Some-or-none recollection: Evidence from item and source memory

Serge V. Onyper
Department of Psychology
St. Lawrence University

Yaofei Zhang, and Marc W. Howard
Department of Psychology
Syracuse University

In press, *Journal of Experimental Psychology: General* November 11, 2009

Abstract

Dual-process theory hypothesizes that recognition memory depends on two distinguishable memory signals. Recollection reflects conscious recovery of detailed information about the learning episode. Familiarity reflects a memory signal that is not accompanied by a vivid conscious experience but nonetheless enables participants to distinguish recently-experienced probe items from novel ones. This dual-process explanation of recognition memory has gained wide acceptance among cognitive neuroscientists and some cognitive psychologists. Nonetheless, its difficulty in providing a quantitatively satisfactory description of performance in item recognition experiments has precluded a consensus not only about the theoretical structure of recognition memory but also about how to best measure recognition accuracy. In two experiments we show that neither the standard formulation of dual-process signal detection theory (DPSD) nor a widely-used single-process model (UVSD) provides a satisfactory explanation of recognition memory across different types of study materials (words and travel scenes). In the variable recollection dual-process model (VRDP), recollection fails for some old probe items, as in standard formulations of dual process signal detection theory, but gives rise to a continuous distribution of memory strengths when it succeeds. The VRDP can approximate both the DPSD and the UVSD. In both experiments it provides a consistently superior fit across materials to the *superset* of the DPSD and UVSD. The VRDP offers a simple explanation of the form of conjoint item-source judgments, something neither the DPSD nor the UVSD can accomplish. The success of the VRDP supports the core assumptions of dual-process theory by providing an excellent quantitative description of recognition performance across materials, response criteria and type of response.

Recognition memory is an essential human cognitive ability that enables one to identify stimuli—people, places and objects—as having been previously experienced. Much of recent theory on the subject of recognition memory centers on what is known as two-process theory, which hypothesizes that recognition relies on two separable memory signals, referred to as recollection and familiarity. A common example should suffice to illustrate the distinction between these two components. All of us have had the experience of a chance encounter with a person who seems like someone we have met, but who we cannot place. After a moment's search of memory, we may succeed in recovering a detailed memory of the person next to you at the checkout counter and remember, say, that she was the mother of one of the guests at your daughter's recent birthday party. The flood of specific, vivid, episodic memories of the birthday party are referred to as recollection. The feeling of knowing that this person is someone you've previously encountered, in the absence of detailed memories for a specific event, is referred to as familiarity. Notably, we have all had the experience that the memory search fails, i.e., familiarity without recollection. This illustrates a key property that is central to theorizing about recollection—that it does not succeed for all probes of memory.

Recognition memory is studied in the laboratory by presenting participants with a list of to-be-remembered stimuli and then presenting them with a test list of probes. Performance is expressed as a hit rate—the proportion of targets successfully identified as part of the list—and false alarm rate—the proportion of lures incorrectly identified as part of the list. Taken in isolation, neither hit rate nor false alarm rate tell us anything about accuracy. Ideally, our measurement of discriminability should not change with changes in the response bias. Any attempt to measure recognition discriminability, let alone the separate contributions of recollection and familiarity is dependent on some model of the relationship between hit rate and false alarm rate across multiple levels of response bias.

Recognition accuracy is often measured across levels of response bias using multiple confidence ratings. Rather than making a binary response to a problem, the participant expresses confidence that a probe item was on the list using, for instance, a seven-point scale. The experimenter can then calculate hit rates and false alarm rates for six different criteria. For instance, at the most stringent criterion we would count only responses of "7" as yes ratings and calculate a hit rate and a false alarm rate. One then repeats the analysis counting "6" or "7" responses as yes ratings and obtains a new hit rate and false alarm rate. Continuing this process across all possible criteria results in six pairs of hit rates and false alarm rates. Plotting the hit rates as a function of the false alarm rates yields a receiver operating characteristic (ROC) curve. The debate about the form of ROC curves has been extremely active, and sometimes heated, in recent years. The discussion has also had implications for memory theory, in particular the question of whether recognition is really subserved by recollection and familiarity or whether it is best described as a single process.

Address correspondence to Serge Onyper, sonyper@stlawu.edu, or Marc Howard, marc@memory.syr.edu.
Supported by NIH grant 1-R01 MH069938 to MWH. Experiment 1 was part of Yaofei Zhang's M.S. thesis at Syracuse University. Experiment 2 was part of Serge Onyper's Ph.D. thesis at Syracuse University. This paper benefitted from discussions with and/or comments from Michael Kahana, Brandy Bessette-Symons, and Caren Rotello.

This debate has broad implications. Two process theory has captured the imagination of much of the cognitive neuroscience community studying memory. Results implicating two distinct physical sources for recognition include studies using scalp EEG (e.g. Duzel, Yonelinas, Mangun, Heinze, & Tulving, 1997; Paller, Hutson, Miller, & Boehm, 2003; Rugg & Curran, 2007), fMRI (e.g. Uncapher & Rugg, 2005; Yovel & Paller, 2004), human neuropsychology studies (e.g. Aggleton et al., 2005; Holdstock et al., 2002; Verfaillie & Treadwell, 1993; Yonelinas et al., 2002), as well as studies of recognition memory in non-human species (Fortin, Wright, & Eichenbaum, 2004; Good, Barnes, Staal, McGregor, & Honey, 2007). Yonelinas (1994, 2002) has developed a specification of dual-process theory, which we will refer to as dual-process signal detection theory (DPSD) that makes predictions about the shape of ROC curves. This work has been the subject (both pro and con) of several high-impact publications in recent years (e.g. Fortin et al., 2004; Sauvage, Fortin, Owens, Yonelinas, & Eichenbaum, 2008; Wais, Wixted, Hopkins, & Squire, 2006; Yonelinas, Otten, Shaw, & Rugg, 2005). However, the DPSD has typically provided a worse quantitative description of empirically-observed ROC curves than the unequal variance signal detection model (UVSD), a competitor model that has typically been identified with single process theory (but see Wixted, 2007a). In this paper we show that neither the DPSD nor the UVSD provide a satisfactory account of the results across different stimulus materials. We reject each of them as a general description of recognition accuracy across different levels of response bias. We show that the variable recollection dual process model (VRDP), a slightly different quantitative implementation of dual-process theory, provides a superior account of the data to the *superset* of the DPSD and UVSD. Before describing the VRDP, we describe the UVSD and the DPSD in some detail.

Two-parameter signal detection models

The UVSD and DPSD both inherit the basic assumption of signal detection theory. Each memory probe generates a continuous decision variable, which we can think of as the strength of evidence that the item was on the list. This strength is compared to a criterion to determine whether to respond "yes" or "no" to the probe. An ROC is generated by assuming that there are multiple criteria used to select the response. The UVSD and DPSD differ in their assumptions about the properties of the distribution of strength induced by old probes.

Figure 1, top row, illustrates typical response distributions, ROC curves and z-transformed ROC curves (zROC) for the UVSD. The UVSD presumes that the distributions of strengths for both old and new items are described by normal distributions. Two parameters control these distributions. The distance between the old and new distributions, in units of the standard deviation of the new item distribution, is described by d'_F. The standard deviation of the old item distribution is controlled by another parameter, σ_F. This second parameter enables the UVSD to generate ROCs that are asymmetric around the diagonal. The UVSD generates the hit rate for the kth criterion using the following equation,

$$P(\text{`yes'} \leq k \,|\, \text{old}) = \Phi\left(c_k, -d'_F/2, \sigma_F\right), \tag{1}$$

where c_k is the response criterion for rating k and $\Phi\left(c, \mu, \sigma\right)$ is the integral of the cumulative probability density function of a normal distribution with mean μ and standard deviation

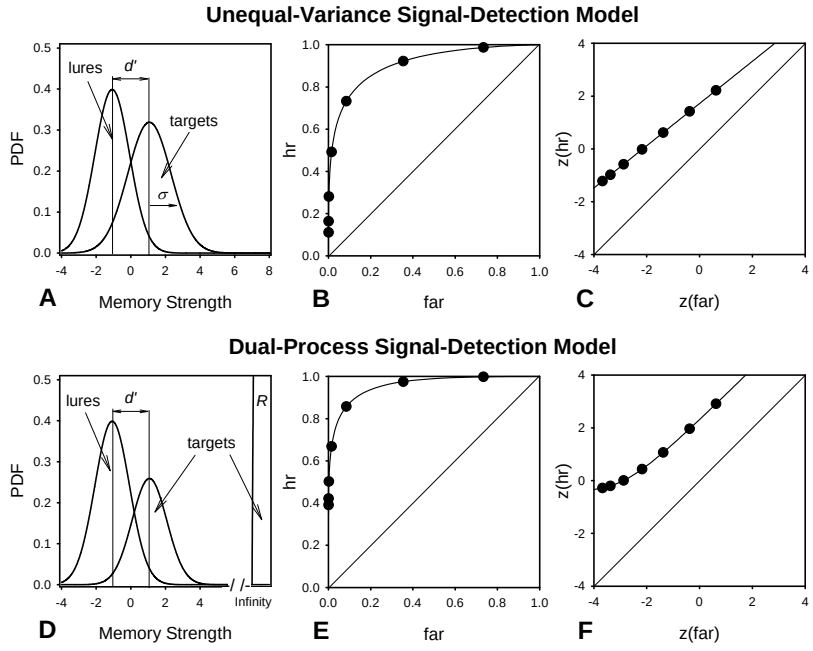

Figure 1. Probability density functions (PDF) and receiver operating characteristics (ROC) and z-transformed ROC (zROC) curves generated from the unequal-variance signal detection model (UVSD, top) and the dual-process signal detection model (DPSD, bottom). **A**,**D**. Model PDFs. **B**,**E**. ROC curve generated by the models. **C**,**F**. zROCs.

σ over the range from $-\infty$ to c.[1] In the UVSD, the false alarm rate for the kth criterion is given by:

$$P\left(\text{`yes'} \leq k \,|\, \text{new}\right) = \Phi\left(c_k, d'_F/2, 1\right). \tag{2}$$

Note that the sign of the mean has changed relative to the old item distribution and that the standard deviation of the new item distribution is one rather than σ_F.

The major difference between the UVSD and the DPSD is that the latter describes the variability in the old item distribution as arising from the fact that a subset of the old items are recollected whereas recollection fails to provide any useful information for the remainder of the old items. The DPSD assumes that familiarity is an equal-variance signal detection process that contributes to the discriminability of non-recollected items. Figure 1, bottom row, illustrates typical response distributions, ROCs and zROCs for the DPSD.

In the DPSD, the hit rate is given by

$$P(\text{`yes'} \leq k \,|\, \text{old}) = R + (1 - R)\,\Phi\left(c_k, -d'_F/2, 1\right) \tag{3}$$

[1]Note that increasingly stringent criteria would be ordered from left to right in Figure 1, but increasingly stringent criteria are given by decreasing numbers in the Equation 1. This accounts for the somewhat counterintuitive negative mean for the old item distribution and positive mean for the new item distribution. This is done here to be consistent with prior treatments of the models and the standard definition of $\Phi()$.

while the equation for the false alarm rate is:

$$P\left(\text{`yes'} \leq k \,|\, \text{new}\right) = \Phi\left(c_k, d'_F/2, 1\right) \tag{4}$$

Comparing the equation for the false alarm rate in the DPSD (Eq. 4) to that for the UVSD (Eq. 2), we see that they are identical. However, the DPSD's equation for the hit rate has at least two important differences from the analogous equation for the UVSD. First, the DPSD fixes σ_F at one. Second, this degree of freedom is replaced by the parameter R that describes the proportion of recollected items. Note that in Eq. 3, recollected items receive a yes response that is independent of the response criterion. That is, according to Eq. 3, recollected items recover a strength of evidence that exceeds every response threshold, resulting in a highest-confidence response for every old probe that is recollected. Items that are not recollected (which happens with probability $1 - R$) get no advantage from recollection. This all-or-none property of Eq. 3 is not a necessary theoretical claim of the DPSD (see e.g., Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996; Parks & Yonelinas, 2007a). It is nonetheless the case that previous Eq. 3 has been widely used in applications of the DPSD and exhibits all-or-none recollection for old items. To anticipate, the VRDP differs from the DPSD solely by relaxing the all-or-none property of Eq. 3. This more detailed description of recollection is consistent with dual process theory more broadly, as well as neurocomputational models of recollection (Elfman, Parks, & Yonelinas, 2008; Norman & O'Reilly, 2003).

As can be seen from the middle column of Figure 1, the UVSD and DPSD produce very similar ROC curves. However, the shape of the ROC curves that the two models induce are not identical, which can be seen more clearly when examining their zROC curves. The UVSD produces linear zROC curves with intercept d'_F and slope $1/\sigma_O$. In contrast, the DPSD generates non-linear zROCs. The asymptotic slope of the DPSD's zROC at extremely liberal criteria (to the right) is unity, due to the fact that the DPSD does not allow the standard deviation of the familiarity distribution to vary. The asymptotic slope of the DPSD's zROC at extremely conservative criteria (to the left of the figure) is zero. This is a consequence of the property of Eq. 3 that recollected items always receive a yes response no matter how stringent the response criterion. In the DPSD, the parameter d'_F controls the intercept of the liberal asymptote. The parameter R controls how far to the left the zROC deviates from the liberal asymptote.

Despite the intuitive appeal of dual-process theory, the verdict from quantitative analyses and examination of the shape of empirically observed zROCs has most often favored the UVSD, although the evidence is somewhat mixed (see below). Dozens of studies have described zROCs as linear with a slope of slightly less than unity (Diana, Reder, Arndt, & Park, 2006; Glanzer, Adams, Iverson, & Kim, 1993; Glanzer, Kim, Hilford, & Adams, 1999; Healy, Light, & Chung, 2005; Heathcote, 2003; Ratcliff, Sheu, & Gronlund, 1992; Rotello, Macmillan, & Reeder, 2004; Slotnick & Dodson, 2005; see Wixted, 2007a for a recent review of the empirical evidence in favor of the UVSD over the DPSD).

In addition to the findings from item recognition, which have tended to support the UVSD, another line of research that has been used to criticize the the DPSD comes from conjoint item-source ratings. In source recognition experiments, participants respond on the basis of the detailed quality of their memory for a probe. For instance, the participant may be presented with a list of words, half of which are presented in a female voice and

half of which are presented in a male voice. A source judgment at test would require participants to select the voice in which a particular probe item was presented. In conjoint item-source ratings (Yonelinas, 1999; Slotnick & Dodson, 2005) participants first give an old-new item rating to each probe, followed by a source rating. The key finding that raises potential problems for the DPSD is the finding that the source ROC calculated for items that have received a highest-confidence rating is curvilinear and closely approximates an equal-variance signal detection model. This is discrepant from the linear ROC one would expect from an all-or-none recollective signal. Although Parks and Yonelinas (2007a) argue that the key data can be addressed within the DPSD by assuming that both recollection and familiarity can support source judgments, they have not published predictions of conjoint item-source ratings illustrating that this assumption is sufficient to provide a satsifactory account of the data. We show that dual-process theory can easily account for these data if the all-or-none property exhibited by Eq. 3 is relaxed.

Most authors have taken the evidence against the DPSD as contradicting the assumption that there are two distinct processes, rather than one variable process, underlying recognition memory (e.g., Glanzer et al., 1999; Heathcote, 2003; Slotnick & Dodson, 2005). Recently, however, Wixted (2007a) argued that two-process theory can be described by the UVSD model if there is some degree of recollection for every item and recollection and familiarity combine in an additive fashion. If familiarity and recollection are both described as normal distributions, then their sum is also a normal distribution, with the standard deviation of the distribution of the sum equal to the sum of the standard deviations of the underlying distributions. Because recollection only succeeds for old items, the result is a normal old item distribution with a greater variance than the new item distribution. This dual-process interpretation of the UVSD radically alters how recollection is conceived by assuming that all old items receive some boost from recollection.

Of course this reinterpretation of dual-process theory is unnecessary if it turns out that the UVSD fails to provide a satisfactory description of the empirical data. While the majority of the zROC curves that have been reported are described as linear, a number of studies have reported zROCs with significant deviations from linearity, consistent with the predictions of the DPSD model (Fortin et al., 2004; Howard, Bessette-Symons, Zhang, & Hoyer, 2006; Sherman, Atri, Hasselmo, Stern, & Howard, 2003; Yonelinas, 1999). This discrepancy may be attributable to study materials; findings that report linear zROCs have almost exclusively used words as stimuli, whereas studies that find curvilinear zROCs have usually used other kinds of materials (Howard et al., 2006 and Sherman et al., 2003 used travel scenes; Fortin et al., 2004 used odors).[2]

There is evidence to suggest that in some instances both the UVSD model and the DPSD model fall short of adequately characterizing recognition accuracy. In order to fit subtle deviations from linearity in zROC curves from item recognition studies, DeCarlo (2002) proposed a mixture signal detection model (see also DeCarlo, 2003a; Hilford, Glanzer, Kim, & DeCarlo, 2002). DeCarlo (2002) explained the mixture of the old item distributions as a consequence of discrete modulation of attention during encoding, such that old items are either encoded into a high-attention state or a low-attention state. The discontinuity in encoding results in a model in which the old item distribution is described as two distinct

[2]Unpublished secondary analyses of the travel scene recognition data in Schwartz et al., (2005) also show non-linear zROC curves.

normal distributions. Similarly, Sherman et al. (2003) found that neither the UVSD nor the DPSD could adequately describe the performance of participants administered the cholinergic antagonist scopolamine. The zROC curve for patients administered scopolamine was nonlinear, but inconsistent with that predicted by the DPSD, with a pronounced "kink" in the middle. To accommodate this result, Sherman et al. (2003) implemented a variant of the DSPD model that included a graded recollection component. A similar some-or-none model has been proposed for associative recognition (Kelley & Wixted, 2001; Macho, 2004, see Yonelinas & Parks, 2007 for a recent review).

The VRDP: a model of some-or-none recollection

Here we argue that a some-or-none dual process model, which we will refer to as the variable-recollection dual-process model (VRDP), provides a more general solution to describing recognition accuracy than either the UVSD or the DPSD. The VRDP model builds upon the dual-process framework of Yonelinas (1994). However, unlike Eq. 3, when an old item is recollected, it does not necessarily receive a highest-confidence response. Instead, recollected items give rise to a distribution of strengths of evidence which are then compared to the response criteria. When considering item recognition in isolation, the VRDP model is mathematically equivalent to mixture signal detection (DeCarlo, 2002). However, unlike the explanation of DeCarlo (2002), which is relatively noncommittal about the psychological source of the two old item distributions, the VRDP makes the strong prediction that two old item distributions being mixed correspond to familiarity and recollection, with correspondingly different qualitative properties. These properties lead to strong predictions when one considers conjoint item-source ratings (this is discussed in detail below).

The VRDP has three key parameters: d'_F, d'_R, and R (plus, of course, criteria). The familiarity distribution is characterized by d'_F (the distance from the lure distribution in the units of the common standard deviation), with a standard deviation of 1. The probability that an old probe is recollected is given by R. The center of the recollective distribution in relation to the familiarity distribution is indexed by d'_R. The equation for the hit rate is given by:

$$P(\text{`yes'} \leq k \,|\, \text{old}) = R\,\Phi\left(c_k, -d'_F/2 - d'_R, 1\right) \;+\; (1-R)\;\Phi\left(c_k, -d'_F/2, 1\right) \tag{5}$$

while the equation for the false alarm rate is:

$$P(\text{`yes'} \leq k \,|\, \text{new}) = \Phi(c_k, d'_F/2, 1) \tag{6}$$

Comparing these to the Equations for the DPSD (Eqs. 3 and 4) we see that the sole difference is that the first term in the hit rate equation includes a distribution, weighted by R, that must be compared to the criterion. That is, rather than recollected items receiving a highest confidence rating, recollected items generate a continuous strength of evidence drawn from a normal distribution. This value is then compared to the response criteria. As a consequence, recollected items are compared to the response criteria and give rise to a range of responses rather than simply receiving a highest-confidence response as in Eq. 3.

The VRDP model can closely approximate both the UVSD and the DPSD for appropriate choices of parameters (Figure 2). When d'_R is small, the old recollective and familiarity distributions overlap considerably (Figure 2A). The ROC generated as a result

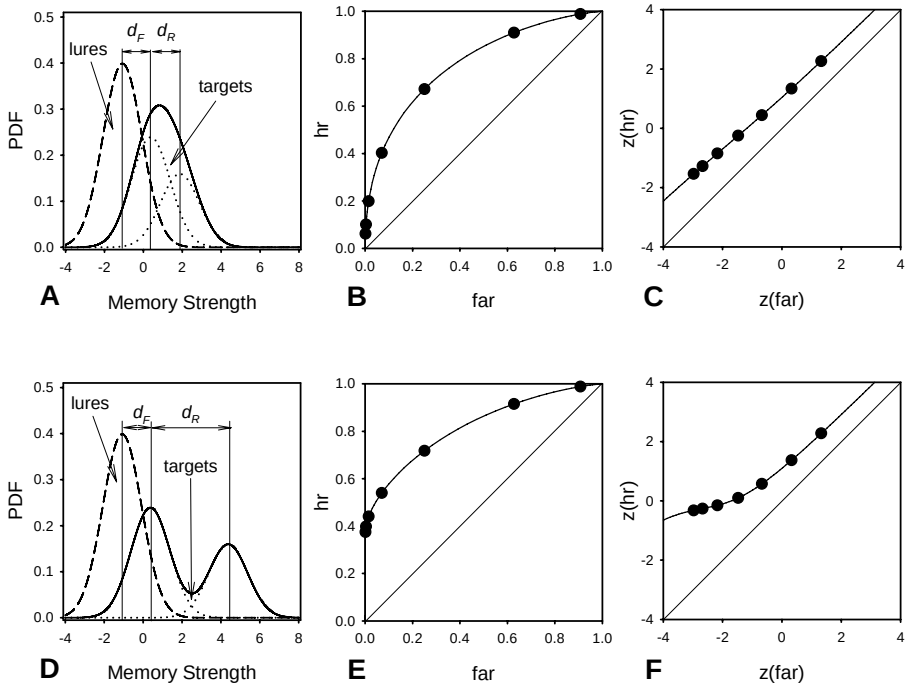

Figure 2. The variable recollection dual-process (VRDP) model can approximate the properties of both the unequal variance signal detection model (UVSD) and the dual-process signal detection model (DPSD). Probability density functions (PDF) and receiver operating characteristics (ROC) and zROCs generated from the VRDP. Parameters were chosen such that the VRDP approximates the unequal-variance signal detection model (UVSD) in the upper panels ($d'_F = 0.75$, $d'_R = 1.50$, and $R = 0.40$) and the dual-process signal detection model (DPSD) in the lower panels ($d'_F = 0.75$, $d'_R = 4.00$, and $R = 0.40$). **A**,**D**. Model PDFs. Dotted lines indicate the location of the underlying target distributions. **B**,**E**. ROC curve generated by the model. **C**,**F**. zROCs.

is curvilinear and asymmetric (Figure 2B); the zROC closely approximates a straight line with a slope of less than one (Figure 2C). On the other hand, when d'_R is large, the recollective and the familiarity distributions are quite distinct (Figure 2D). The resulting ROC is curvilinear and asymmetric, intersecting with the hits axis close to the value of R (Figure 2E). The corresponding zROC curve has a noticeable 'upward' bend, consistent with the predictions of the DPSD model (Figure 2F). In the limit as d'_R grows without bound, all recollected items exceed the (finite) highest-confidence bound and Eq. 3 is recovered. That is, as d'_R goes to infinity, the VRDP becomes identical to the DPSD.

Although the VRDP is clearly more flexible than either the UVSD or the DPSD, it cannot take on an infinite range of values. Figure 3 illustrates the family of zROC curves generated by the VRDP. The VRDP is limited to asymptotic slopes of one in both the liberal and stringent directions (assuming that $d'_R \neq \infty$). The value of d'_F controls the intercept of the liberal asymptote. The value of d'_R controls the difference between the intercepts of the liberal and conservative asymptotes and the value of R controls where the transition takes

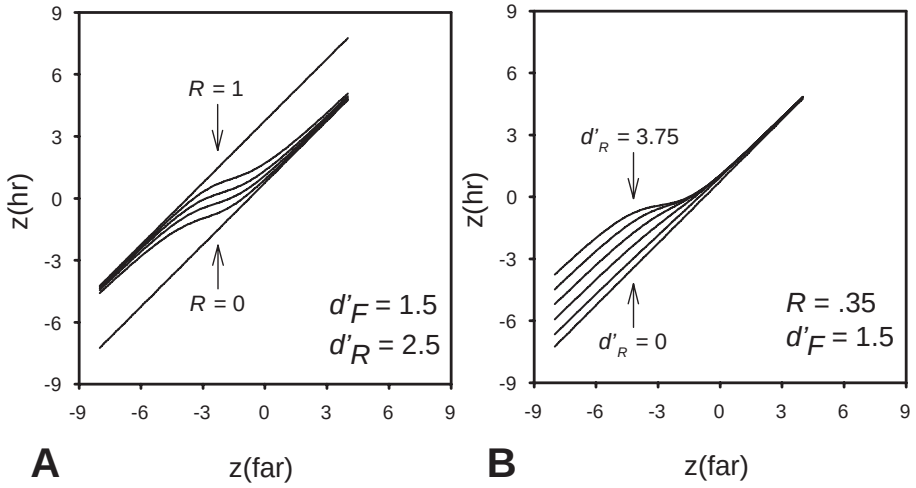

Figure 3. The effect of R and d'_R on zROC curves generated by the VRDP model. **A**: R was increased from 0 to 1 in increments of 0.2 (setting $d'_F = 1.5$ and $d'_R = 2.5$). For $R = 0$ and $R = 1$, the z-ROC curve is a straight line with a slope of unity. For intermediate values of R, the z-ROC curve shifts from the straight line corresponding to the familiarity distribution to the straight line corresponding to the recollective distribution, with the value of R controlling the location and the steepness of the transition. **B**: d'_R was increased from zero to 3.75 in increments of 0.75 (fixing $R = 0.35$ and $d'_F = 1.5$). The value of d'_R controls the intercept of the bottom straight line.

place. Notably, the VRDP can generate "kinked" zROC curves that resemble neither those predicted by the UVSD nor the DPSD (Sherman et al., 2003; DeCarlo, 2007, 2002).

Although the ability to fit item recognition zROCs is an important property of any model of recognition accuracy, item recognition zROCs alone will not be sufficient to resolve all of the debate. It is probably possible to write down a variant of the UVSD that includes a parameter for skew that would be able to generate nonlinear zROCs comparable to those predicted by the DPSD. Even if one accepts that the equations supporting the VRDP provide the best description of item recognition ROC curves, it is still possible that the psychology behind the model is incorrect. For instance, one could imagine a single-process model that happens to give rise to bimodal old item distributions. The mixture model of DeCarlo (2002) is usually understood in this sense, although his work can also be read more broadly. The dual-process interpretation of the VRDP predicts that the recollected probes should elicit conscious recovery of at least some details about the study episode whereas probes that are not recollected should not provide any details about the study episode. Unlike single-process interpretations of a bimodal old-item distribution, the VRDP makes specific predictions regarding the conjoint item-source recognition. In particular, the VRDP predicts that old items recognized on the basis of familiarity should not give rise to source discriminability whereas old items recognized on the basis of some recollection should give rise to source discriminability.

These considerations lead to a two-dimensional variable-recollection model of conjoint

item and source judgments (Figure 4a). In conjoint item recognition, a probe is first judged as old or new, followed by a source discrimination decision. To set a criterion for an item decision, we would draw a horizontal line through the response distributions projected onto the item axis (the y-axis in Figure 4a). Conservative item criteria correspond to the top of the figure. Source judgments correspond to setting a vertical criterion through the distribution. The parameter d'_S describes the discriminability of the recollected items on the source dimension.

When projected onto the source dimension, this two-dimensional VRDP model is mathematically identical to the mixture signal detection description of source recognition proposed by DeCarlo (2003a, see also Hilford et al., 2002). DeCarlo (2002, 2003a) proposed a mixture signal detection model of the old item distributions in both item recognition (DeCarlo, 2002) and source discrimination (DeCarlo, 2003a; Hilford et al., 2002) tasks. In both cases, the explanation for the mixture was a discrete attentional process. If one makes the identification that the item mixture and the source mixture are the same mixture, then the two-dimensional VRDP results. This identification does not necessarily follow from the attentional encoding account proposed by DeCarlo (2002). Indeed, DeCarlo (2003b) himself pursued an explanation of conjoint item-source recognition that did not make this identification. Moreover, Wixted (2007a) accepted mixture signal detection as an explanation of source ROCs in the context of discussing conjoint item-source ratings while maintaining that the UVSD describes the marginal item recognition response distributions. If the identification of the two mixtures is made, however, then the implication is that the high-attention old items retrieve source information in the item recognition task whereas low-attention old items retrieve item information but not source information. At that point, the difference between position of DeCarlo (2002, 2003a) and a dual-process theory becomes one of semantics.

Figure 4a gives a graphical description of the two-dimensional VRDP. The ellipses refer to multivariate normal distributions viewed from above, with the lines indicating a point of equal probability for each distribution, analogous to a topographic map. There are a total of four distributions of evidence, corresponding to new items, old items for which recollection fails and two distributions for which recollection succeeds, separated by the source in which the item was encoded. Distributions higher on the plot have more strength of evidence in an item recognition test. Distributions centered along a vertical line at the center of the figure, the new item distribution and old item distribution for which recollection fails, do not generate source discriminability. Recollected items are both higher along the item axis and more discriminable on the source axis. Old items which fail to generate recollection are higher on the item axis but not discriminable on the source dimension. To generate response probabilities, item criteria are generated by drawing a horizontal line on the graph. Source criteria would be represented as vertical lines at a particular level of item rating (we allow the source criteria to change across levels of item confidence). The probability of making a particular pair of item and source ratings is generated by taking the integral of the distributions over a rectangular region of the two-dimensional space.

The two-dimensional VRDP can be formalized as follows. For an old item presented as part of the left source, the probability of a response of at least $\mathbf{c}$, where $\mathbf{c}$ is a a point

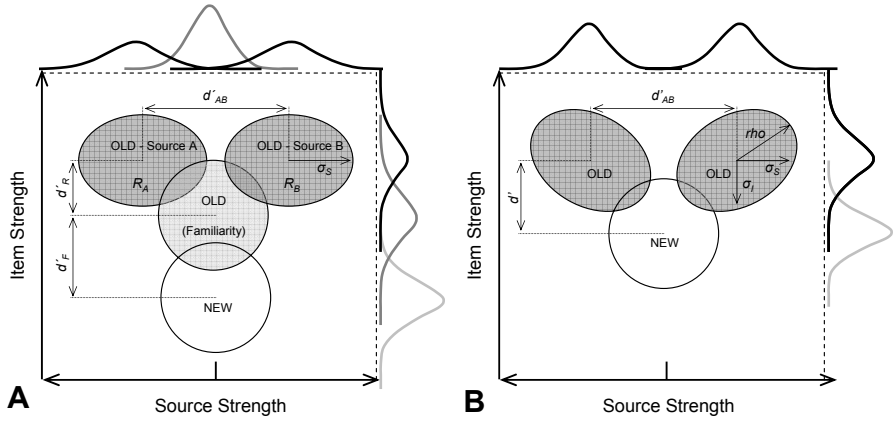

Figure 4. Two dimensional signal detection models for item and source judgments. A. The variable recollection dual process signal detection model (VRDP) leads to a natural model of conjoint item-source accuracy. The horizontal dimension reflects source strength, the vertical axis reflects item strength. Closed figures represent multivariate test probe distributions viewed from above, as in a topographic map. Recollected old probes (dark shading) lead to high item strength and also source discriminability. Old items that are not recollected do not result in source discriminability but do lead to item discriminability (light shading). New items are shown by the unshaded curve. B. The two-dimensional signal detection model proposed by DeCarlo (2003). The fact that item and source discriminability are correlated with one another requires that for very low values of item confidence, old items should show negative source discriminability as the two distributions cross.

describing an item criterion and a source criterion, respectively is given by

$$P(\text{response} \leq \mathbf{c} \,|\, \text{old} \,\&\, \text{left}) = R\,\Phi\left(\mathbf{c}, \boldsymbol{\mu}_L, \boldsymbol{\sigma}_S\right) \;+\; (1-R)\,\Phi\left(\mathbf{c}, \boldsymbol{\mu}_F, \mathbf{I}\right), \tag{7}$$

where Φ now refers to the integral of the multivariate normal distribution, $\boldsymbol{\mu}_L$ describes the mean of the recollective distribution for items presented on the left, located at the point $(-d'_F/2 - d'_R, -d'_S/2)$ and $\boldsymbol{\sigma}_S$ is a matrix with diagonal elements 1 and σ_S and zero off-diagonal terms. That is, we assume that item and source strengths are independent within the recollected distribution, but allow the standard deviation in the source dimension to vary. The second term in Eq. 7 describes the old familiarity distribution, which is centered at $\boldsymbol{\mu}_F$, which is the point $(-d'_F/2, 0)$; the identity matrix $\mathbf{I}$ indicates that the item-source distribution for non-recollected old items is circular. The equation for old items presented on the right source is identical except that the recollected distribution would be located at $\boldsymbol{\mu}_R = (-d'_F/2 - d'_R, d'_S/2)$.

For the new items, the probability of a response is given by

$$P(\text{response} \leq \mathbf{c} \,|\, \text{new}) = \Phi\left(\mathbf{c}, \boldsymbol{\mu}_N, \mathbf{I}\right), \tag{8}$$

where the two-dimensional normal distribution is located at $\boldsymbol{\mu}_N = (d'_F/2, 0)$.

Although conjoint item-source ratings have been used to argue against a straightforward interpretation of the DPSD, the data also argue against a straightforward interpretation of the UVSD. The most straightforward way to implement the UVSD in a conjoint

item-source task is to assume that there is a two-dimensional distribution of item-source strength with correlated dimensions. This follows from the dual-process interpretation of the UVSD proposed by Wixted (2007a) if one makes the identification that recollection should consistently give rise to source discriminability (see also Mickes, Wais, & Wixted, 2009). That is, in Figure 4b, the old items are characterized by a mixture of three elliptical distributions. In a straightforward two-dimensional UVSD, the default assumption is that as item strength for a probe from one source increases, source discriminability also increases. As illustrated in Figure 4b, this assumption leads to two ellipses meeting in a V-shape (DeCarlo, 2003b).

Uncorrelated item-source distributions (e.g. Banks, 2000; Glanzer, Hilford, & Kim, 2004) are falsified by the finding that source discriminability decreases with the item rating given to an item (Slotnick, Klein, Dodson, & Shimamura, 2000; Slotnick & Dodson, 2005). Although the correlated multi-dimensional UVSD (DeCarlo, 2003b) can fit some aspects of the conjoint item-source rating, detailed predictions were not published until quite recently (Hautus, Macmillan, & Rotello, 2008).[3] The correlation between item and source information predicts that with continuing decreases in item ratings, the V-shaped distributions cross over so that the model predicts *reversed* source discriminability for old items receiving a low item rating. Put another way, the V-shape observed at a particular probability is really the upper half of an X-shaped pair of distributions that would be observed at lower probabilities. Hautus et al. (2008) showed this prediction leads to poor fits to the observed data and proposed several adjustments to the model to try and reconcile the correlated UVSD account with the observed pattern of data.

The dual-process assumption of the VRDP implies a qualitative difference in the amount of source discriminability offered by the recollected items and the familiar old items. At high levels of item confidence, the VRDP predicts curvilinear ROCs approximating those predicted by an equal variance signal detection process—imagine taking a horizontal strip through the two recollected distributions at the top of Figure 4a. As item confidence is reduced, the horizontal strips describing source discriminability are dominated more and more by the familiarity distribution. This carries no source information, so the result is an ROC curve along the diagonal. Although these familiar old probes carry no source discriminability, they nonetheless are distinguished from the new probe distribution on the item dimension and can thus support above-chance discriminability in an item recognition task.

Overview of Experiments

In the experiments that follow we will evaluate the fit of the UVSD, DPSD, and VRDP models to words and travel scenes. In Experiment 1, we demonstrate that neither the UVSD nor the DPSD provide a consistently superior description across study materials, while the VRDP provides an excellent fit to both types of materials. Experiment 2 replicates these findings in a study that collects conjoint item and source judgments.

[3] Although Slotnick and Dodson (2005) in some sense fit the UVSD model to their conjoint item-source data, they did so by estimating a separate d' and criteria for each level of item rating. While this is useful in describing the shape of the source ROCs, this does not constitute an internally consistent two-dimensional account of item and source judgments.

Our modeling work touches on both item and source memory. With respect to item recognition, we compare the two parameter models, the UVSD and DPSD to each other across study materials. We find that neither model provides a satisfactory account across materials, with the UVSD consistently outperforming the DPSD for word stimuli and the DPSD consistently outperforming the UVSD for travel scene stimuli. We therefore reject both the UVSD and the DPSD as a common model of recognition accuracy across materials. We then compare the three parameter VRDP to two other three parameter models. One model is a superset of the UVSD and DPSD. That is, d'_F, σ_F and R are allowed to vary (with $d'_R = \infty$). This superset model is extremely flexible, with the ability to attain any zROC curve that either the UVSD or the DPSD can attain, and an additional set that neither the UVSD nor the DPSD, nor indeed the VRDP, can reach. The other three-parameter model we compare to the VRDP is a variant of the UVSD with encoding failure that was suggested by (Wixted, 2007b) to account for curvilinear zROC curves observed with pictorial materials.

In the source memory realm, we show that the VRDP provides an excellent description of conjoint item and source data. In particular, the VRDP is able to provide a good description of source accuracy at different levels of item confidence, a major source of criticism for previous dual-process models of item recognition (Slotnick & Dodson, 2005; Wixted, 2007a). This demonstration also provides support for a key qualitative prediction of the VRDP—source discriminability is restricted to the old items that receive high item confidence ratings, while for a large number of old probes item recognition is above chance even though source discriminability is completely absent.

Experiment 1: Item Recognition of Words and Travel Scenes

In order to resolve the extensive debate about whether the UVSD or the DPSD is a superior model of item recognition, we conducted a test of item recognition with multiple levels of confidence. Because prior results arguing for the superiority of the UVSD relied almost exclusively on data using words as stimuli, type of study material, words or travel scenes, was a between-subjects variable. We evaluated the results at the subject level by fitting the UVSD, DPSD, and VRDP models to the individual participants' response distributions, and at the group level by fitting the models to group zROC curves.

Method

Participants. A total of 267 Syracuse University undergraduates participated in the experiment for course credit. Participants were tested individually. All participants were native speakers of English. The participants studied either words or travel scenes. Participants' data were not analyzed further if they did not follow the task directions as instructed or had item recognition accuracy below a d' of 0.5 as estimated from the intercept of the best-fitting straight line fit to the zROC. Twenty-seven participants were excluded from the word condition and twenty were excluded from the travel scene condition, leaving $n = 116$ for words and $n = 104$ for travel scenes. Although this exclusion criterion may seem somewhat stringent in that a relatively large number of participants are excluded, it is particularly important given the goals of this paper that we exclude participants with poor performance. All of the models under consideration are perfectly able to account for

no discriminability between old and new items; the models make differentiable predictions only for participants who demonstrate robust discriminability.

Materials. Stimuli were presented on Dell desktops with 19-inch flat-screen monitors. The word pool was constructed from the MRC Psycholinguistic Database (Coltheart, 1981). The stimuli were all nouns 5-8 letters in length (mean 6.69), with an average frequency of 3.84, presented in capital letters in the center of the screen. Travel scenes came from the pool previously used in Howard et al. (2006). They included a variety of scenes from around the world depicting various natural, urban, and community landscapes, cultural events, and flora and fauna. Images that were obviously emotionally salient to a large proportion of viewers (e.g., the World Trade Center in New York City) were eliminated from the pool. Any images with writing were also excluded. Travel scenes were presented as digital pixmaps with a resolution of 350 x 232 pixels.

Procedure. Each participant studied six lists of 64 items (words or travel scenes). Presentation order was assembled randomly and independently for each participant. Stimuli remained on the screen for 1000 ms, followed by a 500 ms blank screen. After each list presentation, participants were given 128 probes, 64 new and 64 old, randomly intermixed. In response to each probe, participants rated the confidence of their memory on a scale from 1 (absolutely certain new) to 8 (absolutely certain old). They were instructed to use all eight buttons and cautioned against using extreme ratings exclusively. After each response, the program automatically advanced to the next test item. Prior to the experiment, the participants were familiarized with the procedure and completed a brief practice task. They were encouraged to take short rest breaks after each study-test block. Each participant contributed a total of 786 responses over six lists (384 old test items and 384 new test probes).

Analyses. For quantitative analyses the models were fit to each participant's response distribution via maximum likelihood estimation (MLE). Models were optimized using Microsoft Excel's Solver routine (Microsoft Corporation, 2007) by minimizing $-\sum_i N_i \log p_i$, where N_i is the number of responses in category i and p_i is the probability of response i predicted by the model. The sum runs over all eight response categories for both hits and false alarms. The models were also fit using SAS proc nlp (SAS Institute, 2005). The results from SAS fits were virtually identical to the fits produced by MS Excel 2007, hence only the Excel fits will be reported. For both the SAS fits and the Excel fits, multiple starting parameter values were used and the best-fitting solution was selected. We constrained the mean of the recollective old item distribution of the VRDP to be greater than that of the familiarity old item distribution, $d'_R > 0$. In fitting the VRDP, we also constrain $d'_F \geq 0$ to avoid overfitting.

Negative log likelihood calculated across subjects was used to compare models with the same numbers of parameters. In addition, we report the number of subjects best-fit by each model. These analyses showed convergent results. An appendix reports comparing models with different numbers of parameters.

To assess the ability of the models to describe the qualitative properties of the participants' item recognition performance, we fit the models to the group zROC curves by using a box-constrained quasi-Newton method (R function optim with method "L-BFGS-G") to

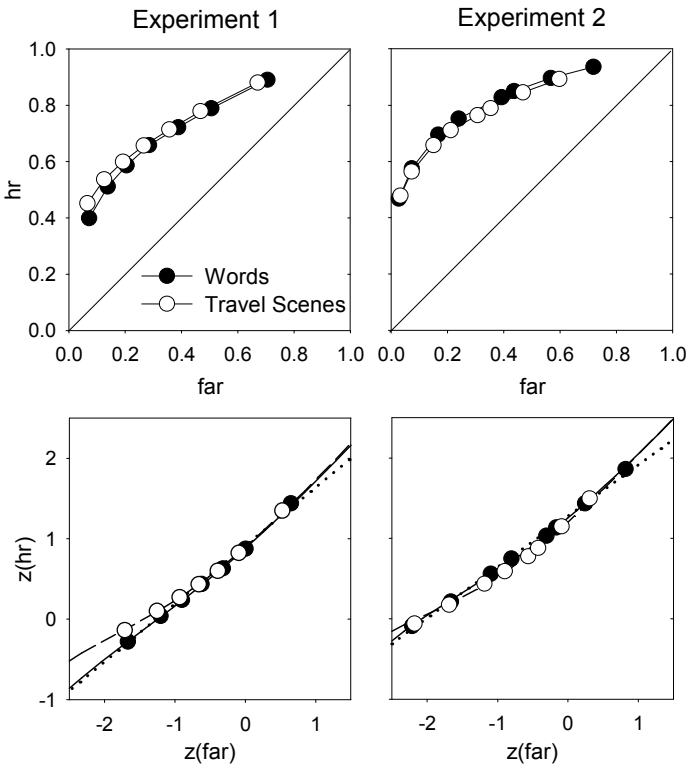

Figure 5. Item recognition ROC curves for Experiments 1, and 2 (upper panels), with zROCs averaged across individuals (lower panels). The zROC panels show the fit of the VRDP model to words (solid line) and travel scenes (dashed line). The dotted line is the best fitting linear regression to the word data.

minimize the squared distance between the model zROC points and the points of the group zROC curve, weighted inversely by the square of the standard error of the mean.

Results and Discussion

Qualitative fits to group zROC curves.

In order to characterize the zROCs from the two types of materials, we plotted average zROC curves by calculating each participant's zROC curve, then averaging each point. Figure 5 shows the results of this analysis, along with the fit to the group curve from the VRDP. For words, we observed a nearly-linear zROC curve (filled symbols) whereas the zROC we observed for travel scenes (open symbols) showed a reliable non-linearity similar to that predicted by the DPSD. To ensure that the VRDP model was able to describe the qualitative properties of recognition accuracy across materials, we fit the VRDP model to the group zROC data. As can be seen from Figure 5, the VRDP model provided a qualitatively satisfactory fit to group zROC curves for both words and travel scenes. The best-fitting parameters for the fit to the word data were $d'_F = .40$, $d'_R = 1.84$, and $R = .47$. The best-fitting parameters for the fit to travel scenes were $d'_F = .45$, $d'_R = 2.48$, and $R = .44$. The pattern in the fitted parameters paralleled the findings from the fits to

			Negative log-likelihood		Number best-fit	
		n	UVSD	DPSD	UVSD	DPSD
Exp. 1	words	104	**133,019**	133,077	**63**	41
	travel scenes	116	141,150	**141,008**	44	**72**
Exp. 2	words	75	**70,143**	70,241	**44**	31
	travel scenes	75	69,517	**69,444**	26	**49**

Table 1: Significant differences between models are denoted by the use of bold symbols to denote the best-fitting model. Note: n = number of participants; UVSD = Unequal-Variance signal detection model; DPSD = Dual-Process signal detection model;

the individual response distributions that we will report shortly. Examination of Figure 5 demonstrates that the VRDP model captured both the nearly-linear nature of the word zROCs and the apparent curvature of the travel scene zROCs. Neither the UVSD nor the DPSD can capture both of these types of zROC curves. As might be expected from Figure 5, the UVSD provided a better fit to the group zROC for the word condition than did the DPSD; the converse was true for the group zROC for the travel scene condition.

Fitting individual response distributions.

In order to determine if the deviation from the qualitative patterns of zROC data predicted by the UVSD and DPSD were statistically reliable, we first compare the fits of the UVSD and the DPSD to performance across materials (see Table 1). For words, the negative log-likelihood was lower for the UVSD than the DPSD model, indicating that the UVSD model produced a better fit to the ensemble of participants. Conversely, for travel scenes, the DPSD model produced a better fit to the ensemble of participants than the fit of the UVSD. One can use the negative log-likelihoods of the two models to calculate the conditional that one model is correct given that one of the two is correct. This calculation is closely analogous to Akaike weights (Wagenmakers & Farrell, 2004), although the models do not differ in number of parameters. According to the conditional probability, the difference in likelihoods between the models is hugely significant. The conditional probability that the DPSD is correct for words (given that one of the two models is correct) is less than 10^{-25}; the conditional probability that the UVSD is correct for travel scenes is $< 10^{-61}$. The difference in the models' abilities to describe the data produced by different types of materials was not due to anomalous results from a small number of participants. Table 1 also reports the numbers of participants best-fit by the UVSD and DPSD. At the individual participant level, the UVSD fit better than the DPSD for 63 out of 104 participants, $p < .02$, whereas for travel scenes the DPSD fit better than the UVSD for 72 out of 116 participants, $p < .01$. The proportion of participants best-fit by each model was significantly different across materials, $\chi^2(1) = 10.37$, $p < .001$.

The foregoing analyses suggest that neither of the two-parameter models widely in use in the recognition memory literature provides a satisfactory account of recognition accuracy across materials. Therefore we reject each of them as an account of the results of Experiment 1. We next compared the VRDP to two three-parameter models suggested by the literature. One is the superset of the UVSD and the DPSD, each of which is widely-

		Negative log-likelihood		Number best-fit	
		VRDP	Superset	VRDP	Superset
Exp. 1	words	**132,883**	132,928	**73**	31
	travel scenes	**140,821**	140,870	**74**	42
	combined	**273,704**	273,798	**147**	73
Exp. 2	words	**70,060**	70,069	40	35
	travel scenes	**69,351**	69,365	42	33
	combined	**139,411**	139,434	82	68

Table 2: Comparison of the VRDP to a three parameter model that is a superset of the UVSD and DPSD.

used in the recognition memory literature. The other is the UVSD with encoding failure for a subset of items. This model was suggested to account for nonlinear zROCs in item recognition of travel scene data by Wixted (2007b)

It would not make sense to accept either of the two-parameter models over the other because we have strong evidence that type of study material interacts with the preferred model. This suggests the possibility that there is not a common model of recognition performance. Perhaps when people recognize words they use the UVSD, but when they recognize travel scenes they utilize the DPSD. Perhaps a proportion of participants use the UVSD and another proportion use the DPSD. Perhaps participants can switch back and forth across materials, or even across lists. Because the understanding of recollection is so different in the two frameworks (Wixted, 2007a; Parks & Yonelinas, 2007a), this possibility is theoretically extremely unappealing. Fortunately, it can be directly tested.

We fit a three-parameter model with d'_F, σ_F and R allowed to vary (with $d'_R = \infty$). This model is the superset of the UVSD and DPSD. It can generate any data that either the UVSD or DPSD can. In addition, it can also exhibit phenomena that neither the UVSD nor the DPSD are capable of exhibiting. For example, the superset model can generate non-linear zROC curves with an asymptotic slope different from one for liberal criteria and an asymptotic slope of zero for stringent criteria. It should be noted that the VRDP cannot exhibit precisely linear zROC curves with slope different from one nor zROCs with asymptotic slopes different from one in the liberal criteria. That is, the set of zROCs the superset model can exhibit is larger than the superset of the zROCs the UVSD and the DPSD can generate, and disjoint in at least two ways from the set of zROCs that the VRDP can generate.

Table 2 illustrates the results of comparing the VRDP to the superset of the UVSD and DPSD. As can be seen from Table 2, the VRDP was more likely than the superset model for both words, conditional $p < 10^{-19}$, and travel scenes, conditional $p < 10^{-21}$ in Experiment 1. In addition, the proportion of participants better fit by the VRDP was significantly greater for both words, $p < .001$ and travel scenes, $p < .002$.

Wixted (2007b) attempted to reconcile the finding that there appear to be non-linear zROCs, as suggested by the DPSD, for travel scene data with the UVSD. He suggested the possibility that in situations where the presentation rate is relatively fast not all items may be encoded (see also DeCarlo, 2002). To address whether this possibility provides a superior

		Negative log-likelihood		Number best-fit	
		VRDP	UVSD-EF	VRDP	UVSD-EF
Exp. 1	words	**132,883**	132,928	**69**	35
	travel scenes	**140,821**	140,951	**87**	29
	combined	**273,704**	273,879	**156**	57
Exp. 2	words	**70,060**	70,103	**52**	23
	travel scenes	**69,351**	69,462	**67**	8
	combined	**139,411**	139,565	**119**	31

Table 3: Comparison of the variable recollection dual process model (VRDP) to the unequal variance signal detection model with encoding failure (UVSD-EF).

account of the data than the VRDP, we compared the VRDP model with a 3-parameter variant of the UVSD which allowed for encoding failure. To fit this model we added a parameter to the UVSD controlling mixing with a distribution of old items identical to the new item distribution.[4] The UVSD with encoding failure has sufficient flexibility to generate non-linear zROCs as well as linear zROCs with arbitrary slope.

Table 3 shows that for Experiment 1, the VRDP provided a much-superior fit to that data compared to that of the UVSD with encoding failure. The UVSD with encoding failure provided a worse fit than did the VRDP for both words, conditional $p < 10^{-19}$, and travel scenes, conditional $p < 10^{-56}$. This difference was not due to a few anomalous participants. The VRDP fit better for 69 out of 104 participants for words, $p < .001$, and 87 out of 116 participants for travel scenes, $p < .001$. This finding suggests that our results cannot be accommodated within the framework of the UVSD by including encoding failure.

Estimates of parameter values derived from fitting the VRDP model to the item response distributions from Experiment 1 are shown in Table 4. There were no differences between the mean estimates of d'_F or R across materials. There was, however a reliable difference between the estimates of d'_R, $t(218) = 3.70$, $p < .001$. For comparison purposes, parameter estimates for the fits of the UVSD and the DPSD models are also presented in Table 4, but caution should be exercised in interpreting these values as neither UVSD nor DPSD provide an acceptable fit to the response distributions across materials. However, it is certainly the case that the conclusions about the separate effects of material on recollection and familiarity depend dramatically on which model one uses to measure accuracy and the interpretation of those parameters.

[4] In general there are five parameters that can be fit in a one-dimensional mixture model in addition to the criteria. In addition to the parameters of the VRDP, R, d'_F, and d'_R, one can also imagine σ_F, from the UVSD, and σ_R being allowed to vary. The UVSD, DPSD and VRDP all reside within this general framework. For instance, the DPSD is achieved with $\sigma_F = 1$, $d'_R = \infty$ and R and d'_F allowed to vary. In this framework, the UVSD can be achieved in two ways, either with $R = 0$ and d'_F and σ_F allowed to vary or with $R = 1$ and d'_R and σ_R allowed to vary. To model the UVSD with encoding failure, we started from the second realization, fixed $d'_F = 0$ and allowed R to vary.

VRDP Model			UVSD Model			DPSD Model		
	words	travel scenes		words	travel scenes		words	travel scenes
				Experiment 1				
d'_F	.30 (.03)	.35 (.03)	d'_F	**1.38 (.07)**	**1.59 (.08)**	d'_F	**.68 (.03)**	**.57 (.03)**
R	.51 (.02)	.49 (.02)	σ_F	**1.49 (.04)**	**1.67 (.04)**	R	**.29 (.02)**	**.38 (.02)**
d'_R	**2.55 (.12)**	**3.17 (.11)**						
				Experiment 2				
d'_F	.51 (.05)	.62 (.04)	d'_F	2.06 (.12)	1.96 (.10)	d'_F	.97 (.04)	.87 (.05)
R	**.59 (.02)**	**.51 (.02)**	σ_F	1.64 (.04)	1.71 (.05)	R	.39 (.02)	.41 (.02)
d'_R	**2.92 (.15)**	**3.60 (.16)**						

Table 4: Parameter values obtained by fitting the VRDP, DPSD, and UVSD models to individual response distributions. Numbers in parentheses are standard errors. Parameter estimates from the UVSD and DPSD models should be interpreted with caution, as in most cases they do not provide optimal fit to data. Bold face is given for parameters significant ($p < .05$) for the comparison between words and travel scenes.

Experiment 2: Conjoint Item and Source Recognition

Experiment 1 demonstrated that the VRDP provides a good description of item recognition accuracy across criteria and for different materials. However, Experiment 1 did not directly evaluate the VRDP's dual-process interpretation of mixture signal detection. One could imagine a psychological motivation for mixture signal detection other than dual-process theory that would have yielded an identical description of the item recognition data (e.g. DeCarlo, 2002). The dual-process interpretation of the VRDP implies a qualitative difference in the type of information available for recollected and non-recollected old items. The dual-process framework predicts that there will be no source information associated with familiarity-based retrieval and that only recollected items can support source judgments.

The present experiment used both item and source judgments on each recognition probe. This experiment had two goals. First, we wanted to replicate the findings of Experiment 1 by fitting the VRDP to the marginal item recognition data. Second, we wanted to evaluate the dual-process interpretation by simultaneously fitting item and source judgments. If the dual-process interpretation is correct, then there should be a large number of old items with no source discriminability, but reliable item discriminability.

Method

Participants.

Data from one hundred fifty Syracuse University undergraduates who participated in the study for course credit were used in the analyses. Participants were tested individually. Half of them were randomly assigned to study lists of words; half to study sets of travel scenes. Participants were replaced if they did not follow the task directions as instructed

(typically using only extreme response keys) or had item recognition accuracy below a d' of 0.5 as estimated from the intercept of the best-fitting straight line fit to the zROC. Eighteen participants were replaced in the word condition; nine in the travel scene condition.

Materials. The stimuli and the apparatus were the same as those used in Experiment 1. However, pilot testing showed that there was a dramatic difference in the source accuracy of visually-presented words and travel scenes. In an attempt to boost item and source accuracy for the words they were presented both visually and auditorially in Experiment 2. Words were digitized for auditory presentation in a male and a female voice.

Procedure. The study phase was the same as in Experiment 1, with the following exceptions. Each participant studied six lists of 50 items (words or travel scenes). Three untested items were added before and after each study set. Each stimulus was shown for 2 seconds; between presentations, a 500-ms fixation cross appeared in the center of the screen. Within each set, half of the items were presented on the left and half on the right side of the screen. To improve retention, stimuli were encased in a frame of a specific color (red or green) and words were also spoken in a male or a female voice. The pairings for each participant were consistent throughout the entire experiment. Pictures of travel scenes were only presented visually. The participants were instructed to form associations between studied items and the source features.

For each test probe (old and new), the participants first rated their memory for the item on a scale of 1 to 9 (i.e., 1 = very sure new; 9 = very sure old), and then rated their memory for the location of the item, also on a 9-point scale (1 = very sure left; 9 = very sure right). All participants were informed that location was redundant with the color of the frame. Participants in the word condition were informed that location was also redundant with the gender of the voice speaking the word. Participants were advised that one reason they may not remember the location of a probe was because the probe was not presented during the study phase. The same set of computer keys was used to make the item and the source judgments. To visually separate the item scale from the source scale, the item scale disappeared once a response was made; immediately after, the source scale appeared on the screen slightly below where the item scale had been. Each participant contributed a total of 600 responses on the item recognition test (300 responses were made to the items presented at study, and 300 to new, never-presented probes) and a total of 600 source responses. Of these, 150 were old items from the left or the right sources each and 300 were in response to new items.

Analyses.

Analyses on the marginal response distributions for the item judgments were conducted using the same methods as Experiment 1. Hautus et al. (2008) demonstrated that a straightforward two-dimensional implementation of the UVSD provides a poor qualitative fit to conjoint item-source data. In addition, several authors have argued that the form of conjoint item-source ratings argue against the DPSD out of hand (e.g., Slotnick & Dodson, 2005; Slotnick et al., 2000; Wixted, 2007a, but see Parks & Yonelinas, 2007a). For these reasons, our interest was not in competitive model-fitting, but rather demonstrating that the dual-process VRDP was able to account for the qualitative pattern of results describing source discriminability across different levels of item ratings while also accounting for the

shape of the item ROC.

The two-dimensional VRDP source model was fit to the conjoint item-source response matrix comprised of group data, by maximizing the negative log-likelihood of the probability of response in each cell given the model predictions. The decision to do the fits at the group level was largely motivated by the dramatic sparsity of large regions of the response matrices, which was even more pronounced at the individual participant level. The fits were calculated using Microsoft Excel's Solver routine. As illustrated in Figure 2a, the free parameters of the VRDP model consisted of d'_F, d_R, R, plus the source discriminability d_S and standard deviation σ_S. In addition, we estimated 8 item criteria and 72 source criteria (i.e., 8 source criteria for each of the nine levels of item confidence).

Results and Discussion

Item recognition: Qualitative fits to group zROC curves. Figure 5 shows ROC curves and zROC curves for the item judgments in Experiment 2. As in Experiment 1, the travel scene data showed non-linear zROCs compared to the more-linear zROCs for word data. In contrast to Experiment 1, the difference between materials in overall accuracy was attenuated, presumably due to auditory and visual presentation of word stimuli. As in Experiment 1, the VRDP model provided an excellent qualitative fit to the group zROC curves (Figure 5, bottom right). The best-fitting parameters for the fit to the word data were $d_F = .64$, $d_R = 2.20$ and $R = .61$. The best-fitting parameters for the fit to travel scenes were $d_F = .73$, $d_R = 2.72$ and $R = .51$. The model was able to capture the nearly linear zROC curve from the word data as well as the curvilinear zROC curve from the travel scene data.

Item recognition: Individual response distributions.

We first compared the ability of the UVSD and DPSD models to describe item recognition across materials using the marginal item recognition response distributions. The results are shown in Table 1. As in Experiment 1, the UVSD model resulted in a smaller negative log-likelihood than the DPSD for words, whereas the opposite was true for travel scenes. The UVSD provided a better fit for 44 out of 75 participants in the word condition, $p = .08$. In contrast, the DPSD fit better for 49 out of 75 participants in the travel scene condition, $p < .001$. These proportion of participants best-fit by each model differed across materials $\chi^2(1) = 7.74$, $p < .01$.

As in Experiment 1, neither the UVSD nor the DPSD provided a superior fit across materials. Rather, the model that provided the best fit changed systematically across study materials. Unlike in Experiment 1, there was not an overall advantage in accuracy for travel scenes in Experiment 2, and the encoding conditions were quite distinct. Despite these procedural differences and a moderation of the difference in the overall accuracy across materials, we confirmed the theoretically unsatisfying result that neither the UVSD nor the DPSD provides a general description of the data.

As in Experiment 1, the VRDP provided a superior fit to the two three-parameter models we considered. Table 2 shows that the VRDP vastly outperformed the three-parameter model that is a superset of the UVSD and DPSD. The conditional probability of the superset model was less than .0002 for words, less than 10^{-6} for travel scenes and less than 10^{-9} for words and travel scenes taken together. As in Experiment 1, the

number of participants best-fit by the VRDP was greater than the number best-fit by the superset model for both words or travel scenes, although this number was not significantly greater than what would be expected by binomial variability. Table 3 shows the results of comparing the VRDP to the variant of the UVSD with encoding failure suggested by Wixted (2007b). The conditional probability of the UVSD with encoding failure was less than 10^{-18} for words, less than 10^{-48} for travel scenes and less than 10^{-110} for the data across materials. The VRDP outperformed the UVSD with encoding failure for 52 out of 75 participants for words, $p < .001$, and for 67 out of 75 participants for travel scenes, $p < .001$. Combined with the results from Experiment 1, these findings strongly suggest that the VRDP provides a better description of the item recognition data than the superset of the UVSD and the DPSD and that encoding failure within the framework of the UVSD is not sufficient to account for our observation of non-linear zROCs in the travel scene data.

Best-fitting parameter values for the VRDP can be found in Table 4. In Experiment 2 both R, $t(148) = 2.70$, $p < .01$, and d_R, $t(148) = 3.07$, $p < .01$, differed reliably between words and travel scenes.

Despite a smaller difference in overall item recognition performance across materials than found in Experiment 1, and other differences in the methods, modeling of individual participants' marginal response distributions replicated the critical aspects of the modeling in Experiment 1. Considering the UVSD and the DPSD, neither model provided a consistently superior fit across materials. The VRDP model also provided a superior fit to a three-parameter variant of the UVSD with encoding failure and even to the superset of the UVSD and DPSD.

Conjoint item and source recognition.

For the data from the word condition, the fits settled on $d_F = .78$, $d_R = 2.48$, $R = .44$, $d_S = 3.13$, and $\sigma_S = 1.09$. For the data from the travel scenes condition, the fits settled on $d_F = .65$, $d_R = 2.42$, $R = .47$, $d_S = 2.29$, and $\sigma_S = .93$. A graphical representation of the parameters, along with the values of the criteria can be seen in Figure 6. It is interesting to note that the best-fitting values of σ_S did not differ dramatically from unity, suggesting that circular distributions may be sufficient. It is also interesting to note that the source criteria are decidedly not constant across levels of item confidence. That is, the value of a particular source criterion at one level of item confidence is not the same as the value of that source criterion at another level of item confidence. This can be seen clearly from examination of Figure 6B, which illustrates the best-fitting parameters for the travel scene condition. The source criteria start closely placed near the center of the distributions at high level of item confidence (at the top of the figure), then spread out for lower-confidence ratings. This reflects the property that most new items received very low-confidence source ratings (i.e., ratings near 5). Hautus et al. (2008) noted that curvilinear source criteria could result from using decision bounds calculated from fixed likelihood ratios. The comparison between the present model the the models explored in Hautus et al. (2008) will be explored more extensively in the general discussion.

The best-fitting ROC curves for the conjoint VRDP model are shown in Figure 7. The VRDP model provided an excellent qualitative fit to the source response matrix conditionalized on the level of item confidence. According to the VRDP, old items that receive a high item confidence rating will be almost exclusively recollected. Under these circum-

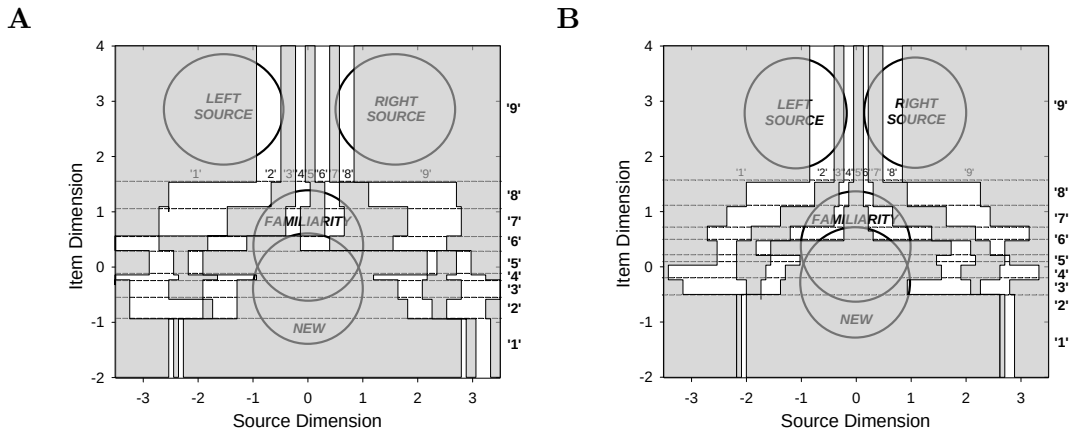

Figure 6. Graphical representation of the best-fitting parameters from the conjoint item-source ratings in Experiment 2. Item criteria are shown as horizontal lines that extend across the entire figure. Shading illustrates regions with the same source rating. That is, source criteria appear as vertical line segments separating regions with different shading. A. Words. B. Travel scenes.

stances, the some-or-none assumption means that the model generates source ROCs that approximate the equal-variance signal detection model. An all-or-none model would predict that the ROCs should be linear. As the item ratings decrease, the model passes through a mixture of the recollection and familiarity distributions, settling on a source ROC with chance discriminability for items that that do not receive a high-confidence item rating.

The data show this same pattern of results, with items that receive an item confidence rating of less than seven or eight showing source discriminability that was not different from chance. This outcome is consistent with the hypothesis that item recognition is characterized by two qualitatively different processes, familiarity and recollection, and that recollection fails for a subset of old items. If recollection contributed useful source information for all old items then there should be some source information for all old items. This assertion is not supported by the data in Figure 7.[5]

However, it remains possible that there are not two qualitatively different processes supporting recognition memory. It is possible that the items that received low item recognition confidence ratings also failed to show item recognition discriminability as well as failing to show source discriminability. In this case, the most parsimonious explanation would be that old probes that received low item confidence ratings were not encoded at all. In order to assess this, we undertook additional analyses.

Discontinuity between item and source discriminability. To assess the memorability of items that attracted different levels of item confidence, we constructed item zROC curves cumulated across different combinations of confidence ratings, separately for words and travel scenes. The zROCs were constructed for responses 1-9, 1-8, 1-7, 1-6, 1-5, 1-4, and

[5]It is possible that recollection does contribute some information to all items, but that this information is ignored for those items that receive an item rating below some threshold. This possibility is discussed more extensively in the general discussion.

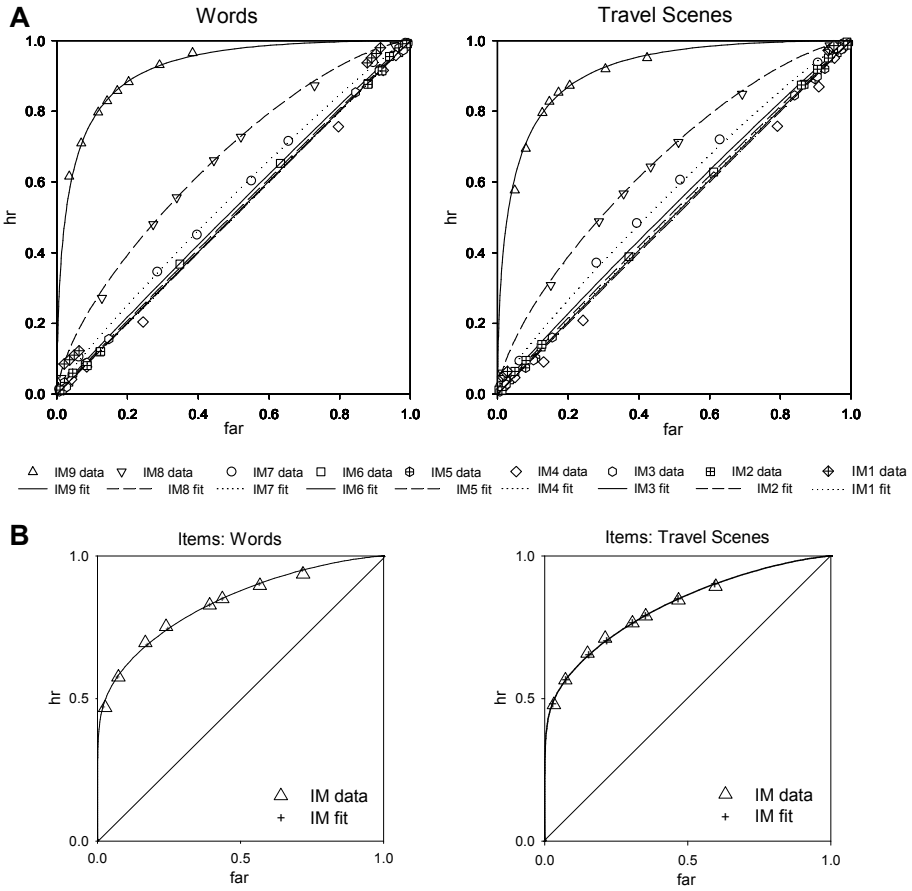

Figure 7. The VRDP model fit to conjoint item and source ROCs. **A**: Source memory ROCs conditional on item recognition. The open symbols are the data. The curves depict the model fit. Data from the word condition of Experiment 2 is on the left. Data from the travel scene condition is on the right. **B**: Fits to the item recognition ROC from the conjoint item-source fit in panel A. hr = hit rate; far = false alarm rate.

1-3 (Figure 8). As can be seen from the figure, for the lower item confidence ratings, the item zROC curves were approximately linear with a slope approaching one. Critically, the intercept for all zROC curves was reliably above zero, indicating that there was reliable item discriminability for probes that attracted lower levels of item confidence. Note that above-chance item discriminability was observed even for items given an item rating of "3" or lower on a nine-point scale.

To illustrate this finding in another way, Figure 9 shows source accuracy (d'), as a function of item recognition accuracy (intercept of a linear regression to the item zROC). Source and item accuracy are reported first for all probes (top right of each panel). Each successive point gives source and item accuracy for probes given a more restricted range of item ratings. For instance, the next point gives source and item accuracy calculated only from probes given an item rating from 1-8. Successive points take an increasingly restricted

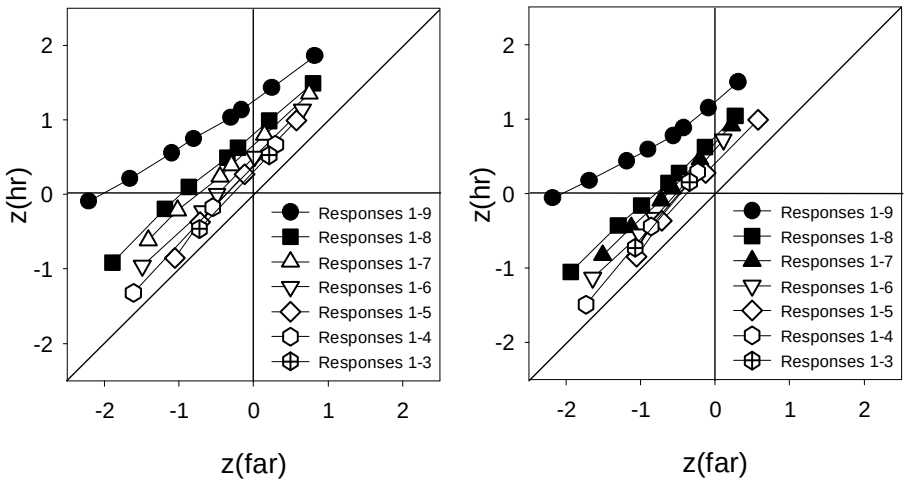

Figure 8. Item recognition zROCs cumulated across different combinations of item confidence ratings in Experiment 2. The zROCs for words are shown on the left; for travel scenes, on the right. Filled symbols: zROC curves for items with above-zero source discrimination. Unfilled symbols: zROCs for responses that carried no source information. hr = hit rate; far = false alarm rate.

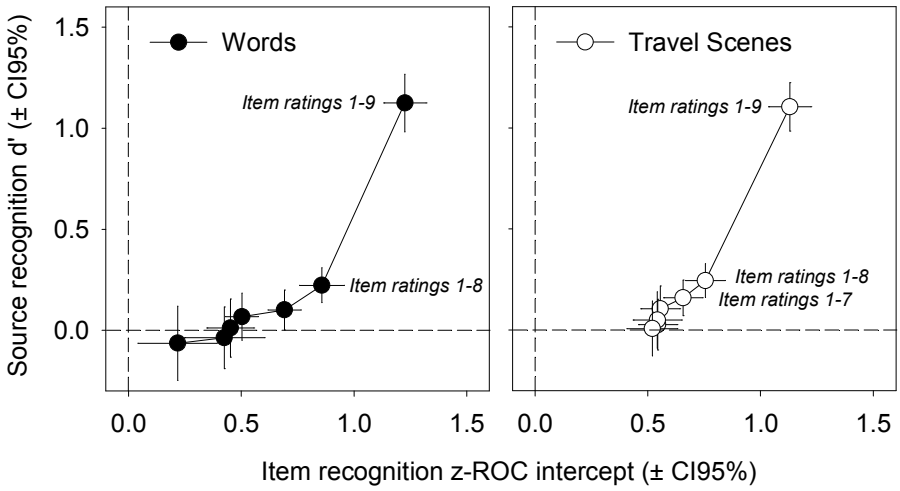

Figure 9. Source discrimination as a function of item recognition accuracy for words (left) and travel scenes (right) in Experiment 2. On the x-axis, an estimate of item discrimination (defined as the intercept of the zROC) is plotted across the cumulated item ratings along with the 95% confidence intervals. On the y-axis, source accuracy, expressed as d' is given. The rightmost point in each panel reflects item and source accuracy for all responses (i.e., item recognition confidence ratings 1 through 9). The second point from the right corresponds to estimates of item and source accuracy calculated only for probes assigned item ratings 1 through 8, and so on, with the leftmost point in each panel corresponding to probes assigned ratings 1 through 3.

range of ratings down to probes given item ratings 1-3 for the point in the lower-left of each panel. While item recognition accuracy at each level of confidence was significantly greater than chance (all ps $< .01$, note that the error bars in Figure 9 are 95% confidence intervals), source accuracy, on the other hand, was at above-chance levels only for ROC curves constructed from ratings 1 through 9 and 1 through 8 for both words and travel scenes and ratings 1 through 7 for words only. Notably, the function relating source discriminability to item discriminability shows that at lower confidence levels, the values cluster around a point on the source discriminability axis, but reliably away from zero on the item discriminability axis. If recollection leads to source discriminability, this result is not what one would expect if all items were recollected, which would predict all old items should have some source discriminability. It is also not consistent with the hypothesis that a subset of studied items were not encoded at all, which would predict that item discriminability for these items would also be zero.

General Discussion

The results of the experiments and model-fitting described above provide strong support for a dual-process, variable-recollection description of recognition accuracy. Model fits to individual response distributions revealed that the unequal-variance signal-detection model (UVSD) provided a better fit to the word recognition data than the dual-process signal-detection model (DPSD). However, the reverse was true for the travel scenes (Table 1). The cause of this difference in the ability of the models to account for the data across materials is almost certainly the existence of zROC curves that diverge from the qualititive predictions of each model (compare Figure 1 with Figure 5). This discrepancy indicates that neither the UVSD model nor the DPSD model provides a satisfactory description of recognition accuracy across materials.

A some-or-none dual-process model (VRDP) provided an excellent qualitative description of the zROCs across materials (Figure 5). As a consequence, it fit the item recognition data better than the *superset* of the UVSD and DPSD across both types of materials (Table 2). The VRDP also provided a superior fit across materials to the UVSD with encoding failure (Table 3), a model proposed by (Wixted, 2007b) to account for nonlinear zROCs observed with pictorial stimuli (Sherman et al., 2003; Howard et al., 2006). The VRDP, like the DPSD, assumes that recollection can fail. However, when recollection succeeds variability in the amount of recollected information can lead participants to select a variety of confidence ratings. With appropriate choice of parameters, the VRDP can approximate both the UVSD and the DPSD, generating zROCs that are nearly linear or zROCs with a pronounced curvature (Figures 2, 5). The dual-process interpretation of mixture signal detection (DeCarlo, 2002) that the VRDP makes leads to strong predictions about the nature of conjoint item/source recognition (Figure 4a). The VRDP provided an excellent qualitative description of empirically-observed conjoint item and source accuracy (Figure 7). In particular, the VRDP predicted that old probes that are not recollected have no source discriminability but have reliable item accuracy on the basis of familiarity. We saw just this result for old items that received lower levels of item confidence (Figures 8, 9).

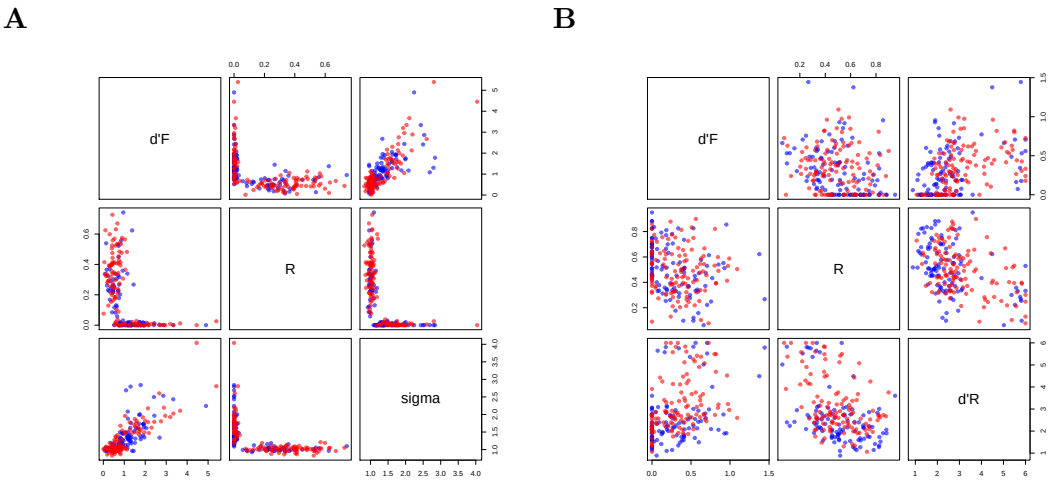

Figure 10. Distribution of parameter values for the superset model (A) and the VRDP (B). Values are shown from Experiment 1. Each subject in the word condition is shown in red; each subject from the travel scene condition is shown in blue. The superset model organized subjects into two clusters—UVSD-like participants and DPSD-like participants (see especially the middle plot in the right column of R as a function of σ_F. In contrast, the VRDP provides a continuous description of variability across participants.

Neither the UVSD, nor the DPSD, nor the UVSD and the DPSD describe the data

We compared the VRDP to a model that is a superset of the UVSD and DPSD. This superset model, which has the same number of parameters as the VRDP, can take on a wide variety of states that the the VRDP cannot. Nonetheless, the VRDP provided a fit superior to that of the superset model across both experiments (Table 2). This is presumably due to the VRDP's ability to generate zROC curves with a kink (Sherman et al., 2003; DeCarlo, 2007), a property not shared with the superset model. However, even if the superset model had provided a numerically superior fit to the VRDP, it would still be theoretically unsatisfactory.

Figure 10 shows the best-fitting parameter values for each subject in Experiment 1 for the superset model (Figure 10A) and the VRDP (Figure 10B). The superset model describes participants as falling into one of two distinct clusters. The superset includes one cluster of UVSD-like participants with $R \simeq 0$, $\sigma_F > 1$ and relatively large values of d'_F. In addition, there is a cluster of DPSD-like participants with $R > 0$, $\sigma_F \simeq 1$ and relatively smaller values of d'_F. Very similar results were observed for Experiment 2. According to the superset model, participants either behave like the UVSD or the DPSD—no participants manifest a combination of the two models. That is, no participants showed both a high degree of recollection $R \gg 0$ and unequal variance ($\sigma_F \gg 1$). Note that because the VRDP can approximate both the UVSD and the DPSD, these are precisely the parameters that yield predictions qualitatively different from the VRDP—yet they are never observed.

Although the parameters of the superset model lead one to conclude that participants must behave like either the UVSD or the DPSD rather than a combination between the two, the number of participants in each cluster of parameters changes across materials!

The fact that the allocation of participants changes across conditions implies that at least a subset of participants must be able to approximate each model. But then why wouldn't these participants be able to draw on both sources of information rather than choose one or the other? In contrast, the VRDP makes the hypothesis that the linear (or nearly-linear) zROCs with slope less than one observed in some data and the non-linear zROCs observed in other experiments are a result of the same cause, but with a different parametric value. We conclude that the superset of the UVSD and DPSD, in addition to showing a numerically inferior fit to the data (Table 2) is also theoretically much less satisfactory than the VRDP.

Measuring recognition accuracy

One of the goals of descriptive models, like the UVSD, DPSD and VRDP, is to provide a means to solve the problem of how to measure recognition accuracy. One of the potential advantages of the VRDP is that it offers the ability to measure separate properties of recognition. However, there is much work to do before the VRDP can be widely adopted as a practical means of measuring recognition accuracy. There is no guarantee that a given ROC curve can be described by a unique set of 3 parameters. When the estimate of R approaches 0 or 1, the target item distribution becomes unimodal and the model loses the ability to distinguish between d'_F and d'_R. A unimodal normal distribution can be fit reasonably well by a model with a high d'_F value and a very small d'_R value (i.e., setting $R = 0$), or a model with a very small d'_F value and a large d'_R value (i.e., setting $R = 1$).[6] The above considerations make clear that in the context of the VRDP, the finding of ROCs consistent with an equal-variance signal detection process is not by itself evidence for an impairment in either familiarity or recollection.

Another question of practical application concerns one's ability to estimate the parameters of the VRDP model by examining the ROC or the response distributions. Unlike the UVSD, which is guaranteed to provide easily interpreted results for any experiment that generates a linear zROC curve, for the VRDP there is at present no substitute for directly fitting the model to the data. It is also possible that the parameters of the model can trade off with each other in practice, especially when there are not a great many responses to constrain the data. Work on the sampling statistics of the VRDP analogous to that undertaken for the UVSD (Macmillan, Rotello, & Miller, 2004) to estimate the effects of decision error on parameter estimates (e.g. Malmberg & Xu, 2006) would also be a welcome development.

We should note that the VRDP is almost certainly not sufficiently complex to provide a truly detailed model of recognition accuracy. Our analyses average over a number of variables known to affect recognition performance, including list number, test position, serial position and word frequency. An accurate model would have to also include these sources of variability, as well as a number of other variables. It is also possible that our results were affected by the relatively large number of response options we used and the large amount of data we collected. Nonetheless, at least for these data, the two-parameter models failed. This is particularly problematic from a measurement perspective. Most researchers typically only use a one-parameter model to measure recognition accuracy (common measures

[6] A third possibility would occur when the recollective and the familiarity distributions lie on top of each other, in which case $d'_F = d'_R$, and R can take on any value.

include d', proportion correct and A'). The investigation of the effects of experimental manipulations on the two-parameter models have barely begun. If recognition discriminability is a three-dimensional quantity, as our results strongly suggest, then this means that we know essentially nothing about the separate effects of experimental manipulations on d'_F, d'_R and R. These distinctions may prove to have important implications for substantive models of recognition performance.

Empirical issues in item recognition

Our conclusions constitute a meaningful departure from the UVSD that has been the dominant measure of recognition accuracy for the last decade and a half. However, our empirical data do not differ substantively from what has been reported previously for either item recognition or source memory. Wixted's (2007a, 2007b) empirical argument that the UVSD model provides a superior fit to that of the DPSD in describing item recognition accuracy was based on the review of studies that have almost exclusively considered words as stimulus materials. The data from our word condition can be included in this category; in both experiments, the UVSD model outperformed the DPSD model when fit to the word data. Our empirical results for travel scenes, and the finding that the DPSD outperforms the UVSD for travel scene data, are also consistent with what has been reported previously (e.g., Howard et al., 2006; Sherman et al., 2003). Our source recognition data are comparable to other studies as well—for example, the source ROCs separated by item strength (Figure 7) appear very similar to the ones described by Slotnick and Dodson (2005). The consistency of our observations with previous reports makes it unlikely that our conclusions stem from some peculiarity in how the data were collected. Moreover, the size of our sample makes an explanation based on random variation unlikely: across the two experiments there are more than 250,000 item recognition responses and an additional 45,000 source responses.

Our results demonstrate that there are reliable differences in the shape of zROC curves for words and travel scenes. According to the VRDP model, these differences are largely attributable to a larger d'_R for travel scenes than for words. Indeed, analysis of model parameters showed a reliable difference in the magnitude of d'_R across both experiments. In Experiment 2, words were encoded both auditorially and visually. Overall levels of item recognition (and source) accuracy were comparable for travel scenes and words in Experiment 2, but there was still a reliably larger d'_R for travel scenes than for words. In Experiment 2, the probability of recollection was greater for words than for travel scenes, perhaps due to the additional set of source features available for the words, which were presented both auditorially and visually. It is interesting to speculate why d'_R tends to be larger for the travel scenes. It is possible that the travel scenes possess more unique perceptual details, resulting in greater discriminability during test if those details are retrieved.

Regardless of the source of the differences between words and travel scenes, the fact that the DPSD and the UVSD differentially fit across materials suggests that some of the apparent debate in reconciling other studies with one another may be at least partially attributable to differences in study materials. Fortin et al. (2004) found that the data from an odor recognition task performed by rats were fit better by the DPSD than the UVSD model. Notably, Fortin et al. (2004) found nearly-linear ROCs (and thus highly non-linear zROCs) for rats presented the odor recognition task at a delay. This finding cannot be reconciled with the dual-process theory interpretation of the UVSD. In contrast, Wais et al.

(2006) found ROCs consistent with an equal-variance signal detection model with a delay in a word recognition task with human adults. These experiments differ radically not only in their procedure and subjects, but also in the nature of the stimulus materials, which the present results have shown are sufficient to alter the quality of model fits even under carefully controlled circumstances.

The VRDP can explain both the linear ROCs observed by Fortin et al. (2004) and the symmetric curvilinear ROCs observed by Wais et al. (2006) as disruptions of familiarity. The VRDP can explain the linear ROCs observed with a delay by Fortin et al. (2004) as resulting from a d'_F of zero and a very large value of d'_R with a non-zero value of R. If recollection is a some-or-none process, as postulated by the VRDP, then the Wais et al. (2006) findings are also perfectly consistent with the gradual decrease of a recollective process, only with a more modest value of d'_R. Notably, we observed relatively modest values of d'_R for words relative to travel scenes even without a delay in the current study. Within the context of the VRDP, the apparent discrepancy between the Wais et al. (2006) data and the Fortin et al. (2004) findings disappears. We should caution that while both the findings of Fortin et al. (2004) and Wais et al. (2006) are consistent with a decrease in the efficacy of familiarity with a delay, one is not forced into this interpretation by the VRDP. Strictly speaking, the presence of symmetric ROCs for patients with disruptions to the hippocampus observed by both Fortin et al. (2004) and Wais et al. (2006) do not by themselves imply that the hippocampus is essential for recollection in the context of the VRDP. Because the VRDP is a mixture of two equal-variance signal detection models, a more symmetric ROCs could, in principle at least, be either the result of depressed recollection (R near zero) or intact recollection with a high probability of recollection accompanied by a small d'_R. The choice must be made on the basis of considerations external to those provided by the model as a measurement tool. These considerations lead us to strongly favor the interpretation that hippocampal damage is causing a disruption of recollection rather than familiarity in the case of hippocampal damage (see also Farovik, Dupont, Arce, & Eichenbaum, 2008; Robitsek, Fortin, Koh, Gallagher, & Eichenbaum, 2008; Sauvage et al., 2008).

Alternative theoretical approaches

The question of whether the UVSD or the DPSD provides a better account of recognition memory (Parks & Yonelinas, 2007a, 2007b; Wixted, 2007a, 2007b) appears to have a clear resolution—they are both insufficient to describe recognition accuracy. This does not mean, however, that the VRDP is correct. There are undoubtedly alternative approaches outside the scope of models that have thus far been explored.

Nonlinear zROC curves falsify a key qualitative prediction of the UVSD. There are a number of ways these might be reconciled. One may appeal to guessing as a means to introducing nonlinearities to distort an otherwise linear zROC (Ratcliff, McKoon, & Tindall, 1994; Malmberg & Xu, 2006). The challenge to such an approach from the present findings is why guessing appears to have little to no effect for words but a large effect when studying travel scenes.

Another approach to account for our item recognition findings comes from the idea that the distribution of strengths of the evidence supporting an item recognition judgment is non-normal but unimodal. A particularly promising aspect of this line of research is that it could be conducted in the context of substantive models of recognition accuracy.

Although the Shiffrin and Steyvers (1997) REM model can make distributions of evidence that are nearly normal, it is capable of exhibiting a much wider range of behavior. In addition, BCDMEM (Dennis & Humphreys, 2001) also generates non-normal strength-of-evidence distributions. The variety of zROCs generated by these models may be sufficient to account for the variety of findings observed here. The challenge for models of this class will be to account for the finding that a large proportion of old item probes generate essentially no source discriminability but robust item discriminability in a single-process framework (Figure 9).

One possibility is that study events are described as a series of features, some of which are relevant for an item decision and some of which are relevant for a source judgment. When faced with an item judgment, a different set of features is consulted than when presented with a source discrimination judgment. Heathcote, Raymond, and Dunn (2006) explored the potential of task-dependent cue matching to explain various phenomena ascribed to recollection. If features are discretely encoded and/or retrieved, then the presence of a set of old items with no source discriminability but above-chance item discriminability makes sense—those are items that failed to encode or retrieve the features relevant for the source discrimination.

Task-dependent cue matching of discrete features may be usefully contrasted with the VRDP by examining the predictions made in situations where multiple source dimensions can be queried (in the current study, there were multiple source dimensions but they were redundant with each other). According to the VRDP, recollection is a precondition for retrieving any conscious details of the experience. Because the evidence induced for recollected items is continuous, there is no guarantee that all source features are retrieved perfectly, but there should be strong correlations between success in retrieving one source dimension and another when taken across items that do and do not succeed in triggering recollection. Indeed, positive correlations between source dimensions are observed in multifeatural source studies (Meiser & Bröder, 2002; Meiser, Sattler, & Weisser, 2008; Starns & Hicks, 2005) with some evidence that positive correlations are restricted to items given remember responses when the remember-know procedure is utilized (Meiser & Bröder, 2002; Meiser et al., 2008). Although a single-process explanation of our results based on task-dependent cue matching of discrete features may be possible, the challenge would be provide a principled account of correlations among source features for some probes without that explanation amounting to a qualitatively different process.

Another way to attempt to account for the finding that a large population of old items show no source discriminability but above-chance item discriminability without appealing to dual-process theory is to alter straightforward multivariate signal detection theory (DeCarlo, 2003b) to attempt to overcome its counterfactual predictions. Hautus et al. (2008) demonstrated that a straightforward application of multivariate signal detection theory fails to describe basic qualitative features of conjoint item-source ratings. As discussed above, the correlation between item and source information observed at high levels of item confidence induces a reversal in source discriminability at very low levels of item confidence. The result is source ROCs conditionalized on item ratings that go below chance (see Figure 7 in Hautus et al., 2008). To rectify this problem and provide an increasingly good description of the data, Hautus et al. (2008) incrementally induced three changes to the DeCarlo (2003b) model: likelihood ratio bounds, the assumption of source guessing for items not

given a "yes" item response, and the assumption that some old items were sampled from a third distribution.[7] The location of this additional distribution was quite similar to the old-familiarity distribution used here.

The modeling approach of Hautus et al. (2008) in describing conjoint item-source ratings includes several potential advances over the methods used here. Rather than fitting separate source criteria for each level of item confidence, as was done here, likelihood ratios provide a principled way to generate curved source criteria across levels of item confidence. Interestingly, although we used a very different approach to generating source criteria, the results of our fitting result in at least a coarsely similar set of source criteria contours over the higher item confidence ratings (compare Figure 6 with Figure 8 in Hautus et al., 2008). However, the other modifications introduced by Hautus et al. (2008) are designed to overcome a weakness inherent in the multidimensional single-process account that generates the item-source distributions. This weakness in the multidimensional single-process model may be the factor that necessitated the guessing modification. We note that this weakness is not present if one starts from the assumptions of the VRDP about the output of the memory system.

The idea is that participants know that there are new items included in the probes, so that when presented with an old item that they have not rated as old, they do not process the probe for source information—which would result in below-chance source information if they used the DeCarlo (2003a) model to generate strength—and instead make a guess. Although Hautus et al. (2008) reached a different conclusion, this account seems to us to contradict the findings of Starns, Hicks, Brown, and Martin (2008) who observed that participants gave above-chance source ratings to old probes they rated as new if the response criterion was very conservative. Why wouldn't the participants in the Starns et al. (2008) study also forestall processing for the items they judged "new"? In any event, the guessing account proposed by Hautus et al. (2008) makes a straightforward prediction—if the item criteria are adjusted to be sufficiently liberal, the multidimensional item-source distribution should eventually give rise to below-chance source performance. Because the VRDP does not predict below-chance source performance under any circumstances, there is no need for a post-hoc explanation to explain why below-chance performance on the source test is never observed. If two-dimensional signal detection were a viable model of the output of the memory system, it should be possible to create a version of the item-source judgment that would give rise to worse-than-chance performance. This may be possible by changing the proportion of old items, or by not testing source for new items. Worse than chance performance on a source test for low item confidence old items would definitively falsify the VRDP.

The third modification adopted by Hautus et al. (2008) was to allow mixing between the unimodal multivariate distributions for old probes and an additional distribution the location and variance of which was allowed to vary freely, which turned out to have little-

[7] It has been suggested that the way to correct this misprediction of two-dimensional SDT is to simply assume that participants do not attempt a source rating for items they have already rated as old. This amounts to a mixture model, with the mixture being done at the stage of retrieval. The present approach uses a mixture interpreted as a difference in recollection and familiarity, with the benefit that it also provides the VRDP model with the ability to account for the item recognition data from travel scenes in Experiments 1 and 2 (see also Sherman, et al, 2003; Howard et al., 2006).

to-no source discriminability but above-chance item discriminability. This is very much like the approach used here.

Although Hautus et al. (2008) describe their model as "single-process," and note that it does not require recollection, they have maintained that degree of parsimony at the cost of introducing a post-retrieval guessing strategy and adding an additional discrete attentional process (DeCarlo, 2002, 2003a). The approach used here starts as a dual-process account. The addition of guessing and discrete encoding used by Hautus et al. (2008) require several more parameters than the VRDP to describe the distributions. Future work should directly compare the conjoint item-source distributions of the VRDP with those of the Hautus et al. (2008) model.

Independence and process purity

The VRDP provides strong evidence for the core assumption of dual-process theory—that recollection fails to contribute to memory for a substantial proportion of previously-experienced stimuli. A number of recent criticisms of dual-process theory address other aspects of dual-process theory. For instance, some criticisms of dual process theory are really criticisms of the all-or-none approximation (see Eq. 3) that is typically adopted in fitting the model to data (Dunn, 2004; Rotello, Macmillan, Reeder, & Wong, 2005; Wixted & Stretch, 2004). The VRDP assumes that in item recognition, recollection is continuous and subject to a signal detection decision, so that criticisms of previous treatments of the DPSD that rely on the all-or-none property of Eq. 3 do not apply to the VRDP.

Other criticisms of dual-process theory are directed at the assumption that recollection and familiarity are independent (e.g. Dunn, 2004). In our view, anatomical considerations make strict independence implausible, although not impossible, to support. Familiarity is typically hypothesized to depend on the functioning of extra-hippocampal medial temporal lobe (MTL) regions, whereas recollection is hypothesized to depend critically on the hippocampus proper (e.g. Eichenbaum, Yonelinas, & Ranganath, 2007; Norman & O'Reilly, 2003). Extra-hippocampal MTL regions, in particular the entorhinal cortex, provide the cortical input to the hippocampus. In turn, the hippocampus sends its output to these same regions. It seems implausible to us that the integrity of information about an item in extra-hippocampal regions, which presumably supports familiarity, would not have *some* effect on the functioning of the hippocampus. Similarly, it seems implausible that a strong output from the hippocampus, believed to support recollection, would not have *some* effect on the activity observed in extra-hippocampal MTL regions. Finally, we note that many studies that have purported to show neuropsychological manipulations that differentially affect recollection or familiarity rely on a specific model of how these putative processes map onto behavioral observations. If those behavioral models are incorrect, then the conclusions about recollection and familiarity they engender may need to be reevaluated, especially insofar as the empirical data reveals relatively complex changes in the shape of the ROC curves (e.g., Sauvage et al., 2008; Bowles et al., 2007).

In addition to the empirical criticisms of the DPSD, Wixted (2007b) also criticized the process-purity assumption of the DPSD on theoretical grounds. That is, according to the DPSD, participants respond to a probe on the basis of either recollection or familiarity rather than on the basis of their sum. If two sources of information are available, why wouldn't they be combined in order to make the decision? In addressing this question, we

first note that it is possible to build an additive model that is mathematically identical to the VRDP when just the item distributions are considered. If recollection contributed precisely d'_R worth of strength to each recollected old probe, then one would recover precisely the same equations used here if recollection and familiarity were summed prior to making a decision. This seems an unsatisfactory solution, however, in that it requires on the one hand that the contribution of recollection change in magnitude across participants, conditions and materials, but nonetheless show precisely zero variability for each participant in each condition.[8] The same argument would have to be made along the source dimension to account for the fact that our fits of conjoint item-source discriminability ended up making essentially circular two-dimensional distributions for the recollected old probes.

While we are sympathetic to the theoretical concerns expressed by Wixted (2007a), if one embraces the basic assumptions of signal detection theory, the empirical data seems to suggest that subject either rely on recollection or familiarity, rather than their sum, to make a decision. This may tell us something about the processes that give rise to recollection and familiarity. For instance, it is possible that the act of reconstructing the detailed context provided by an old item disrupts the familiarity signal such that it can no longer contribute to the decision. Another, more radical, alternative is that the VRDP does not reflect a signal detection process. That is, perhaps recollected items do not generate a strength that is then projected onto a decision access to generate a response. Perhaps recollection is the result of a multinomial process (Batchelder & Riefer, 1990) that is then mapped onto a variety of responses (Malmberg, 2002). That is, maybe recollection *per se* is not variable, but subjects vary their responses to recollected items randomly among a set of alternative responses. Viewed from this perspective, the difference in d'_R across words and travel scenes is understood not as recollected travel scenes generating more strength than recollected words. Rather, this can be understood as participants deciding to assign higher confidence ratings to recollected travel scenes than words. Perhaps this decision reflects a belief that a vivid recollection of a trial-unique travel scene is more diagnostic than recollection of a word that has been experienced many times before. While this approach would seem to fly in the face of much evidence in favor of signal detection as an explanation of recognition decisions (Mickes, Wixted, & Wais, in press; Wixted & Stretch, 2004), it would go a long way toward solving the problem of why the response distributions all appear to be normal with the same variance in mixture signal detection.

Toward a substantive model of item recognition

None of the signal detection models under consideration here make, in isolation, concrete predictions about how memory is encoded or retrieved. As such, by themselves they describe the data more than they explain it (Murdock, 2006). Of course an accurate description of recognition accuracy across criteria is a prerequisite for a more concrete model of episodic memory. Widespread acceptance of the VRDP could place meaningful constraints on the development of substantive models of recognition memory. Dennis and Humphreys (2001) categorized substantive mathematical models of recognition memory as item-noise (e.g. Shiffrin & Steyvers, 1997) or context-noise (e.g. Dennis & Humphreys,

[8] Although the results are not presented here, we have confirmed that the four-parameter model that consists of the VRDP with additional variability in the recollected distribution fails to fit the data as well as the VRDP as measured by the AIC and likelihood ratio testing.

2001) models according to whether they rely on interference from item representations or context representations (Criss & Shiffrin, 2004, used both to account for recognition performance). Recollection has been described as the recovery of spatial (Eichenbaum et al., 2007) or temporal (Schwartz, Howard, Jing, & Kahana, 2005) context. If the VRDP is correct, item-noise aspects of recognition could be addressed within the familiarity process, whereas contextual recovery is an excellent candidate for recollection. The fundamental constraint on substantive models offered by the VRDP is the requirement that recollection fails completely for a large proportion of old probes. When it succeeds, however, a great deal of information on which to base an item recognition decision becomes available. The threshold nature of recollection has been underappreciated by recent substantive models of item recognition (but see Elfman et al., 2008; Norman & O'Reilly, 2003).

Conclusions

We have proposed and evaluated a model that provides a general framework for measuring recognition memory performance, the variable-recollection signal detection model (VRDP). The model postulates that two distinct cognitive mechanisms, recollection and familiarity, govern item recognition. Familiarity is described as a continuous, normally-distributed process, whereas recollection is described as a continuous process that fails for some items. The VRDP model approximates the UVSD and the DPSD model as limiting cases, although it also provided a superior fit to the superset of the UVSD and DPSD. The VRDP is consistent with the qualitative properties of conjoint item-source judgments. A key prediction—that a large population of old probes recovers no source discriminability but reliable source information—was confirmed for both words and travel scenes. The VRDP may prove useful as a means to measure recollection and familiarity for applications in cognitive psychology, cognitive neuroscience and neuropsychology.

References

Aggleton, J. P., Vann, S. D., Denby, C., Dix, S., Mayes, A. R., Roberts, N., et al. (2005). Sparing of the familiarity component of recognition memory in a patient with hippocampal pathology. *Neuropsychologia*, *43*(12), 1810-23.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science*, *11*(4), 267-273.

Batchelder, W. J., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*(4), 548-564.

Bowles, B., Crupi, C., Mirsattari, S. M., Pigott, S. E., Parrent, A. G., Pruessner, J. C., et al. (2007). Impaired familiarity with preserved recollection after anterior temporal-lobe resection that spares the hippocampus. *Proceedings of the National Academy of Science, USA*, *104*(41), 16382-7.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach, 2nd edition*. New York: Springer.

Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: a comment on Dennis and Humphreys (2001). *Psychological Review*, *111*(3), 800-7.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: theoretical developments with applications to recognition memory. *Psychological Review*, *109*(4), 710-21.

DeCarlo, L. T. (2003a). An application of signal detection theory with finite mixture distributions to source discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 767-78.

DeCarlo, L. T. (2003b). Source monitoring and multivariate signal detection theory, with a model for selection. *Journal of Mathematical Psychology*, *47*, 292-303.

DeCarlo, L. T. (2007). The mirror effect and mixture signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 18-33.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(2), 452-78.

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: a review of arguments in favor of a dual-process account. *Psychonomic Bulletin and Review*, *13*(1), 1-21.

Dunn, J. C. (2004). Remember-know: a matter of confidence. *Psychological Review*, *111*(2), 524-542.

Duzel, E., Yonelinas, A. P., Mangun, G. R., Heinze, H. J., & Tulving, E. (1997). Event-related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences*, *94*(11), 5973-8.

Eichenbaum, H., Yonelinas, A., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*, 123-152.

Elfman, K. W., Parks, C. M., & Yonelinas, A. P. (2008). Testing a neurocomputational model of recollection, familiarity, and source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 752-68.

Farovik, A., Dupont, L. M., Arce, M., & Eichenbaum, H. (2008). Medial prefrontal cortex supports recollection, but not familiarity, in the rat. *Journal of Neuroscience*, *28*(50), 13428-34.

Fortin, N. J., Wright, S. P., & Eichenbaum, H. (2004). Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature*, *431*(7005), 188-91.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*(3), 546-67.

Glanzer, M., Hilford, A., & Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology*, *30*(6), 1176-95.

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*(2), 500-513.

Good, M. A., Barnes, P., Staal, V., McGregor, A., & Honey, R. C. (2007). Context- but not familiarity-dependent forms of object recognition are impaired following excitotoxic hippocampal lesions in rats. *Behavioral Neuroscience*, *121*(1), 218-23.

Hautus, M. J., Macmillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 889-905.

Healy, M. R., Light, L. L., & Chung, C. (2005). Dual-process models of associative recognition in young and older adults: Evidence from receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 768-788.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1210-30.

Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory and Language*, *55*, 495-514.

Hilford, A., Glanzer, M., Kim, K., & DeCarlo, L. T. (2002). Regularities of source recognition: Roc analysis. *Journal of Experimental Psychology*, *131*(4), 494-510.

Holdstock, J. S., Mayes, A. R., Roberts, N., Cezayirli, E., Isaac, C. L., O'Reilly, R. C., et al. (2002). Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? *Hippocampus*, *12*(3), 341-51.

Howard, M. W., Bessette-Symons, B. A., Zhang, Y., & Hoyer, W. J. (2006). Aging selectively impairs recollection in recognition memory for pictures: Evidence from modeling and ROC curves. *Psychology and Aging*, *21*, 96-106.

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

Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 701-22.

Macho, S. (2004). Modeling associative recognition: a comparison of two-high-threshold, two-high-threshold signal detection, and mixture distribution models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 83-97.

Macmillan, N. A., Rotello, C. M., & Miller, J. O. (2004). The sampling distributions of Gaussian ROC statistics. *Perception and Psychophysics*, *66*(3), 406-21.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 380-7.

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin & Review*, *13*(1), 99-105.

Meiser, T., & Bröder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 116-37.

Meiser, T., Sattler, C., & Weisser, K. (2008). Binding of multidimensional context information as a distinctive characteristic of remember judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 32-49.

Mickes, L., Wais, P. E., & Wixted, J. T. (2009). Recollection is a continuous process: implications for dual-process theories of recognition memory. *Psychological Science*, *20*(4), 509-15.

Mickes, L., Wixted, J. T., & Wais, P. E. (in press). A direct test of the unequal-variance signal-detection model of recognition memory. *Psychonomic Bulletin & Review*.

Murdock, B. B. (2006). Decision-making models of remember-know judgments: comment on Rotello, Macmillan, and Reeder (2004). *Psychological Review*, *113*(3), 648-56.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611-46.

Paller, K. A., Hutson, C. A., Miller, B. B., & Boehm, S. G. (2003). Neural manifestations of memory with and without awareness. *Neuron*, *38*(3), 507-16.

Parks, C. M., & Yonelinas, A. P. (2007a). Moving beyond pure signal-detection models: comment on Wixted (2007). *Psychological Review*, *114*(1), 188-202.

Parks, C. M., & Yonelinas, A. P. (2007b). Postscript: Comment on Wixted (2007). *Psychological Review*, *114*(1), 201-202.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472-91.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory ROC functions and implications for GMMs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 763-785.

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518-535.

Robitsek, R. J., Fortin, N. J., Koh, M. T., Gallagher, M., & Eichenbaum, H. (2008). Cognitive aging: a common decline of episodic recollection and spatial memory in rats. *Journal of Neuroscience*, *28*(36), 8945-54.

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: a two-dimensional signal-detection model. *Psychological Review*, *111*(3), 588-616.

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review*, *12*(5), 865-73.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Science*, *11*(6), 251-7.

Sauvage, M. M., Fortin, N. J., Owens, C. B., Yonelinas, A. P., & Eichenbaum, H. (2008). Recognition memory: opposite effects of hippocampal damage on recollection and familiarity. *Nature Neuroscience*, *11*(1), 16-8.

Schwartz, G., Howard, M. W., Jing, B., & Kahana, M. J. (2005). Shadows of the past: Temporal retrieval effects in recognition memory. *Psychological Science*, *16*(11), 898-904.

Sherman, S. J., Atri, A., Hasselmo, M. E., Stern, C. E., & Howard, M. W. (2003). Scopolamine impairs human recognition memory: Data and modeling. *Behavioral Neuroscience*, *117*, 526-539.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM — retrieving effectively from memory. *Psychonomic Bulletin and 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*, 151-170.

Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1499-1517.

Starns, J. J., & Hicks, J. L. (2005). Source dimensions are retrieved independently in multidimensional monitoring tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1213-20.

Starns, J. J., Hicks, J. L., Brown, N. L., & Martin, B. A. (2008). Source memory for unrecognized items: predictions from multivariate signal detection theory. *Memory & Cognition*, *36*(1), 1-8.

Uncapher, M. R., & Rugg, M. D. (2005). Encoding and the durability of episodic memory: a functional magnetic resonance imaging study. *The Journal of Neuroscience*, *25*(31), 7260-7.

Verfaillie, M., & Treadwell, J. R. (1993). Status of recognition memory in amnesia. *Neuropsychology*, *7*(1), 5-13.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192-6.

Wais, P. E., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2006). The hippocampus supports both the recollection and the familiarity components of recognition memory. *Neuron*, *49*(3), 459-66.

Wixted, J. T. (2007a). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152-76.

Wixted, J. T. (2007b). Spotlighting the probative findings: Reply to Parks and Yonelinas (2007). *Psychological Review*, *114*, 203-209.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*(4), 616-41.

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*(6), 1341-54.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: a formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1415-34.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441-517.

Yonelinas, A. P., Dobbins, I. G., Szymanski, M. D., Dhaliwal, H. S., & King, S. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition: An International Journal*, *5*, 418-441.

Yonelinas, A. P., Kroll, N. E., Quamme, J. R., Lazzara, M. M., Sauvé, M. J., Widaman, K. F., et al. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity. *Nature Neuroscience*, *5*(11), 1236-41.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience*, *25*(11), 3002-8.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychological Bulletin*, *133*(5), 800-32.

Yovel, G., & Paller, K. A. (2004). The neural basis of the butcher-on-the-bus phenomenon: when a face seems familiar but is not remembered. *Neuroimage*, *21*(2), 789-800.

		Δ AIC			Δ BIC		
		UVSD	DPSD	VRDP	UVSD	DPSD	VRDP
Exp. 1	words	63	180	0	0	117	419
	travel scenes	426	141	0	284	0	397
	combined	489	321	0	168	0	699
Exp. 2	words	17	213	0	0	196	313
	travel scenes	182	36	0	146	0	294
	combined	198	248	0	0	50	461

Table 5: AIC: Akaike's information criterion; BIC: Bayesian information criterion; UVSD: unequal variance signal detection model; DPSD: dual-process signal detection model; VRDP: variable recollection dual process model.

Appendix: Comparison of models with different numbers of parameters

The arguments in the body of the paper do not rely on direct comparison between models with different numbers of free parameters. Here we report comparisons between models with different numbers of parameters. Table 1 shows the results of calculations using Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for the UVSD, DPSD and VRDP to the item recognition response distributions from Experiments 1 and 2. Briefly, the AIC describes the VRDP as providing a vastly superior fit to the other two models in all conditions of both experiments. The corrected AIC (AICc) leads one to identical conclusions as the AIC owing to the large number of data points contributing to the response distribution across participants. The BIC shows contradictory results, favoring the best-fitting two-parameter model for all conditions. Collapsed across conditions, the BIC favors the DPSD for Experiment 1 (perhaps due to the fact that there were more participants in the travel scene condition) and the UVSD for Experiment 2.

The two measures of goodness-of-fit provided discrepant findings for the data in Table 1. The DPSD model is nested within the VRDP, hence it is appropriate to use a likelihood ratio (LR) test to compare these two models, providing an additional means to compare the two. If the null hypothesis is true, the LR statistic for the comparison between the two models ought to be distributed as chi-square with degrees of freedom equal to the number of participants. The results for Experiment 1 indicated that the VRDP model resulted in a much better fit that the DPSD for words, $\chi^2(104) = 360.5$, travel scenes, $\chi^2(116) = 282.5$, and combined across materials, $\chi^2(220) = 643.0$, all ps $< .001$. The results for Experiment 2 were largely consistent. The VRDP model resulted in a much better fit that the DPSD for words, $\chi^2(75) = 425.6$, $p < .001$, and combined across materials, $\chi^2(150) = 496.8$, $p < .001$, although for travel scenes, the models did not differ reliably, $\chi^2(75) = 71.2$, $p > .5$. Although we have not undertaken a formal model complexity analysis (Pitt, Myung, & Zhang, 2002), the AIC and LR tests provide strong support for the VRDP.

While the AIC and LR provide strong support, the BIC provides comparably strong support in favor of one or the other of the of the two-parameter models. There are a

couple of potential reasons for this discrepancy. First, however, we note that the results from the BIC are discrepant not only with the AIC and the LR ratio test, but also with the comparison of the VRDP with the superset of the UVSD and DPSD reported in the main body of the text (Table 2). If the BIC results reported in Table 5 are to be taken at face value, then we would expect that either the UVSD or DPSD is correct for a given participant, or perhaps a given list. However, the superset of the UVSD and DPSD was definitively rejected by non-parametric analyses.

One potential reason for the discrepancy between the AIC and the BIC for these data is that the assumptions used to derive the AIC and BIC do not hold for the VRDP. In particular, the derivation of both the AIC and BIC assumes that the likelihood of the data given the model falls off rapidly around the best-fitting parameters (Burnham & Anderson, 2002; Kass & Raftery, 1995). For instance, the BIC attempts to estimate the value of an integral across parameter space under the assumption that the likelihood of the data falls off rapidly around the best-fitting set of parameters. The difficulty in calculating this integral in practice leads to a step in which the value of the likelihood at the best-fitting point is taken to be the dominant term in the entire integral (e.g., Kass & Raftery, 1995). There are non-trivial states, in particular those consistent with an equal-variance signal detection model, for which the VRDP is degenerate. That is, under those circumstances an infinite range of parameter values generate the same model predictions. If the best-fitting parameters land in this region of the parameter space, the integral will not fall off rapidly around the best-fitting point. A similar step is present in the derivation of the AIC (Burnham & Anderson, 2002). In short, the assumptions used to derive the AIC and BIC do not hold for the VRDP, suggesting that caution should be exercised in using such statistics for specific models.