on latent mnemonic strengths. When a tested item is new, its strength is distributed as a unit normal. Studied-item strength distributions have both a larger mean (denoted d') and variance (denoted σ^2). The resulting ROC plots are curved and have a degree of asymmetry that is determined by the value of σ^2. When $\sigma^2 > 1$, the asymmetry favors greater hit rates at lower false alarm rates, as in Fig. 1(B). Hence, UVSD currently serves as the iconic single-process account of recognition memory.

A popular dual-process model is Yonelinas' (1994) *dual-process signal detection* (DPSD) model. In this model, performance is a mixture of recollection and familiarity. Recollection is modeled as a high-threshold component which occurs with probability R if the tested item was studied and does not occur otherwise. Familiarity is modeled as an equal-variance signal detection process; i.e. $\sigma^2 = 1$. This model also produces curved ROC plots that display the appropriate asymmetry (see Fig. 1(D)). The degree of asymmetry is determined by the value of R. When $R = 0$ the ROC is symmetric; the ROC becomes more asymmetric as R increases above zero.

The UVSD model and the DPSD model serve as base models for subsequent development. In the next section, we take a slight digression and discuss the difference between variability in process and variability in items. Following that, we develop several recognition memory models. These are then compared on benchmark data from Pratte et al. (2010) with model selection performed by the deviance information criterion (DIC, Spiegelhalter, Best, Carlin, & Linde, 2002). The results are complex. We find better support for dual-process models over single process counterparts. These dual-process models, however, show a fair, though not complete, degree of correlation between recollection and familiarity. This result suggests that there are not two independent processes in recognition memory.

1. Separating sources of variability

It is critical to separate variability due to process from that due to items and participants. To understand variability in process, consider the hypothetical situation in which a single participant is tested on a single item. If we could perform this test over and over, without any learning or testing effects, the variability in response would reflect the mnemonic process for this participant and item combination. The underlying patterns in this variability, such as whether the resulting confidence rating ROCs are curved or straight lines, or are asymmetric, reflect deep structural properties that are not a function of the person or item. An example of such a property is that recognition memory is mediated by a mixture of recollection and familiarity. In fact, the elucidation of these structural properties, uncontaminated by variability in people or items, are the main object of study. Separate from these deep structural properties is variation from items and participants. For instance, the probability of recollection in DPSD may vary across items.

The ramifications of this view are subtle, especially when discussing the UVSD model. Mickes et al. (2007) and Wixted (2007), for example, speculate that the asymmetry in ROC plots reflects item variability rather than a deep structural property. They note that items will differ in sensitivity, and when data are averaged, this difference will result in greater variability across studied items. This scenario, however, is not the only possibility. It may be that the asymmetry is a deep structural property that does not reflect item variability but is, instead, a signature of process variability. In fact, the dual-process model is a process-variability explanation because recollection is assumed to occur and add asymmetry even in the absence of item variation. Pratte et al. (2010) assessed these explanations of ROC asymmetry using a hierarchical version of the UVSD model (to be discussed subsequently). They found that ROC asymmetry was even more pronounced when item variation was modeled, indicating that the asymmetry reflects process variability. In the next section we develop several single- and dual-process models that separate participant, item, and process variation. By separating process from these other sources, we can see which process model, uncontaminated by participant and item effects, provides for the best account of the data.

2. Model development

We develop twelve variants of single and dual process models. These models are presented below and their relations are shown in Fig. 2. All models are presented and analyzed in a Bayesian framework. Primers on Bayesian hierarchical models for psychologists are provided in Lee (1997) and Rouder and Lu (2005). In the following sections, we provide the core specification of the models. Technical elements, such as the specification of the priors, expressions for conditional posterior distributions, and the corresponding Markov chain Monte Carlo sampling steps, are provided in the Appendix. The analysis software is available as a package for the **R** data analysis language. The package, `hbmem`, may be downloaded from http://www.cran.r-project.org/ or http://pcl.missouri.edu.

The following notation is used throughout: Let $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$ index the participants, items, and confidence ratings, respectively. Let y_{ij} denote ith participant's confidence rating to the jth item. Pratte et al. (2010) showed that the number of intervening study and test trials between the study and test of an item (denoted *lag*) has a significant effect on studied-item strength, and that a linear function between lag and mean strength provided a good fit. Let l_{ij} index the (zero-centered) lag for the ith participant tested on the jth studied item.

2.1. Equal-variance signal detection model

We start with the equal-variance signal detection (EVSD) model because it is a restriction of both UVSD and DPSD. Elements of

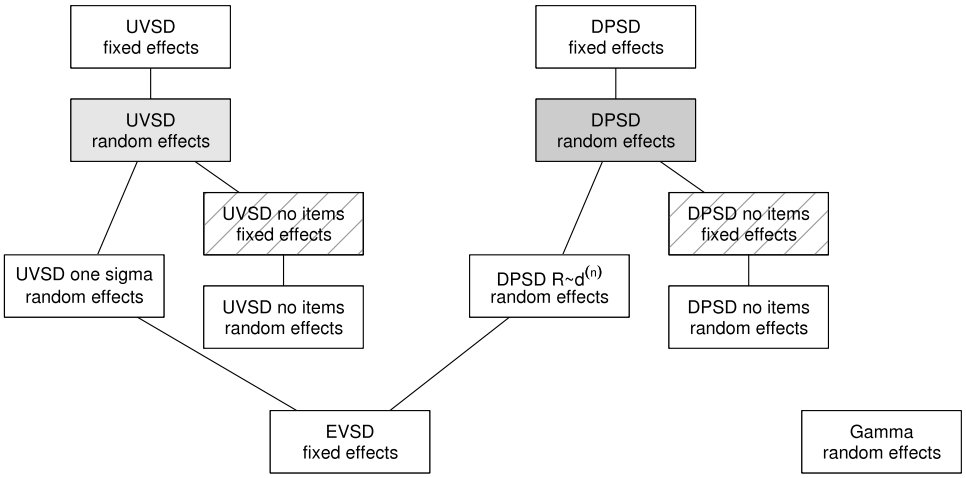

Fig. 2. Relationships between models. Lines denote nested relationships. Shaded boxes denote the best-fitting DPSD model (right) and best-fitting UVSD model (left). Boxes with hatched lines denote models comparable to typical analysis.

the development of this model aid in the development of UVSD and DPSD models. According to signal detection theory, each item at test gives rise to some latent mnemonic strength denoted X_{ij}. Participants are assumed to place criteria on the latent strength dimension such that the probability of making the kth response is equal to the probability that the latent strength falls between the $k-1$ and kth criterion. Typically, the mean strength of new items is set to 0.0 in order to fix the location of the strength space. It is more convenient in accounting for item and participant effects to allow this mean to be a free parameter and center the space on a criterion (see Rouder, Lu, Sun, Speckman, & Naveh-Benjamin, 2007). A graphical representation of the EVSD model with this parametrization is shown in Fig. 3(A). Latent strengths are given by:

$$X_{ij} \sim \begin{cases} \text{Normal}(d_{ij}^{(n)}, 1) & \text{new}, \\ \text{Normal}(d_{ij}^{(s)}, 1) & \text{studied}, \end{cases}$$

where $d_{ij}^{(s)}$ and $d_{ij}^{(n)}$ are means for the ij person-by-item combination when the item is studied and new, respectively. The probability that the ith person makes the kth response to the jth item is:

$$\Pr(y_{ij} = k|\ \text{new}) = \Phi\left(d_{ij}^{(n)} - C_{i(k-1)}\right) - \Phi\left(d_{ij}^{(n)} - C_{ik}\right),$$

$$\Pr(y_{ij} = k|\ \text{studied}) = \Phi\left(d_{ij}^{(s)} - C_{i(k-1)}\right) - \Phi\left(d_{ij}^{(s)} - C_{ik}\right),$$

where $C_0 = -\infty$, $C_{K/2} = 0$ ($C_{(K+1)/2} = 0$ if K is odd), $C_K = \infty$, and Φ is the CDF of the standard normal. In this parametrization sensitivity d' is the distance between the new and studied-item distributions $d'_{ij} = d_{ij}^{(s)} - d_{ij}^{(n)}$. Overall new/studied response biases manifest as concurrent shifts in the distributions, and can be measured as $\text{bias}_{ij} = (d_{ij}^{(s)} + d_{ij}^{(n)})/2$.

The means of the new and studied-item distributions can not be estimated without some restriction, as each participant-by-item combination occurs only once, and only as *either* new or studied. To make the model estimable, additive structures are placed on the means:

$$d_{ij}^{(n)} = \mu^{(n)} + \alpha_i^{(n)} + \beta_j^{(n)}, \tag{1}$$

$$d_{ij}^{(s)} = \mu^{(s)} + \alpha_i^{(s)} + \beta_j^{(s)} + \theta^{(s)} l_{ij}, \tag{2}$$

where $\mu^{(n)}$ and $\mu^{(s)}$ are grand means, $\alpha_i^{(n)}$ and $\alpha_i^{(s)}$ are participant effects, and $\beta_j^{(n)}$ and $\beta_j^{(s)}$ are item effects. Parameter $\theta^{(s)}$ is the linear effect of study-test lag.

There are two approaches to modeling participant and item effects. The first is to assume that these effects are *fixed effects*; that is, they are unconstrained and free to take any value (this approach is sometimes referred to as *full individual differences*). The second approach is to assume that the effects are *random effects*, that is, they are samples from a *parent* distribution which places constraint on the distribution of values (this approach is sometimes referred to as *structured individual differences*). Both approaches may be implemented by placing the following structure on participant and item effects:

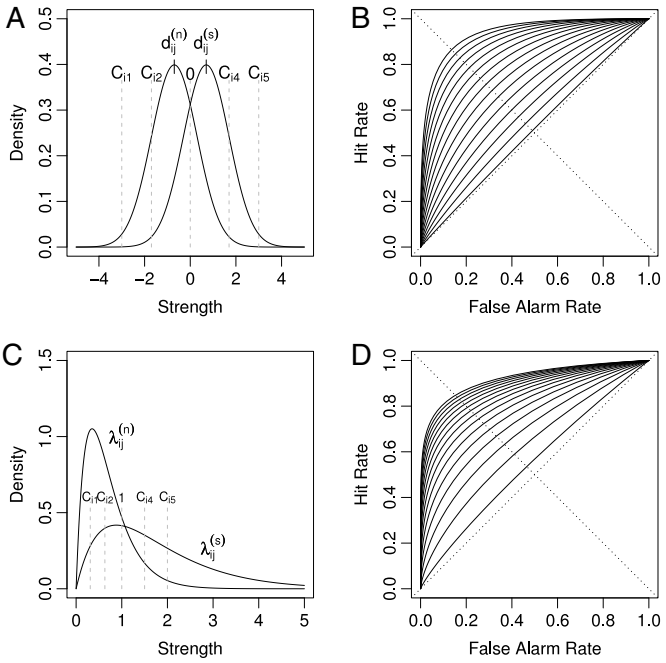

Fig. 3. Hierarchical single process models. (A & B) Hierarchical equal-variance signal detection model and resulting symmetric ROC curves. (C & D) Hierarchical gamma signal detection model and resulting asymmetric ROC curves.

$$\alpha_i^{(n)} \sim \text{Normal}(0, \sigma_{\alpha,n}^2), \tag{3}$$

$$\beta_j^{(n)} \sim \text{Normal}(0, \sigma_{\beta,n}^2), \tag{4}$$

$$\alpha_i^{(s)} \sim \text{Normal}(0, \sigma_{\alpha,s}^2), \tag{5}$$

$$\beta_j^{(s)} \sim \text{Normal}(0, \sigma_{\beta,s}^2). \tag{6}$$

To treat the effects as fixed, the variances $(\sigma_{\alpha,n}^2, \sigma_{\beta,n}^2, \sigma_{\alpha,s}^2, \sigma_{\beta,s}^2)$ are set exceptionally large such that the normals place little

constraint on the values. Conversely, to treat the effects as random, these variances are free parameters that are estimated from the data. Random-effects modeling has been preferred for hierarchical modeling of cognitive phenomena because it provides a more parsimonious model as well as a means of generalizing results to a larger class of participants and items (e.g., Lee, 2006; Morey, Rouder, & Speckman, 2008, 2009; Pooley, Lee, & Shankle, 2011; Rouder et al., 2007; Rouder, Morey et al., 2008; Rouder, Tuerlinckx et al., 2008).

In this report, we fit both fixed and random effects models for two reasons. First, the constraint in random effects models, while previously assumed, has never been assessed in this context. A model-selection comparison of fixed and random effect versions allows assessment of the random-effects constraint. Second, and perhaps more importantly, the fixed-effect models simulate common practice in the field in which model parameters are estimated separately and independently for each participant without constraint. Including fixed-effect models allows for a more natural comparison of our models with more common approaches.

The structures in Eqs. (3)–(6) serve as *priors* on effects. These priors have an implicit independence assumption; that is, in the absence of data the effects are assumed to be independent. Consequently, any correlations between these effects in the data may be underestimated with these priors. It is possible to develop correlated priors; for example, Rouder et al. (2007) develop a Wishart prior that includes correlations in signal detection models. We chose the independence priors in (3)–(6) to insure that any observed correlations, which we interpret as markers of psychological process, reflect correlations in data without bias from the prior.

EVSD ROC curves are shown in Fig. 3(B). Each line differs in sensitivity (d') and may be thought of as the result of different participant-by-item combinations. One feature of these curves is that they are symmetric around the negative diagonal and this feature stands in contrast to observed data. A second feature is that the curves form an orderly field. These lines have a coherent relationship that holds across participants, items, conditions, and experiments. In this regard, the predictions are highly constrained and the model exhibits a large degree of parsimony.

2.2. Unequal-variance signal detection models

The hierarchical extension of UVSD is identical to the equal-variance version developed above except that the variance of the studied-item distribution is a free parameter denoted σ^2 (see Fig. 1(A)). In the most general case, we let σ^2 vary across participants and items. Because σ^2 must be positive, additive effects are placed on the log of σ_{ij}^2 as follows:

$$\log(\sigma_{ij}^2) = \mu^{(\sigma)} + \alpha_i^{(\sigma)} + \beta_j^{(\sigma)} + \theta^{(\sigma)} l_{ij}, \tag{7}$$

where $\mu^{(\sigma)}$ is a grand mean, $\alpha_i^{(\sigma)}$ are participant effects, $\beta_j^{(\sigma)}$ are item effects, and $\theta^{(\sigma)}$ is the linear effect of lag. These random effects are constrained to follow the analogous normal priors in Eqs. (3)–(6). We developed fixed and random effect versions of these priors on $d^{(n)}$, $d^{(s)}$, and σ^2. The fixed- and random-effects versions of UVSD are depicted as the two top-left models, labeled "UVSD fixed effects" and "UVSD random effects", respectively, in Fig. 2.

The UVSD ROC curves are more complicated than those of EVSD. Fig. 1(B) shows the case for various values of d' when $\sigma = 1.1$ and $\sigma = 1.5$. As can be seen, increasing σ provides for greater ROC asymmetry. Although these curves provide for asymmetry, there are two undesirable properties. First, when sensitivity is low, ROC curves may dip below the diagonal. This dip implies that performance is *worse than chance*. To our knowledge no such pattern has ever been demonstrated in typical recognition memory tasks in which ROCs are drawn for studied and new items (cf. Heathcote, Raymond, & Dunn, 2006). The second consequence of unequal variances is that the curves do not display the same degree of orderliness as those for EVSD. For instance, the UVSD curves cross each other whereas the EVSD ones do not. In this sense, UVSD is far less parsimonious than EVSD.

There are several interesting restrictions on UVSD. In our previous work (Morey, Pratte, & Rouder, 2008; Pratte et al., 2010), we assumed that σ^2 was an invariant of the memory processing system that did not vary across people, items, or lags. We can test whether this assumption is warranted by comparing the general UVSD model to one in which σ_{ij}^2 is constrained to be constant across people, items, and lags. This constrained model is depicted in Fig. 2 in the box labeled "UVSD one sigma". A second set of constraints is motivated by current practice. The typical approach is to average data over items to produce participant-specific effects. Averaging over items implicitly assumes there are no item or lag effects, and so we implemented UVSD with no item or lag effects in $d^{(n)}$, $d^{(s)}$, or σ^2. The fixed-effects version of this no-item effects restriction is the closest Bayesian analog of the typical analysis, and is shown as the box with hatched lines in Fig. 2.

2.3. Dual-process signal detection models

A hierarchical version of the Yonelinas' dual-process signal detection model is shown in Fig. 1(C). The structure of the EVSD component is identical to that developed above (see Eqs. (1)–(6)) — means of the new and studied-item distributions are the additive combinations of participant and item effects, and all but the middle criteria are free to vary across participants. In the hierarchical version, recollection, R_{ij}, varies across participants, items, and lags. As before, recollection can not be estimated for every participant-by-item combination without constraint. We place an additive structure on the probit link (i.e., the quantile function of the standard normal):

$$\Phi^{-1}(R_{ij}) = \mu^{(r)} + \alpha_i^{(r)} + \beta_j^{(r)} \theta^{(r)} l_{ij},$$

where $\mu^{(r)}$ is a grand mean, $\alpha_i^{(r)}$ are participant effects, $\beta_j^{(r)}$ are item effects, and $\theta^{(r)}$ is the linear effect of lag. We implement versions with these participant and item effects as fixed or random in a manner analogous to that in Eqs. (3)–(6). The fixed- and random-effects version of DPSD are depicted as the two top-right models, labeled "DPSD fixed effects" and "DPSD random effects", respectively, in Fig. 2.

According to the hierarchical dual-process model, the probability that the ith person makes the kth response to the jth item is:

$$\Pr(y_{ij} = k|\text{ new}) = \Phi\left(d_{ij}^{(n)} - C_{i(k-1)}\right) - \Phi\left(d_{ij}^{(n)} - C_{ik}\right),$$

$$\begin{aligned}\Pr(y_{ij} &= k \in \{1, \ldots, K-1\}|\text{ studied})\\ &= \left(1 - R_{ij}\right)\left[\Phi\left(d_{ij}^{(s)} - C_{i(k-1)}\right) - \Phi\left(d_{ij}^{(s)} - C_{ik}\right)\right],\end{aligned}$$

$$\Pr(y_{ij} = K|\text{ studied}) = R_{ij} + \left(1 - R_{ij}\right)\Phi\left(d_{ij}^{(s)} - C_{i(K-1)}\right),$$

Fig. 1(D) shows ROC curves generated from dual-process models with recollection probabilities of either 0.2 or 0.4. Like the UVSD model, the predictions of the dual-process model are far less constrained than those of EVSD. ROC curves need not lie in any uniform field and may cross. This loss of parsimony is exactly what a *dual*-process model implies — that ROC curves can not be accounted for with a simple one-parameter model.

The right branch of Fig. 2 shows the DPSD models considered here. As with UVSD, there are several interesting constraints on the above DPSD model. The first constraint is to assume a linear

relationship between recollection and familiarity (labeled "DPSD $R \sim d^{(n)}$" in Fig. 2). This constraint implies that recollection and familiarity are the result of a single process and that the dual-process models may be overspecified. As with UVSD, the typical approach to fitting DPSD is to average data over items to produce participant specific effects. Averaging implicitly assumes there are no item or lag effects. To mimic averaging, we implemented DPSD with no item or lag effects in $d^{(n)}$, $d^{(s)}$, or R. The fixed-effects version of this no-item-effects restriction is the closest Bayesian analog of the typical analysis, and is shown in Fig. 2 as the box with hatched lines.

2.4. Gamma signal detection model

One of the features of EVSD that we find most attractive is the constraint it places on fields of ROC curves. This constraint is greatly diminished in UVSD and DPSD models. The downside of EVSD is that it is not compatible with the universally observed asymmetry in ROC curves. We develop a novel hierarchical signal detection model that retains the parsimony of EVSD while predicting ROC asymmetry. In our new model, latent strengths are distributed as gamma distributions rather than as normal distributions. The effect of study is to increase the scale of the distribution; participants set criteria on the latent space as before.

Fig. 3(C) shows the hierarchical gamma signal detection (GSD) model. Latent mnemonic strengths are given by:

$$X_{ij} \sim \begin{cases} \text{Gamma}(2, \lambda_{ij}^{(n)}) & \text{new}, \\ \text{Gamma}(2, \lambda_{ij}^{(s)}) & \text{studied}. \end{cases}$$

The shapes of both gamma distributions are fixed to 2.0. This value was chosen not for any deep theoretical reasons, but simply because gamma distributions with this shape provide about the right degree of asymmetry. We tried a few other values of shape, but none provided a better fit. To fix the scale of the space, the middle criterion is fixed to 1.0, and, as in EVSD, the remaining criteria are free to vary across participants. The scales of the new and studied-item distributions, $\lambda_{ij}^{(n)}$ and $\lambda_{ij}^{(s)}$, respectively, are free to vary across participants and items. Sensitivity in the gamma model may be defined as the ratio of studied to new-item scales $d'_{ij} = \lambda_{ij}^{(s)}/\lambda_{ij}^{(n)}$. Additive models are placed the log of scale parameters so that these scales are restricted to be positive:

$$\log(\lambda_{ij}^{(n)}) = \mu^{(n)} + \alpha_i^{(n)} + \beta_j^{(n)},$$

$$\log(\lambda_{ij}^{(s)}) = \mu^{(s)} + \alpha_i^{(s)} + \beta_j^{(s)} + \theta^{(s)} l_{ij}.$$

These random effects are given the same hierarchical structures shown in Eqs. (3)–(6).

Fig. 3(D) shows GSD ROC curves for several levels of sensitivity. As can be seen, the curves produce a similarly structured field as EVSD, however, they are asymmetric. If these lines were fit with UVSD, there would be a positive correlation between d' and σ^2. Such a relationship is often but not always observed in the literature (e.g., Glanzer, Kim, Hilford, & Adams, 1999; Ratcliff, Sheu, & Grondlund, 1992; Yonelinas & Parks, 2007).

It may seem that the gamma is a rather arbitrary choice for a signal detection model. Indeed it is. This insight, however, applies to the normal as well — there is no real reason to chose the normal distribution over alternatives (see Egan, 1975; Lockhart & Murdock, 1970; Rouder, Pratte, & Morey, 2010). We consider the gamma superior *a priori* to the normal on pragmatic grounds. It explains the asymmetry in the data while predicting constrained and orderly ROC fields. Therefore, it is a reasoned alternative worthy of study. The gamma model is similar to the extreme-value model proposed by DeCarlo (1998). DeCarlo's model also yields asymmetric ROC curves which form a orderly field. Furthermore, DeCarlo's model has a single parameter to describe the effect of study. In fact, the two models are so similar that each serves as a reasonable surrogate for the other.

2.5. Parameter recovery

The UVSD and DPSD models with participant and item effects on σ^2 and recollection, respectively, are novel developments that merit benchmarking. We use simulation analysis to explore parameter recoverability for both models. First, data were generated from the UVSD model with the same design matrix as the experiment presented below (97 people, 480 items at test). Grand means, random effect variances, and criteria were also the same as those estimated in the experiment (effects were randomly generated from their parent distributions). The results of this simulation are shown in Fig. 4. The top panels show estimates of participant effects on the new-item mean, studied-item mean, and log variance as a function of their true values. The bottom panels show the same for items. Clearly, with a large experiment the model performs extremely well.

The hierarchical DPSD model is benchmarked through simulation in the same manner. The results for new-item mean, studied-item mean, and recollection effects for participants are shown in the top panels of Fig. 5; the same is shown for items in the bottom panels. As with UVSD, the hierarchical DPSD model provides for accurate parameter recovery for this design. The model's performance in other designs (e.g., fewer participants or items) is quite good, but is not explored here, as it should be assessed through simulation on a case-by-case basis.

3. Model comparison

Fig. 2 shows the twelve models of interest. Selecting amongst them is an intellectually difficult exercise. Model-selection must take into consideration both how well a model accounts for the data, and how complex (or flexible) it is. One approach to quantifying complexity is to count parameters. Popular fit statistics such as AIC and BIC penalize models by the number of parameters they include. Unfortunately, these measures are not appropriate for hierarchical models as their constraint is not well-captured by counting parameters.

To understand how constraint is not captured by the number of parameters in hierarchical models, consider the simple linear model in Eq. (1). If no hierarchical structure is imposed on the effect parameters, then there are $1 + I + J$ parameters in the additive components. If the item and participant effects are treated as random effects and the variances in Eqs. (3) and (4) are treated as free parameters, then two new parameters have been added. These additional parameters, however, *decrease* the complexity of the model as they will constrain the effect parameters to be more similar to each other. In this very real sense, adding parameters to make a hierarchical structure, decreases rather than increases complexity.

Because hierarchical structures reduce model complexity by adding parameters, methods that rely on counting parameters are not appropriate. A better alternative in this context is the *deviance information criterion* (DIC, Spiegelhalter et al., 2002), a model selection statistic specifically designed for selecting among hierarchical Bayesian models estimated with MCMC sampling. The DIC statistic for a model is its *deviance* plus a penalty term for the number of its *effective parameters*, denoted *pD*. Deviance is a measure of how well the model accounts for the data (in Bayesian models, the posterior mean of a deviance distribution is calculated). The number of effective parameters is an estimate of how many *unconstrained* parameters there are in the model. This measure equals the true number of parameters when prior distributions are non-informative, and becomes smaller as hierarchical structures add constraint.

Although DIC is convenient and more appropriate than some other methods, it is not without faults and critiques. One of

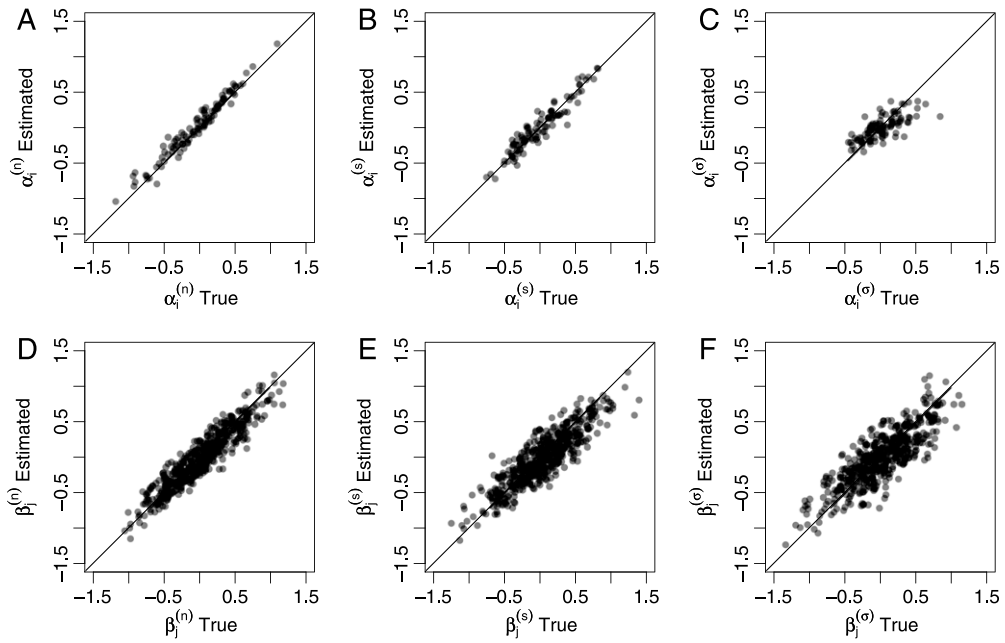

Fig. 4. Parameter recovery for the hierarchical unequal-variance signal detection model. Data were generated from the hierarchical UVSD model and these data were fit with the hierarchical UVSD model. Panels A, B, & C show estimated participant effects as a function of their true values for the new-item mean, studied-item mean, and variance σ^2, respectively. Panels D, E, & F show the same for items.

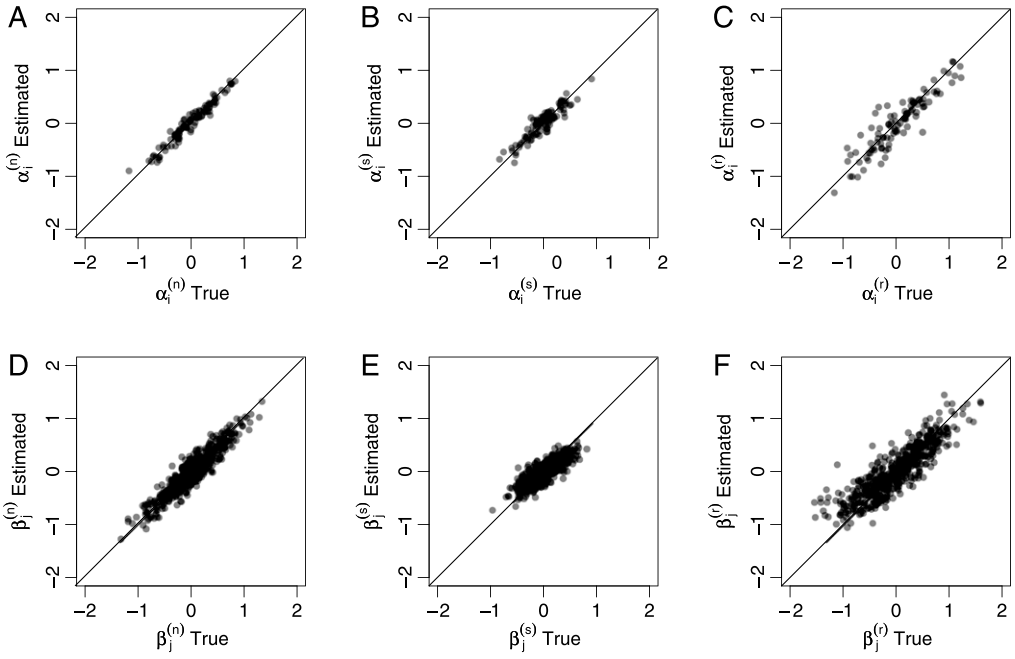

Fig. 5. Parameter recovery for the hierarchical dual-process signal detection model. Data were generated from the hierarchical DPSD model and these data were fit with the hierarchical DPSD model. Panels A, B, & C show estimated participant effects as a function of their true values for the new-item mean, studied-item mean, and recollection, respectively. Panels D, E, & F show the same for items.

our main concerns is that DIC is miscalibrated in that it tends to overstate the evidence toward more complex models. Like AIC and inference by *p*-values, this overstatement increases with sample size (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007). A more compelling model selection statistic would be Bayes factor (Kass & Raftery, 1995), but we know of no tractable method of performing the necessary integration for the models in Fig. 2.

One of our main goals in this paper is discriminating between UVSD and DPSD. We assessed how well this can be performed with DIC via simulation. Data were generated by both random-effects UVSD and DPSD models with the true values based on model fits to the subsequent experiment. Each simulated data set was fit with both UVSD and DPSD, and the DIC difference was calculated. Histograms of these differences are shown in Fig. 6. Clearly the DIC statistic accurately discriminates between data generated from the two models. Although this approach is not as extensive as the bootstrap assessment offered by Wagenmakers, Ratcliff, Gomez, and Iverson (2004), it provides confidence that these models are indeed discriminable with sample sizes characteristic of our data.

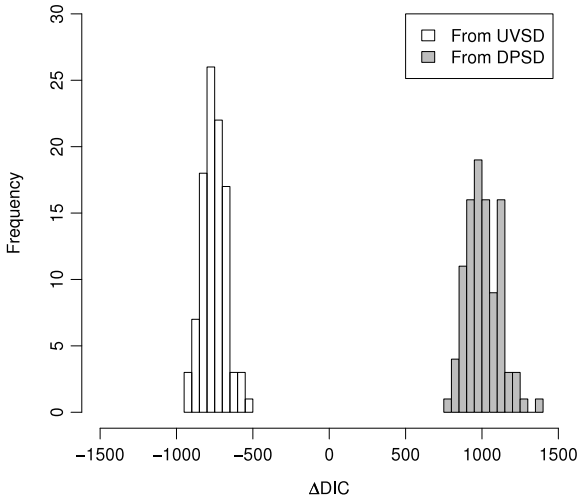

Fig. 6. Discriminating between UVSD and DPSD as generating model. Data were simulated from the hierarchical UVSD model (100 simulations) or the hierarchical DPSD model (100 simulations) with grand means and effect variances set to those estimated in the experiment. For each simulation the difference in DIC (ΔDIC) between the UVSD and DPSD model fits was calculated such that negative values indicate that UVSD provides a better account. The histogram of ΔDIC from these 200 simulations shows that when the data were generated from UVSD (white) DIC prefers the UVSD model; when the data were generated from DPSD (grey) DIC prefers the DPSD model.

Table 1
Model selection statistics.

Model			DIC Statistics		
Family	Effects	Restriction	DIC[a]	Deviance	Penalty (pD)
DPSD	Random	–	0	128732	1716
DPSD	Fixed	–	134	128633	1948
UVSD	Random	–	137	128864	1721
UVSD	Fixed	–	316	128727	2036
DPSD	Random	$R \propto d^{(n)}$	722	129838	1331
UVSD	Random	one σ^2	1009	130098	1359
GSD	Random	–	1016	130101	1363
EVSD	Random	$\sigma^2 = 1$	1799	130886	1361
DPSD	Random	no items	5300	135137	610
DPSD	Fixed	no items	5327	135135	639
UVSD	Random	no items	5344	135182	609
UVSD	Fixed	no items	5377	135163	662

[a] DIC is Deviance + Penalty - DIC for DPSD random effects model.

4. Experiment

The data used to assess the models is from a large-scale confidence-rating experiment reported by Pratte et al. (2010). In this experiment 97 participants studied 240 words by viewing them for 1850 ms and reading each aloud. Following study each participant was tested on the same 480-item set using a 6-point confidence rating scale. Half of the items were chosen randomly to be studied for each participant.

5. Results

The twelve models were fit using the methods discussed in the Appendix with the `hbmem` package in **R**. Table 1 shows DIC statistics for each model. The DIC value for each model is the difference between that model's DIC value and the DIC value for the general DPSD model with random effects. The following patterns are evident.

DPSD is selected over alternatives Every DPSD model was preferred to its corresponding UVSD counterpart. This trend holds regardless of restriction and regardless of whether effects are considered fixed or random. DPSD also outperformed the EVSD and gamma single-process alternatives.

The dependence of recollection on familiarity The general DPSD model with random effects provided the best account of the data. Fig. 7 shows how recollection and familiarity co-vary. Fig. 7(A) shows the scatter plot for participant effects on recollection as a function of participant effects on sensitivity, where the latter is given by $\alpha_i^{(s)} - \alpha_i^{(n)}$. Fig. 7(D) shows the same for items. As can be seen, there is a positive relationship. Fig. 7(B) and (E) show the scatter plots for recollection and new-item or baseline familiarity; Fig. 7(C) and (F) show the same for studied-item familiarity. Almost all of the relationship between sensitivity d' and recollection is reflected in baseline familiarity. Items that pre-experimentally have less familiarity tend to be recollected at higher rates. It is important to note that these correlations do not reflect influences from the priors. The prior structure assumes independence between parameters, and, in our experience, independent priors lead to attenuated estimates of correlations (see Rouder et al., 2007). Hence, if anything, true correlations are larger than those reported.

Based on these relationships, we fit a DPSD model in which recollection was a linear function of baseline familiarity:

$$\Phi^{-1}(R_{ij}) = \mu^{(r)} + \phi_\alpha \alpha_i^{(n)} + \phi_\beta \beta_j^{(n)},$$

where ϕ_α and ϕ_β are the linear slopes relating the new-item strength effects to the recollection effects for participants and items, respectively. The lines in the middle panels of Fig. 7 are from this restricted model with estimated slopes of $\phi_\alpha = -1.03$ and $\phi_\beta = -1.49$. Although this linear model provides a good account of the relationship, the DIC value (see Table 1) suggests that the decrement in fit is not offset by the gain in parsimony when compared to the general DPSD model. The conclusion is that although baseline familiarity and recollection are substantially related, a deterministic linear relationship is not sufficient.

Although the UVSD model provides an inferior account of the data, it is nevertheless interesting to ask whether UVSD parameters are correlated. In out previous work (Pratte et al., 2010) we show that for these data there is a positive relationship between baseline strength and studied-item strength across people (reflecting a response bias), and a negative correlation across items (reflecting a mirror effect). Fitting the UVSD model with σ^2 free to vary across people and items reveals a positive correlation between d' and σ^2 for both participants ($r = .33$, $t(95) = 3.35$, $p < .05$) and items ($r = .68$, $t(478) = 20.42$, $p < .05$). Accordingly, both the DPSD and UVSD model analyzes imply that ROC asymmetry (R and σ^2, respectively) is positively related to overall performance.

Random vs. fixed effects Models with random effects are selected over models with fixed effects. This trend holds for both DPSD and UVSD, and both when item effects are included and when they are excluded. The DIC statistics reveal that the advantage of the random-effect models is not in fit. Naturally, fixed-effect models fit better as they are less constrained. The DIC penalty terms reveal that the gain in parsimony for treating people and item effects as samples from parent distributions, however, more than offsets the loss in fit.

To better explicate the constraint in random effects modeling, we plotted participant and item effect estimates from the random-effects model as a function of those from the fixed-effects model. Fig. 8 shows the case for the UVSD model, and the displayed trends are similar for DPSD. The main difference between random and fixed effects is at the extremes — random effect estimates are less extreme than their fixed effect counterparts. This trend reflects the natural correlate of assuming these effects are from a common parent distribution with normal tails. The interpretation from the random effects model is that the extremes in fixed-effect

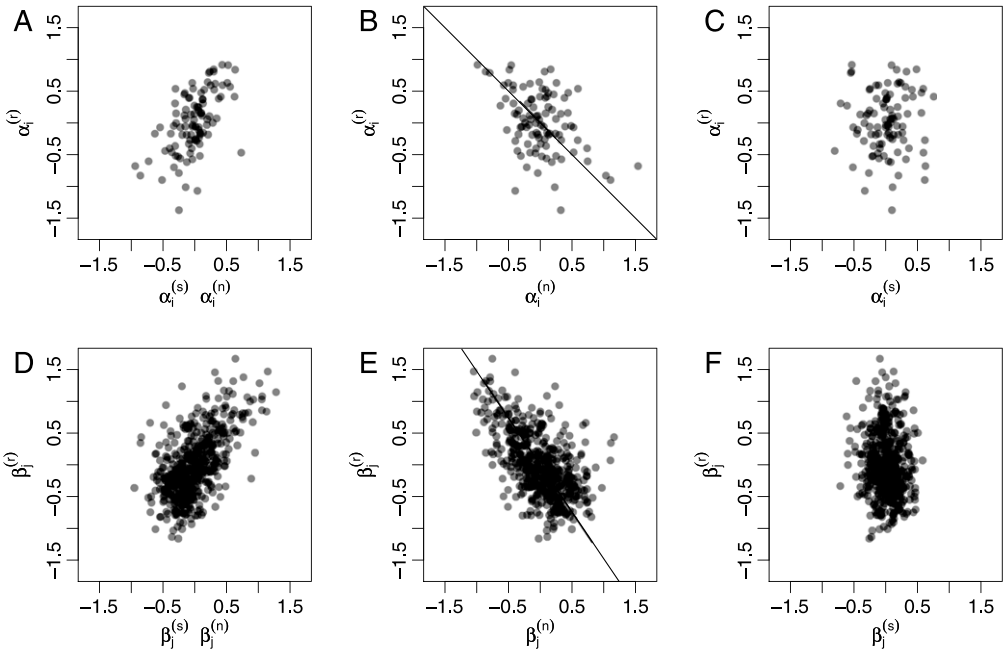

Fig. 7. Correlations between DPSD effects in recollection and familiarity. (A) Participant effects in recollection as a function of those in sensitivity d'. (B) Participant effects in recollection as a function of those in new-item (baseline) familiarity. The solid line is the estimated linear relationship from the DPSD model in which recollection is forced to be a linear function of baseline familiarity. (C) Participant effects in recollection as a function of those in studied-item familiarity. Panels D–F show the same relationships for item effects.

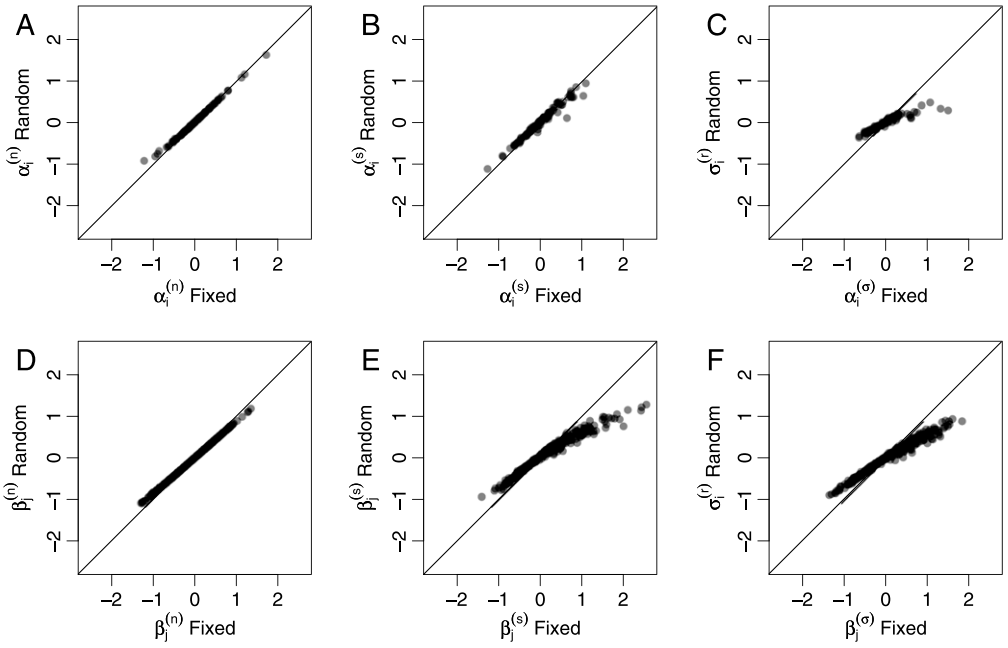

Fig. 8. Comparing fixed and random effect estimates from the UVSD fit to the experiment. (A) Participant effects on the new-item mean from the random-effects model as a function of those from the fixed-effects model. (B) Participant effects on the studied-item mean from the random-effects model as a function of those from the fixed-effects model. (C) Participant effects on $\log(\sigma^2)$ from the random-effects model as a function of those from the fixed-effects model. Panels D, E, & F show the same relationships for item effects.

estimates are elements of over-fitting. The constraint in random effects is expressed as a shrinkage of extreme estimates to group means. This shrinkage effect is greatest for item estimates (bottom row) because there are many more items than participants. The shrinkage is also greater for studied-item parameters than new-item parameters because the effect of study is modeled in two parameters ($d^{(s)}$ and σ^2) rather than in one ($d^{(n)}$).

The presence of item effects Two important questions remain concerning item effects: (1) should item effects be included in the models, and (2) is accounting for item effects driving the advantage of DPSD over UVSD? To answer both questions we fit fixed- and random-effect UVSD and DPSD models without any item effects or lag effects. These analyzes, which are comparable to averaging data over items, are shown in the last four rows of Table 1. Item effects are so prevalent that models without item effects perform worse than EVSD (with item effects). This last comparison is important because EVSD has been considered insufficient for recognition memory for over three decades.

Results for the models without item effects makes it evident that the inclusion of item effects is not driving the selection of DPSD over UVSD. Consider the comparison of the general models vs. the comparison of models without item effects. For the former, the DIC favors DPSD by 137; for the later, DIC favors DPSD by 44. The difference in these comparisons highlights the fact that one of the advantages of modeling items is that it facilitates model selection. When accounting for items, the data are richer (greater numbers of degrees of freedom) and the models are more complex. The balance, however, is that model misspecification of structural properties becomes more apparent and so model selection is easier.

In both UVSD and DPSD, study affects two parameters. In our previous work with UVSD (Morey, Pratte et al., 2008; Pratte et al., 2010), we implemented a single value of σ^2 for all participant-by-item conditions. The current research shows this choice is not optimal as the model with participant, item, and lag effects on σ^2 is selected over our previous model.

6. Conclusion

The results show some of the advantages of taking into account item and participant effects while modeling recognition memory performance. The substantive conclusion offered here is that the dual-process signal detection model outperforms the others. Even so, we are hesitant to fully embrace DPSD, and do not believe the evidence we provide is overwhelming. We offer the following caveats to aid in the judicious interpretation of our results.

Our focus in this paper has been on statistical modeling rather than on exploring the best manipulations to test theoretical positions. We would be more convinced of DPSD or UVSD if one could selectively influence key parameters, especially while accounting for participant and item variation. Several researchers have made convincing arguments that previous demonstrations of selective influence are methodologically or conceptually flawed (Dunn, 2008; Wixted, 2007; Yonelinas & Parks, 2007). We believe the hierarchical models developed here, when combined with appropriate manipulations, offer a principled and powerful approach to assessing selective influence.

A second caveat is rooted in the substantial relationships in DPSD parameter estimates. Recollection is correlated with baseline familiarity suggesting that familiarity and recollection may not be that distinct. There may be reasons for this dependence that retain the spirit of the dual-process model. Alternatively, however, the correlation may be diagnostic of a single-process structure. The challenge from this point of view is to explain why the model with the linear constraint fared worse than the general model. One possibility is that we simply have chosen the wrong transforms for regression. In our DPSD model, recollection parameters enter through a probit transform and familiarity is assumed to be normally distributed. Both of these choices are arbitrary and perhaps other choices would reduce the noise in the scatter plots of Fig. 8. The ability to detect misspecification in transform increases as the data span larger ranges in accuracy. Though these ranges are large in our experiments, they may be made even larger through manipulation.

A third caveat reflects the nature of DIC as a model selection statistic. DIC has, as a benefit, that it reflects constraint from hierarchical structure. One negative property is that it does not reflect this constraint in a consistent manner. As discussed previously, DIC does not appropriately penalize complexity as sample size increases. We feel that DIC is most useful for selecting across models of relatively comparable complexity, such as between comparable UVSD and DPSD models. We are most concerned about the DIC evaluation of nested models with large differences in effective parameters and with large sample sizes, as we have here. As noted, DIC is based on the same logic as AIC, and as a consequence, overstates the evidence against nested submodels. In this regard, we take the selection of the general DPSD model over the one where recollection is a linear function of familiarity with qualification. Similarly, we qualify the rejection of the gamma model via DIC — it too may reflect DIC's bias toward more complex models in cases with large sample sizes. Ideally, more consistent model selection techniques, such as Bayes factor or minimum description length (Grunwald, Myung, & Pitt, 2005), would be available for this inference.

Appendix

Here we provide full specification for each model, and overview how each is estimated.

A.1. Equal-variance signal detection model

For the EVSD model non-informative priors are placed on all grand means and effect variances:

$$\mu \sim \text{Normal}(0, \sigma_0^2),$$
$$\sigma^2_{effect} \sim \text{InverseGamma}(a, b),$$

where $\sigma_0^2 = 100$ and $a = b = .01$. Flat priors are placed on criteria, and criteria are constrained to order.

Sampling model parameters conditioned on the multinomial data is difficult. Instead, following Albert and Chib (1995) data augmentation is used by introducing latent variables w_{ij}. The mapping between the multinomial data and continuous latent data is:

$$(y_{ij} = k) \Longleftrightarrow (c_{i(k-1)} \leq w_{ij} < c_{ik}).$$

Latent variables w_{ij} are distributed as normals:

$$w_{ij} \sim \begin{cases} \text{Normal}(d_{ij}^{(n)}, 1), & \text{New}, \\ \text{Normal}(d_{ij}^{(s)}, 1), & \text{Studied}. \end{cases}$$

The conditional posterior distribution of w_{ij} is

$$w_{ij} \mid \cdot \sim \begin{cases} \text{TN}_{(c_{i(y_{ij}-1)}, c_{i(y_{ij})})}(d_{ij}^{(n)}, 1), & \text{New}, \\ \text{TN}_{(c_{i(y_{ij}-1)}, c_{i(y_{ij})})}(d_{ij}^{(s)}, 1), & \text{Studied}, \end{cases}$$

where $\text{TN}_{(a,b)}(\mu, \sigma^2)$ is a Normal(μ, σ^2) distribution truncated below at a and above at b.

Once w_{ij} are sampled, sampling the grand means and effect parameters in Eq. (1)–(2) is simply a matter of sampling standard linear model parameters conditioned on the latent normal data. There are several techniques for doing so that minimize autocorrelation, including blocked sampling and piecewise sampling with Metropolis-Hastings decorrelating steps (see Gelman, Carlin, Stern, & Rubin, 2004). Here we use the latter option as it is faster for large experiments. After sampling the linear components the criteria may be sampled. Sampling the criteria conditioned on the latent data leads to a great deal of autocorrelation. This autocorrelation may be mitigated instead by sampling the criteria conditioned on the multinomial data with Metropolis–Hastings sampling. Details of sampling the criteria can be found in Morey, Pratte et al. (2008).

A.2. Unequal-variance signal detection model

Sampling from the posterior of the UVSD model is the same as sampling from the EVSD model, with the following exceptions:

$$w_{ij} \sim \begin{cases} \text{Normal}(d_{ij}^{(n)}, 1), & \text{New}, \\ \text{Normal}(d_{ij}^{(s)}, \sigma_{ij}^2), & \text{Studied}. \end{cases}$$

$$w_{ij} \mid \cdot \sim \begin{cases} \text{TN}_{(c_{i(y_{ij}-1)}, c_{i(y_{ij})})}(d_{ij}^{(n)}, 1), & \text{New}, \\ \text{TN}_{(c_{i(y_{ij}-1)}, c_{i(y_{ij})})}(d_{ij}^{(s)}, \sigma_{ij}^2), & \text{Studied}. \end{cases}$$

Sampling the linear components of the new-item means conditioned on w_{ij} is straightforward. To our knowledge, however, the additive model on log variance in Eq. (7) has not been developed previously. Let $\nu_{ij}^2 = \log(\sigma_{ij}^2) = \mu^{(\nu)} + \alpha_i^{(\nu)} + \beta_j^{(\nu)} + \theta^{(\nu)} l_{ij}$. The log conditional posterior distribution of the linear components on the studied-item mean and variance is:

$$\begin{aligned} &\frac{-1}{2}\Bigg[\sum_{i=0}^{I}\sum_{j=0}^{J}\Bigg(\mu^{(\nu)} + \alpha_i^{(\nu)} + \beta_j^{(\nu)} + \theta^{(\nu)} l_{ij} \\ &\quad + \frac{\left(w_{ij} - \mu^{(s)} - \alpha_i^{(s)} - \beta_j^{(s)} - \theta^{(s)} l_{ij}\right)^2}{e^{(\mu^{(\nu)} + \alpha_i^{(\nu)} + \beta_j^{(\nu)} + \theta^{(\nu)} l_{ij})}}\Bigg) \\ &\quad + \frac{(\mu^{(s)})^2}{\sigma_0^2} + \frac{(\alpha_i^{(s)})^2}{\sigma_{\alpha,s}^2} + \frac{(\beta_j^{(s)})^2}{\sigma_{\beta,s}^2} + \frac{(\theta^{(s)})^2}{\sigma_0^2} + \frac{(\mu^{(\nu)})^2}{\sigma_0^2} \\ &\quad + \frac{(\alpha_i^{(\nu)})^2}{\sigma_{\alpha,\nu}^2} + \frac{(\beta_j^{(\nu)})^2}{\sigma_{\beta,\nu}^2} + \frac{(\theta^{(\nu)})^2}{\sigma_0^2}\Bigg]. \end{aligned}$$

The grand mean and effects on the the mean $d^{(s)}$ have normal conditional posteriors:

$$f(\mu^{(s)}|\ldots) \sim \text{N}(m, v)$$

$$v = \left(\sum_{i=0}^{I}\sum_{j=0}^{J}(e^{-\nu_{ij}^2}) + \frac{1}{\sigma_0^2}\right)^{-1}$$

$$m = v\sum_{i=0}^{I}\sum_{j=0}^{J}\frac{w_{ij} - \alpha_i^{(s)} - \beta_j^{(s)} - \theta^{(s)} l_{ij}}{e^{\nu_{ij}^2}}$$

$$f(\alpha_i^{(s)}|\ldots) \sim \text{N}(m, v)$$

$$v = \left(\sum_{i=0}^{I}\sum_{j=0}^{J}(e^{-\nu_{ij}^2}) + \frac{1}{\sigma_{\alpha,s}^2}\right)^{-1}$$

$$m = v\sum_{i=0}^{I}\sum_{j=0}^{J}\frac{w_{ij} - \mu^{(s)} - \beta_j^{(s)} - \theta^{(s)} l_{ij}}{e^{\nu_{ij}^2}}$$

$$f(\beta_j^{(s)}|\ldots) \sim \text{N}(m, v)$$

$$v = \left(\sum_{i=0}^{I}\sum_{j=0}^{J}(e^{-\nu_{ij}^2}) + \frac{1}{\sigma_{\beta,s}^2}\right)^{-1}$$

$$m = v\sum_{i=0}^{I}\sum_{j=0}^{J}\frac{w_{ij} - \mu^{(s)} - \alpha_i^{(s)} - \theta^{(s)} l_{ij}}{e^{\nu_{ij}^2}}$$

$$f(\theta^{(s)}|\ldots) \sim \text{N}(m, v)$$

$$v = \left(\sum_{i=0}^{I}\sum_{j=0}^{J}(l_{ij}^2 e^{-\nu_{ij}^2}) + \frac{1}{\sigma_0^2}\right)^{-1}$$

$$m = v\sum_{i=0}^{I}\sum_{j=0}^{J}\frac{l_{ij}(w_{ij} - \mu^{(s)} - \alpha_i^{(s)} - \beta_j^{(s)})}{e^{\nu_{ij}^2}}.$$

Conditional posteriors of the grand mean and effects on ν^2 do not have conditionals with known forms and so are sampled independently with random-walk Metropolis-Hastings algorithms. The effects on $d^{(s)}$ and on ν^2 are given Metropolis-Hastings decorrleating steps.

A.3. Dual-process signal detection model

As with EVSD, sampling from DPSD is made easier by conditioning on latent data rather than the multinomial data. For the new-item condition, latent data are sampled exactly as they are in the hierarchical EVSD model:

$$\omega_{ij}^{(n)}|\ldots \sim \text{TN}_{(c_{i(y_{ij}-1)}, c_{i(y_{ij})})}(d_{ij}^{(n)}, 1).$$

In addition to sampling latent data for the signal detection components, latent data are sampled for recollection in the manner common for estimating linear models on binomial probabilities (Albert & Chib, 1995). These latent data, denoted $\omega_{ij}^{(r)}$, are positive on trials for which recollection occurred and negative otherwise. On trials for which the response was not equal to K (i.e., not "sure studied") recollection assuredly did not occur and so the response must be the product of the signal detection component. For these trials the mapping between multinomial and latent data is as follows:

$$\left.\begin{aligned} \omega_{ij}^{(s)}|\ldots &\sim TN_{(C_{i(y_{ij}-1)}, C_{i(y_{ij})})}\left(d_{ij}^{(s)}, 1\right) \\ \omega_{ij}^{(r)}|\ldots &\sim TN_{(-\infty,0)}\left(\phi^{-1}(R_{ij}), 1\right) \end{aligned}\right\} y_{ij} \neq K.$$

Responses to studied items equal to K could have arisen from recollection or from familiarity that was above the $K-1$ criterion. If recollection occurred (i.e., $\omega_{ij}^{(r)} \geq 0$) then we have no information about familiarity. Alternatively, if recollection did not occur ($\omega_{ij}^{(r)} < 0$) then familiarity must have been above the $K-1$ criterion:

$$\left.\begin{aligned} \omega_{ij}^{(s)}|\ldots &\sim N\left(d_{ij}^{(s)}, 1\right), \omega_{ij}^{(r)} \geq 0 \\ \omega_{ij}^{(s)}|\ldots &\sim TN_{(C_{i(K-1)}, \infty)}\left(d_{ij}^{(s)}, 1\right), \omega_{ij}^{(r)} < 0 \end{aligned}\right\} y_{ij} = K.$$

In a similar manner, if familiarity is above the highest criterion then we have no information about whether recollection occurred or did not. Alternatively, if familiarity is below the $K-1$ criterion and yet the response is K, then recollection must have occurred:

$$\left.\begin{aligned} \omega_{ij}^{(r)}|\ldots &\sim N\left(\phi^{-1}(R_{ij}), 1\right), \omega_{ij}^{(s)} \geq C_{i(K-1)} \\ \omega_{ij}^{(r)}|\ldots &\sim TN_{(0,\infty)}\left(\phi^{-1}(R_{ij}), 1\right), \omega_{ij}^{(s)} < C_{i(K-1)} \end{aligned}\right\} y_{ij} = K.$$

The latent data $\omega_{ij}^{(n)}$, $\omega_{ij}^{(s)}$, and $\omega_{ij}^{(r)}$ have marginal normal distributions with unit variance and means $d_{ij}^{(n)}$, $d_{ij}^{(s)}$, and $\phi^{-1}(R_{ij})$, respectively. Sampling the additive components conditioned on these normally-distributed data is straightforward. Criteria are sampled with a Metropolis-Hastings algorithm as they are in the EVSD model.

A.4. Gamma signal detection model

Latent data for GSD are posited exactly as in EVSD except that they are distributed as gamma distributions with shape 2:

$$w_{ij} \sim \begin{cases} \text{Gamma}(2, \theta_{ij}^{(n)}), & \text{New}, \\ \text{Gamma}(2, \theta_{ij}^{(s)}), & \text{Studied}. \end{cases}$$

and with conditional posteriors that are truncated gammas rather than truncated normals. We present here the conditional log posterior distribution of the components for the studied-item

scales. Those for the new-item scales are equivalent, and criteria are sampled as they are in EVSD.

$$\sum_{i=0}^{I}\sum_{j=0}^{J}\left(\frac{-w_{ij}^{(s)}}{e^{\left(\mu^{(s)}+\alpha_i^{(s)}+\beta_j^{(s)}+\theta^{(s)}l_{ij}\right)}}-2\left(\mu^{(s)}+\alpha_i^{(s)}+\beta_j^{(s)}+\theta^{(s)}l_{ij}\right)\right)$$
$$-\frac{1}{2}\left(\frac{(\mu^{(s)})^2}{\sigma_0^2}+\frac{(\alpha_i^{(s)})^2}{\sigma_{\alpha,s}^2}+\frac{(\beta_j^{(s)})^2}{\sigma_{\beta,s}^2}+\frac{(\theta^{(s)})^2}{\sigma_0^2}\right)$$

References

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

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

Dunn, J. C. (2008). The dimensionality of the remember-know task: a state-trace analysis. *Psychological Review*, *115*(2), 426–446.

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

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

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley, Reprinted by Krieger, Huntington, NY, 1974.

Grunwald, P., Myung, I. J., & Pitt, M. A. (2005). *Advances in minimum description length: theory and applications*. Cambridge: MIT Press.

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.

James, W. (1890). *Principles of psychology (Volume I)*. New York: Holt.

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

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 1–26.

Lee, P. M. (1997). *Bayesian statistics: an introduction*. New York: Wiley.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*, 100–109.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin and Review*, *14*, 858–865.

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.

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

Pooley, J. P., Lee, M. D., & Shankle, W. R. (2011). Understanding memory impairment with memory models and hierarchical Bayesian analysis. *Journal of Mathematical Psychology*, *55*(1), 47–56.

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

Ratcliff, R., 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 and Review*, *12*, 573–604.

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

Rouder, J. N., 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., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*, 5976–5979.

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.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237.

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

Schacter, D. L. (1990). *Perceptual representation systems and implicit memory: toward a resolution of the multiple memory systems debate*. Hillsdale, NJ: Erlbaum.

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

Tulving, E., & Craik, F. I. M. (2000). *Oxford handbook of memory*. New York: Oxford.

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

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

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.