# Recollection and familiarity in recognition memory: Evidence from ROC curves

Andrew Heathcote[1], Frances Raymond[1] and John Dunn[2]

[1]School of Psychology, Aviation Building,

The University of Newcastle, Australia

[2]School of Psychology,

University of Adelaide, Australia

Contact Email: andrew.heathcote@newcastle.edu.au

PENULTIMATE DRAFT OF A PAPER IN *JOURNAL OF MEMORY AND LANGUAGE*

# Abstract

Does recognition memory rely on discrete recollection, continuous evidence, or both? Is continuous evidence sensitive to only the recency and duration of study (familiarity), or is it also sensitive to details of the study episode? Dual process theories assume recognition is based on recollection and familiarity, with only recollection providing knowledge about study details. Single process theories assume a single continuous evidence dimension that can provide information about familiarity and details. We replicated list (Yonelinas, 1994) and plural (Rotello, Macmillan & Van Tassel, 2000) discrimination experiments requiring knowledge of details to discriminate targets from similar non-targets. We also ran modified versions of these experiments aiming to increase recollection by removing non-targets that could be discriminated by familiarity alone. Single process models provided the best trade-off between goodness-of-fit and model complexity and dual process models were able to account for the data only when they incorporated continuous evidence sensitive to details.

Dual process theories propose that recognition memory is based on two qualitatively different kinds of memorial processes – recollection and familiarity (Yonelinas, 2002). Recollection is viewed as a discrete or all-or-none outcome that recovers details about the study episode through associations between the test item and aspects of the general study context, other studied items, and the physical characteristics of the studied item itself. For example, recollection of the study context can support list discrimination judgments (i.e., deciding which list an item was studied in), recollection of other studied items can support associative recognition judgments (i.e., deciding whether a pair of items was studied together), and recollection of the physical characteristics of a study item can support source memory (e.g., deciding if an item was heard in a male or female voice at study). Recollection can also play a role in item recognition. If recollected details are consistent with the test item, it may be classified as a target (*recollect-to-accept*), whereas, if recollected details are inconsistent with the test item, it may be classified as non-target (*recollect-to-reject*).

If recollection fails, decisions are based solely on familiarity. In contrast to recollection, familiarity "is assumed to be a relatively fast process that reflects the global familiarity or strength of an item" (Yonelinas, 1999a, p.1416). It provides a continuous value that conveys undifferentiated information about the duration, frequency and recency of prior exposure to a test item. Although familiarity conveys no information concerning specific details of the study episode, it can often be used as a reasonably reliable indicator of prior exposure, and so can support decisions in an item recognition paradigm (i.e., discriminating studied and unstudied items). Familiarity may also play a role in paradigms such as list discrimination, but only when study recency provides a cue for list membership (e.g., discriminating lists studied 5 minutes and 5 days ago, Yonelinas, 1999a).

The proposal that recognition memory is based on recollection and familiarity can be tested through examination of the shapes of receiver operating characteristic (ROC) curves (e.g., Yonelinas, 1999a). ROC curves plot, across different levels of decision confidence, the probability of a "yes" answer to the question posed by the recognition task for one type of test item (e.g., a target) against the probability of answering "yes" for another type of test

item (e.g., a non-target). In Yonelinas's (1994) dual process signal detection theory confidence ratings are based on criteria placed on familiarity. In item recognition paradigms larger familiarity values are associated with higher confidence that a test item is old (i.e., was studied) as, on average, familiarity for studied items is greater than familiarity for unstudied (new) items. Dual process signal detection theory also assumes that high confidence old responses can result from recollect-to-accept processes and that high confidence new responses can result from recollect-to-reject processes. Other versions of dual process theory have been proposed (see Yonelinas, 2002 for a review), but here we focus on dual process signal detection theory and refer to it simply as dual process theory.

We report tests of dual process theory using ROC data from a list discrimination paradigm, replicating Yonelinas (1994), experiment one, and an item recognition paradigm, replicating Rotello, Macmillan and Van Tassel (2000), experiment one. In these paradigms recollect-to-accept processes can be used to make decisions about targets and recollect-to-reject processes can be used to make decisions about non-targets that are very similar to targets. In the list discrimination paradigm, targets and similar non-targets correspond to items presented in two different study lists separated by a short pause. In the item recognition paradigm, similar non-targets differ only in plurality from studied items, half of which are in plural and half in singular form. For example, if *hand* is a studied, target, item then *hands* is a similar non-target item. In both the list and plural discrimination paradigms similar non-targets can be rejected on the basis of recollected details, concerning either the list context or the plurality of the studied item respectively. Both paradigms also included new non-target items that had not been studied in either list or in either plurality. For these items, recollection is assumed to play no role and decisions are based purely on familiarity, which would be less for these items than for either targets or similar non-targets.

In both the list discrimination and plurals paradigm, targets and similar non-targets have been chosen to have nearly equal levels of familiarity. In the list discrimination paradigm, targets and similar non-targets are drawn from different lists distinguished at study only by a short pause. In the plurals paradigm, similar non-targets differ from targets only in terms of

their plural form (e.g., hand vs. hands)[1]. When targets and similar non-targets have equal familiarity, dual process theory assumes that they can only be discriminated by discrete recollection. As a result, this theory predicts that the ROC curve relating targets to similar non-targets is linear (see Appendix A for details). Rotello et al. (2000) found an almost exactly linear ROCs of this type, but Yonelinas (1994) did not (see Figure 1, right panels). Linear ROCs have been found in related paradigms, such as associative recognition (Yonelinas, 1997) and source identification (Yonelinas, 1999a), that also equate familiarity. However, these results appear to be exceptions, with the majority of findings indicating non-linear ROCs in associative and source recognition paradigms (e.g., Glanzer , Hilford, & Kim, 2004; Healy , Light & Chung, 2005; Hilford , Glanzer, Kim and DeCarlo, 2002; Kelley & Wixted, 2001; Qin , Raye, Johnson & Mitchell, 2001; Slotnick , Klein, Dodson & Shimamura, 2000; Verde & Rotello, 2004).

It is possible that strategic factors may affect the shape of the relevant ROC curves. Such strategic factors could influence whether participants attempt recollection, and whether they use recollected details to accept or reject test items. In Rotello et al.'s (2000) first experiment participants were instructed to recollect-to-reject (i.e., respond "no" with high confidence if they recalled studying the test item in its alternative plurality), whereas Yonelinas (1994) did not give these instructions. Consistent with the influence of strategic factors, in a second experiment where Rotello et al. did not give recollect-to-reject instructions non-linear target versus similar non-target ROCs were found.

We investigated the role of strategic factors in two further experiments that replicated the original list and plural discrimination experiments with the exception that no new non-target items were presented at test. To distinguish the two sets of experiments, we refer to those that include new non-targets as "with-new" experiments and those that exclude new targets as "no-new" experiments. We speculated that new non-targets may increase reliance on familiarity, which can be used to discriminate these items from both targets and similar non-targets, and similarly discourage the use of recollected details, since such details are unlikely to be recollected for new non-targets. Hence, we hypothesised that the probability of

recollection would be greater in the no-new experiments compared to with-new experiments and that more linear target versus similar non-target ROC curves would result. As a corollary, we also hypothesised that target versus similar non-target discrimination would be better in the no-new than with new experiments due to the increase in information provided by recollection.

For both versions of the plurals paradigm we directly instructed participants to use a recollect-to-reject strategy, following Rotello et al.'s (2000) first experiment. Following Yonelinas's (1994), we did not give direct recollect-to-reject instructions in either list discrimination experiment.. We thus hypothesised that recollect-to-reject decisions would be more common in the plural discrimination experiments than in the list discrimination experiments.

*Single Process Theories*

In contrast to dual process theories, single process theories of recognition memory assume that decisions are based on a single continuous evidence dimension. These theories postulate that evidence is not restricted to familiarity; it can also be derived from flexible cue matching processes that are responsive to task demands. For example, Humphreys, Bain and Burt (1989) distinguished between two conceptually distinct types of evidence which they called *generalized strength* and *episode specific strength*. Generalized strength, like familiarity in dual process models, is "an amalgam of the frequency, recency and duration of exposure" (Chalmers & Humphrey, 1998, p.612). Generalized strength does not vary with test instructions, although instructions and task demands may cause decisions to be based to a greater or less degree on this form of memory. Episode specific strength, in contrast, is sensitive to test instructions through the types of cues that are used to probe memory. Generalized strength might itself be a product of a cue matching process, such as a match to the current (test) context, or it might arise from different processes, such conceptual implicit memory which has been proposed as the basis of familiarity in dual process theory (Yonelinas, 2002).

Evidence in single process theories can either be directly proportional to combined match

values (e.g., the global memory models, see Humphreys, Pike, Bain & Tehan, 1989) or based on a likelihood transformation of combined match values (e.g., Dennis & Humphreys, 2001; Shiffrin & Steyvers, 1997). Although theories that base evidence on a combination of flexible cue matching processes and familiarity processes are referred to as "single-process", and we will maintain this usage here, this term is really a misnomer given that evidence can be based on values obtained from more than one cue matching operation and potentially more than one type of process, and that these values can be combined and transformed by sophisticated decision mechanisms, such as likelihood. Evidence as conceived by single process theories has also been called familiarity. For clarity we will use the term "familiarity" in the sense intended by dual process theories and describe a continuous dimension that supports recognition decisions by the more general term "evidence".

Single process theories have been closely aligned with signal detection models of measurement in choice tasks (see Green & Swets, 1966; Macmillan & Creelman, 1991) as both assume that decisions are based on a single continuous strength-of-evidence dimension (see Wixted & Stretch, 2004, for further discussion). By examining the ability of a signal detection model to account for ROC data, it is possible to test the single process account. Dual process signal detection theory (Yonelinas, 1994) also incorporates a signal detection decision process, but one that limits continuous evidence to familiarity. Hence, according to dual process theory, equally familiar items can only be discriminated using recollection. In the present context, if targets and similar non-targets have equal levels of familiarity, the corresponding ROC curve must be linear. In contrast, according to single process theories, equally familiar targets and non-targets can be discriminated using continuous evidence. Consequently, single process theories predict that the corresponding ROC curve should be non-linear.

We also propose and test a specific version of the signal detection model developed to account for the list discrimination and plurals paradigms (see Appendix B for mathematical details). The basic idea behind this model is that discrimination of targets and similar non-targets is based on the difference of strengths of two cue matching processes. For this reason,

we refer to it as the cue-matching model. In a list discrimination paradigm, it is assumed that cues representing both study list contexts are used to probe memory. Evidence that a test item appeared on the target list is proportional to the difference between the match strength of the item to the target list cue and the match strength of the item to the non-target list cue. Similarly, differences between matches to singular and plural cues are used to provide evidence in the plurals paradigm. The set of difference values defines a single strength-of-evidence dimension. Relatively large values on this dimension support a target decision while relatively small values support a non-target decision. We will describe such continuous cue-dependent measures that provide a basis for judgments concerning task relevant details of test items as specific strength. The notion of specific strength is more general than Humphreys et al.'s (1989) episodic specific strength in that it can be based on specific details of a study item (e.g., its plurality) as well as specific details of the study episode (e.g., study list).

In single process theories, both generalized and specific strength may contribute to the total strength-of-evidence, with the relative weight of each depending on task demands. For example, if some non-target test items in a list discrimination task are new (i.e., not studied in either list), evidence could consist of a weighted combination of generalized strength, relevant to the discrimination of old from new items, and specific strength, relevant to the discrimination of target from non-target list items. If there are no new non-targets at test then we assume that evidence depend only on specific strength. This leads the cue-matching model to predict that discrimination between targets and similar non-targets should be better in the no-new than in the with-new experiments. This occurs because the addition of generalized strength makes evidence noisier and hence less reliable.

*ROC Analysis*

Most single process theories assume that evidence arising from cue match values is based on the sum of a large number of randomly varying factors, such as matches to many different memory traces in the global memory models (see Humphreys, Pike, Bain & Tehan, 1989). This implies that evidence based on cue matches will be approximately normally distributed, and a similar logic applies to familiarity as well as to specific strength. When evidence is

normally distributed, the ROC curve has a characteristic concave[2] shape. In addition, item recognition ROC curves are typically asymmetrical (for a summary, see Ratcliff, Gronlund & Sheu, 1992)[3]. In order to explain this, single process theories assume that the strength-of-evidence of studied items is more variable than that of unstudied items because, among other reasons, the effects of study are likely to vary across trials due to fluctuations in attention and encoding processes. Dual process theory, in contrast, assumes that familiarity has the same variance for both studied and unstudied test items (Yonelinas, 1994) and explains the asymmetry of ROC curves due to recollection. This results in an increase in the y-intercept to a value equal to the probability of recollection, with the asymmetry of the ROC curve increasing as recollection increases. The ROC curves for similar non-targets versus new non-targets reported by Yonelinas (1994) and Rotello et al. (2000) were so asymmetric that they dipped below the main diagonal at the rightmost point (see Figure 1, left panels). This "dip effect" has a plausible explanation in terms of recollect-to-reject processing, whereby some similar non-targets are correctly recollected to have been studied either in their alternative plurality or in the non-target list, leading to high confidence non-target responses. When such details are not recollected, the greater familiarity of similar non-targets causes them to be more often mistakenly classified as targets, so the remainder of the ROC is above the main diagonal.

The dip effect can also be accommodated by a signal detection model if the variance of the evidence distribution for similar non-targets is greater than that for new non-targets. In this case, the ROC asymmetry associated with unequal variance can result in a dip effect as long as the mean strength-of-evidence of similar non-targets is not much greater than the mean of new non-targets. This illustrated in Figure 2 which plots the probability and cumulative probability functions for the best fitting unequal variance signal detection model based on the average data from Rotello et al. (2000), experiment one. The new non-target distribution is used as a reference with a mean fixed at zero and a standard deviation fixed at one. The similar non-target (plural new) distribution has close to the same mean (-0.07) but has a much larger standard deviation (1.75). As a result, as the decision criterion shifts to the

left, the proportional increase in the cumulative response rate is less for the high variance distribution than for the low variance distribution resulting in the slope of the ROC curve to be less than one. If this is marked enough then the curve will dip below the main diagonal, as was the case for both Rotello et al's data and the fitted signal detection model.

Although the signal detection model can describe the dip effect, how can a single process memory theory explain this pattern of evidence distribution parameters? We show in Appendix B that the cue matching model predicts that similar non-target variance is greater than new non-target variance, and that their means can be close to equivalent, as observed in the fits illustrated in Figure 2. However, it is difficult to make exact predictions about evidence variability when ROC data are averaged across participants, because averaging confounds individual differences in mean evidence variability with variability of evidence within a participant. Averaging also risks confounding by ceiling effects, which have a larger effect on target than similar non-target conditions, and so can cause target variance to be underestimated relative to similar non-target variance[4]. We fit models to our data at the individual level to avoid potential problems caused by averaging.

*Extending the Dual Process Model*

The concave ROC curves found by Yonelinas (1994) in list discrimination, and in the majority of associative and source recognition studies (e.g., Glanzer et al., 2004; Healy et al., 2005; Hilford et al., 2002; Kelley & Wixted, 2001; Qin et al., 2001; Slotnick et al., 2000; Verde & Rotello, 2004) are inconsistent with the predictions of dual process theory, since familiarity should be approximately equal for targets and non-targets in these paradigms. In contrast, single process theories are consistent with these results as they allow targets and non-targets to differ in specific strength. In light of these results we propose and test an extended dual process model that replaces an evidence dimension based purely on familiarity with one that can also take account of specific strength. Including specific strength in evidence allows the extended dual process model to accommodate concave ROC curves for test items with equal familiarity.

The extended model is consistent with evidence supporting dual process theory (e.g.,

Yonelinas, 2002), because it allows both recollection and familiarity to play a role in recognition decisions. It is also consistent with single process theories that model recall data (e.g., Gillund & Shiffrin, 1984), as their recall mechanisms could, at least in principle, account for recollection. We use the acronym DP to denote the original dual process model (Yonelinas, 1994) and the acronym DP-s to denote the extended dual process model that allows specific strength to contribute to continuous evidence. For our experiments, the crucial difference between these models is that the DP-s model allows evidence to differ between the target and similar non-target conditions, whereas the DP model does not.

A second potential extension of dual process theory concerns its assumption that familiarity has the same variance for both targets and non-targets. This assumption implies that ROC asymmetry is associated with convex[5] z-ROC curves. However, in item recognition paradigms, Glanzer, Kim, Hilford and Adams (1999) and Heathcote (2003) found asymmetric ROC curves but no evidence of convex z-ROC curves (but see Yonelinas, 1999a, 1999b). Heathcote suggested that dual process theory should be revised to allow studied items to have greater familiarity variance than unstudied test items. We use the acronym DP-su to denote an extended dual process theory that allows for both unequal familiarity variance and specific strength.

*ROC Models and Model Testing*

In summary, we test three variants of dual process theory; the original DP model (Yonelinas, 1994), the DP-s model, and the DP-su model. We compare each of these models with two single process models. The first, denoted by the acronym SP, is the unequal variance signal detection decision model assumed by most single process memory models. The second, denoted by the acronym SP-c, is derived from the cue-matching theory detailed in Appendix B. The SP-c model differs from the SP model only in the SP-c model target and similar non-target evidence standard deviations are assumed equal, whereas they can be unequal in the Sp model. All five models are defined in Appendix A and their basic features are summarized in Table 1[6].

The five models in Table 1 differ in the number of parameters that are to be estimated

from data. Models with more parameters may provide a better fit simply because they are more flexible. To compensate for this greater flexibility, we compared the models using criteria that combine goodness-of-fit, as measured by the model's maximum (negative) log-likelihood ($l$), with a complexity penalty. The best model is the one with the lowest criterion value. We use two such criteria, the Akaike Information Criterion (AIC) that adjusts for the number of model parameters, $p$, and the Bayesian Information Criterion (BIC) that also adjusts for the number of observations ($n$). That is,

$$l = \sum_i \sum_j f_{ij} \ln(p_{ij}) \qquad AIC = -2l + 2p \qquad BIC = -2l + p\ln(n)$$

The summation is over $i$=1..$k$ conditions (e.g., targets and non-targets) and $j$=1..$m$ response categories (i.e., confidence levels for target and non-target choices), where $f_{ij}$ and $p_{ij}$ are, respectively, the number of observations and corresponding probabilities predicted by the model. Myung and Pitt (1997) have noted that AIC tends to favour more complex models when fits are based on a large number of observations, as was the case for our data. Neither criterion takes into account complexity due to differences in the functional form of models[7].

We also report a measure of goodness-of-fit, $G = 2\sum_i \sum_j f_{ij}\left(f_{ij}/F_{ij}\right)$, where $F_{ij}$ is the expected frequency based on the maximum likelihood estimates of the model. $G$ is approximately distributed as $\chi^2$ with $n$-$p$ degrees of freedom[8] and so its sum over $N$ participants is also distributed as $\chi^2$ with $N(n$-$p)$ degrees of freedom. Differences in the summed $G$ values can be used to construct $\chi^2$ tests with $Nq$ degrees of freedom of whether the addition of $q$ parameters to a model causes a significant increase in fit. For dual process models we focus on comparisons between DP and DP-s and between DP-s and DP-su, which test whether specific strength improves the fit of the DP model, and whether unequal evidence variance improves the fit of the DP-s model. For single process models we compare SP-c and SP models, to test whether unequal target and similar non-target evidence variance improves fit.

## Experiment 1: List Discrimination

Experiment one was designed to replicate experiment one from Yonelinas (1994) using the list discrimination paradigm. We modified the design in two ways. First, in the

original experiment, at each of six experimental sessions, 800 words were randomly sampled from the 1000 high frequency concrete nouns that make up the Toronto word pool. We chose to use a larger word pool in order to avoid repeating items across sessions. Second, the original experiment compared short (10 item) and long (30 item) study list lengths. As both conditions produced similar results, we used a single study list length of 20 items. As a result, we required only three experimental sessions to obtain the same number of observations as collected by Yonelinas (1994).

## Method

### *Participants*

Ten members of the staff and student body of the Faculty of Medicine and Dentistry at the University of Western Australia participated in this study. Data from four participants were excluded due either to failure to complete the three sessions (1), or unacceptably high error rates (3). We report data from the remaining six participants. Participants were offered thirty dollars reimbursement upon completion of the experiment to assist with travel/parking expenses. Following explanation of the task, written consent was obtained prior to commencement of the experiment.

### *Apparatus and stimuli*

Stimuli were 1440 words drawn from the Toronto word pool (Friendly, Franklin, Hoffman & Rubin, 1982) and augmented by words selected from the MRC Psycholinguistic Database, Version 2.00 (Coltheart, 1981). The latter items were all concrete nouns and had the same frequency distribution as items in the Toronto word pool. The average Kučera and Francis (1967) word frequency rating for all the words was 73. For each participant, 480 items were randomly selected without replacement to be used in each of the three sessions. An additional 50 words were used for practice at the beginning of the first session.

Experimental tasks were administered with an IBM desktop PC using Cedrus SuperLab Pro (v2.0) software. Stimuli were presented on a 15" CRT monitor. Responses were collected via a Superlab Pro RB830 8-button response pad. The six top-most buttons, arranged in two horizontal arcs, were labelled (from left to right) "sure new", "probably new", "guess new",

"guess old", "probably old" and "sure old". The two remaining buttons were not used.

*Procedure*

There were three experimental sessions each conducted on a separate day. Each session consisted of eight study-test cycles. Each cycle consisted of two 20-item study lists followed by two 30-item test lists. Each test list was composed of 10 items from each of the study lists and 10 new items. No item appeared in more than one study list or test list. Before the first session, an additional study-test cycle was administered as practice. Data from this cycle was not included in analyses. Instructions appeared on the screen prior to presentation of the experimental stimuli. Participants were advised that they would see two study lists, followed by two tests lists, and that they were to try to remember each of the words in the study list as well as the list in which it had appeared. Before each test list, one of the two study lists in the present cycle was nominated as the target list. Each study list was nominated as the target list once in each cycle with the order of nomination randomized across cycles. Participants were told to rate their degree of confidence that a test item had been presented in the target list. A 6-point scale was used where 1=sure new, 2=probably new, 3=guess new, 4=guess old, 5=probably old, and 6=sure old. Participants were instructed to respond as quickly, but as accurately as possible. Instructions for the practice and experimental tasks were identical.

Each study item was presented for two seconds and there was no inter-stimulus interval. There was a five second pause between each pair of study lists. At the end of the second study list the prompt for the first test appeared. Test items remained on the screen until a response was made. Participants had the opportunity to rest for as long as they liked between each study-test cycle. Each participant completed each of three sets of eight study-test cycles on a separate day with one to three days between each session.

## Results and Discussion

Figure 3 presents the ROC curves and fits of the DP and SP-c models obtained by averaging both the data and the best fitting model predictions over participants. Our results replicate those of Yonelinas (1994), showing a dip effect for the similar versus new non-target ROC. The SP-c model accurately fits both the target versus new non-target and the

similar non-target versus new non-target ROC curves. In contrast, the DP model systematically underestimates the target versus new non-target ROC curve and overestimates the similar non-target versus new non-target ROC curve. In Table 2, and following tables of results, we report $G'=G/\min(G)$, where $\min(G)$ is minimum $G$ value for set of models. By definition $\min(G)$ occurs for the most complex, and hence best fitting, model in the set (DP-su in all cases). $G'$ results in Table 2 indicate that the fit of the DP model is worse by a factor of almost 4 than the SP-c fit.

Table 2 also shows the performance of each model after adjusting for complexity using the *AIC* and *BIC* criteria. In Table 2, and following tables, we report $AIC'=AIC-\min(AIC)$ and $BIC'=BIC-\min(BIC)$, where min() is minimum criterion value for a set of models. Hence, the model selected by *AIC* has $AIC'=0$ and the model selected by *BIC* has $BIC'=0$. Overall, *BIC* selected the SP-c model, indicating that it provided the best compromise between goodness of fit and model complexity among the set of five models. These results indicate that recollection, and unequal variance between targets and similar non-targets, do not provide a sufficient increase in fit to justify the additional estimated parameters that they require. The SP model had only a slightly better fit than the SP-c model, by a factor of 1.13, but this increase was significant, $\chi^2(6)=18.1$, $p=.006$.

The more complexity tolerant *AIC* method selected the most complex model, the DP-su model, which incorporates both recollection and evidence of the type assumed by single process theories, with both unequal variance and specific strength. Comparison with the DP and DP-s fits shows that allowing specific strength provides an improvement in fit by a factor of 2.5, $\chi^2(6)=377$, $p<.001$. Comparison of the DP-s and DP-su models shows that once specific strength is allowed, unequal variance improves fit by a further factor of 2.6, $\chi^2(6)=154$, $p<.001$. Hence both modifications of the DP model were supported.

The DP-su model estimates that recollection occurred for about a quarter of trials. The average probabilities of correct responses for targets and similar non-targets were 0.70 and 0.83 respectively. The DP-su model attributes 39% and 25% respectively of these correct responses to recollection. Hence, when recollection is assumed to occur, the majority of

correct responses are based on continuous evidence. Consist with this attribution, targets and similar non-targets *d*' estimates for the DP-su model differed by 1.4, indicating good discrimination on the basis of evidence.

# Experiment 2: Plural Discrimination

Experiment two replicates experiment one from Rotello et al. (2000). Our design differed slightly from Rotello et al. in that we used five study-test cycles, as opposed to three in the original experiment. We also tested only 16 items of each type (target, similar non-target, new non-target) for each list, whereas Rotello et al. tested 24 of each type. Therefore, across all study-test cycles, we collected 80 observations for each item type, slightly more than the 72 observations collected in the original experiment.

## Method

### *Participants*

Twenty-four undergraduate psychology students from the University of Newcastle participated in the study in exchange for course credit. Data from five participants were removed from further analysis because of a failure to use the middle (guess) confidence ratings and because of low accuracy indicating a lack of engagement with the task. Data are thus reported from a total of 19 participants.

### *Apparatus and stimuli*

Stimuli were drawn from a pool of 374 singular concrete nouns selected from the MRC Psycholinguistic Database, Version 2.00 (Coltheart, 1981), for which plural forms could be created by adding an "s" (e.g., *tree-trees*). All stimuli were 3-12 characters in length, with a mean word frequency of 71 (Kučera & Francis, 1967). For each participant, 286 singular–plural word pairs were selected randomly from the pool for study. Either the singular or plural form from each pair was selected at random, with an equal number of singular and plural words in each study list. The study words were divided into six sets, one list of 24 words was used for practice and five sets of 48 words for experimental study lists. The remaining words in the study set were used as untested buffers added to the beginning and end of the study lists, giving a practice study list of 26 words, and experimental study lists of

52 words.

For each test list, 16 studied words (8 for practice) were randomly selected to be presented in the same plurality as at study (old items), 16 study words (8 for practice) were selected at random to be presented in the alternate plurality as at study (similar lures), along with 16 (8 for practice) unstudied words (new lures). Test word presentation order was randomized and half of the words were singular and half were plural for each condition. New lure items were not repeated in subsequent test lists.

Experimental tasks were administered with an IBM compatible desktop PC using Cedrus SuperLab Pro (v2.0) software. Stimuli were presented on a 17" monitor and responses collected via a SuperLab Pro RB830 8-button response pad. The six top-most buttons, arranged in two horizontal arcs, were labelled (from left to right) "sure old", "probably old", "guess old", "guess new", "probably new" and "sure new". The two remaining buttons were not used.

*Procedure*

The task was explained, and written consent obtained, prior to commencing the practice first session. Participants were informed they would be presented with six study lists, each followed by a test list, and that the first study-test cycle was for practice. Each study word was presented for 3 sec, with an inter-stimulus interval of 1 sec. Participants were informed that they should try to remember the words, as their memory for them would be tested. Immediately following each study list, participants were presented with instructions on the screen indicating that they would be presented with the test list as soon as they pressed any button. They were instructed to decide if each test word had been presented in the study list using a 6-point rating scale (1=sure old, 2=probably old, 3=guess old, 4=guess new, 5=probably new, 6=sure new). Following each test phase, participants were asked to press any button when they were ready to begin the next study-test cycle.

Before the experiment began, participants were instructed to pay particular attention to the plurality of study and test words and to respond old only to a test word if it had been studied in exactly the same form. All participants were given the example that if they could

remember studying the word *cats*, but they were presented with the test word *cat*, they could be confident that *cat* was a new item. Participants were also informed that one-third of the test words had not been studied in either plurality and they were to classify these as new items as well. After the practice cycle, participants received feedback showing their accuracy and the number of times they had used each response category, and were reminded to make sure that they used all confidence levels.

## Results and Discussion

Figure 4 presents the ROC data and fits of the DP and SP-c averaged over participants. Comparison with Figure 3 indicates a stronger dip effect in plural discrimination than list discrimination, which is consistent with a comparison of results from Rotello et al. (2000) and Yonelinas (1994) (see Figure 1). The SP-c model was able to accommodate the stronger dip effect, whereas the DP model shows the same systematic over and under estimation as occurred in list discrimination. The DP model fares better in terms of goodness-of-fit than in list discrimination, but is still worse than the SP-c model by a factor of $1^2/_3$. Both *AIC* and *BIC* criteria select the SP-c model first and the SP model second. These results favouring the cue matching model are stronger than in list discrimination, as even the more complexity tolerant *AIC* method indicates that recollection, and unequal variance between targets and similar non-targets, do not provide a sufficient increase in fit to justify the additional estimated parameters that they require.

The relative fits of SP and SP-c models were similar to those found in experiment one. The SP model fit better than the SP-c fit by a factor of only 1.14, but in this case, the increase in fit was not significant, $\chi^2(19)=25.7$, $p=.14$. This supports the cue matching theory from which the SP-c model was derived. Consistent with this model evidence variance for targets and similar non-targets were almost equal.

According to *AIC* the DP-su model was the highest ranked dual process model and according to *BIC* the DP-s model was highest ranked. Comparison with the fits of the DP and DP-s models shows that addition of specific strength improves the fit by a factor of 1.55, $\chi^2(19)=118$, $p<.001$. Comparison of the DP-s and DP-su models shows that if specific

strength is allowed, allowing unequal variance improves fit by a further factor of 1.9, $\chi^2(38)=101$, $p<.001$.

Given that both modifications of the DP model received support we examine the DP-su model parameter estimates to determine the relative roles of recollection and evidence in arriving at correct decisions. All dual process models estimated lower rates of recollection in this experiment than in experiment one. The relative probabilities of recollect-to-accept and recollect-to-reject were similar in both experiments. Both results are inconsistent with our hypothesis that the recollect-to-reject instructions given in the plural discrimination experiment would increase both the overall level of recollection and the relative level of recollect-to-reject processing compared to the list discrimination experiment, where no such instructions were given.

Parameter estimates for the DP-su model indicated that evidence plays a much larger role in discriminating targets and similar non-targets than recollection. Target and similar non-target *d'* estimates differed by 1.45, a slightly greater difference than in list discrimination. The average probabilities of a correct response for targets and similar non-targets were 0.67 and 0.71 respectively. The DP-su model attributes 21% and 20% respectively of these correct responses to recollection. Hence, about 80% of correct responses are attributable to continuous evidence. This value is also slightly higher than for list discrimination, perhaps because targets were more familiar than similar non-targets in the plural discrimination paradigm. However, the increase is only small, and so an explanation of plural discrimination performance purely in terms of familiarity seems unlikely.

## Experiments 3 and 4

Experiments three and four replicate experiments one and two respectively, except that no new non-targets were tested. In the absence of new non-targets, the modelling environment is altered slightly as detailed in Appendix A. Since there are now only two test item types, then without loss of generality, the evidence distribution of similar non-targets may be taken as a reference distribution. Thus, the mean and variance of this distribution is assumed to be zero and one, respectively. As a consequence, the DP model is a pure discrete recollection model.

This version of the DP model predicts linear target versus similar non-target ROCs, as found by Rotello et al. (2000).

## Method

### *Participants*

In the list discrimination task (experiment three), the participants were eight members of the staff and student body of the Faculty of Medicine and Dentistry at the University of Western Australia. They were paid $30 for their participation and none had previously participated in experiment one. In the plural discrimination task (experiment four), participants were twenty-four undergraduate psychology students from the University of Newcastle. They received course credit and none had previously participated in experiment two.

### *Stimuli and apparatus*

Word pools and apparatus were the same as in experiments one and two. In list discrimination 320 items were randomly selected without replacement to be used in each of the three sessions for each participant. An additional 50 words were used for practice at the beginning of the first session.

### *Procedure*

The procedure for list discrimination was the same as that used in experiment one with the exception that each test lists contained no new lures. As a consequence, each test list consisted 20 items consisting of a random mixture of 10 old items and 10 similar lures. Participants were told that the test list would consist only of items that had been presented in the two study lists. They were instructed to use the six-point rating scale and to respond "old" if a test word had been presented in the target list (target) and to respond "new" if the word had been presented in the alternative study list (similar non-target).

The procedure for plural discrimination was the same as that used in experiment two, except that the set of new non-targets presented at test were replaced by an equal number of similar non-targets. For each experimental list, 16 of the studied words (8 for practice) were selected randomly to be presented in the same plurality at test (targets) and the remaining 32

studied words (16 for practice) were presented in their alternate plurality (similar non-targets). Thus, the probability of an target at test was the same as in experiment two. Participants were also informed that $^1/_3$ of the test items would be targets and that the remainder would be non-targets.

## Results and Discussion

Figures 5 and 6 show that target versus similar non-target ROCs were clearly concave in all experiments. As a result the DP model badly misfits these ROCs. In the no-new experiments, the best the DP model could do was to join the end points of the curve[9]; in the with-new experiments, misfit is evident for the end points as well. The SP-c model provided an accurate account of these ROCs in all experiments, despite having one less parameter than the DP model in the no-new experiments, and was selected by both the *AIC* and *BIC* criteria in both no-new experiments. The *G'* measure indicated that in the no-new experiments, the SP-c model fit was almost 10 times better than the DP model fit in the list discrimination, and about 3 $^1/_3$ times better in the plural discrimination.

From a single process perspective, these results support the cue matching theory underlying the SP-c model as it was selected ahead of the SP model by both *AIC* and *BIC* in both no-new experiments. As in the with-new experiments, allowing unequal evidence variance for targets and similar non-targets did not greatly increase fit in the no-new experiments. Neither the increase in fit by a factor of 1.25 for list discrimination, nor the increase by a factor of 1.19 for plural discrimination, was significant, $\chi^2(6)=11.5$, $p=.07$, and $\chi^2(23)=29.4$, $p=.17$, respectively. The SP model fits did consistently estimate greater evidence variance for targets than similar non-targets in both no-new experiments, but only by factors of 1.13 and 1.09 for list and plural discrimination respectively. This supports the cue matching theory from which the SP-c model was derived.

Examination of Figures 5 and 6 also confirms the cue matching theory's prediction that accuracy is greater in the no-new than with-new experiments, as indicated by both no-new ROCs being above the corresponding with-new ROCs. We used hit (H) and false alarm (FA) rates for targets and similar non-targets respectively to calculate $d'=z(H)-z(FA)$ in order to

measure discrimination between targets and similar non-targets and tested the cue matching model's prediction of greater discrimination for no-new than with-new experiments using one-tailed t-tests on these *d*' scores. The mean *d*' was significantly greater in no-new (2.12) than with-new (1.53) list discrimination experiments, t(9)=1.89, *p*=.046. The mean *d*' was also significantly greater in with-new (1.53) than no-new (1.07) plural discrimination experiments, t(38)=1.87, *p*=.035.

Comparison of *G*' for the DP-s and DP models in Tables 4 and 5 reveals that allowing specific strength in evidence, and hence curvature in target versus similar non-target ROCs, produced a large increase in fit, by a factor of 11, $\chi^2(12)=519$, $p<.001$, for list discrimination and a factor of 5.1, $\chi^2(46)=498$, $p<.001$, in plural discrimination. Once specific strength was included, allowing unequal evidence variance produced only a small and non-significant increase in fit, by factors of 1.28, $\chi^2(6)=11$, $p=.09$, and 1.22, $\chi^2(23)=21$, $p=.56$, for list and plural discrimination respectively. These results for dual process models are consistent with the results for single process models in indicating that the difference in evidence variance between targets and similar non-targets was negligible. They contrast with dual process model results for with-new experiments, where allowing unequal evidence variance produced a much larger increase in fit. This occurred because new items clearly had much lower evidence variance than target and similar non-target items, and so in the with-new experiments the DP-s model, which assumes equal variance for all test items types, was inadequate.

Given the DP-s model was favoured in the no-new experiments we examine its parameters to quantify the role played by discrete recollection. In experiment three, the average probability of a correct response was 0.85 and 0.84 for target and similar non-target trials respectively. The DP-s model attributes 15% and 8% respectively of these correct responses to recollection. In experiment four, the average probability of a correct response was lower, at 0.76 and 0.77 for targets and similar non-targets respectively. The DP-s model attributes 33% and 25% respectively of the correct responses to recollection. Consistent with this pattern, the average *d*' estimate for the DP-s model was greater in list discrimination

(1.89) than in plural discrimination (0.98). In contrast to the with-new experiments, these results are not consistent with better discrimination based on continuous evidence in the plurals paradigm than the list discrimination paradigm due to greater familiarity for targets than similar non-targets.

We hypothesised that dual process models could predict an increase in performance in no-new compared to with-new experiments if the exclusion of new test items encouraged recollection. However, estimates of recollection derived from the DP-s model indicate lower recollection rates in the no-new experiments compared to with-new experiments. Recollection estimates reduced from 34% and 29% of trials, on average, for with-new list and plural discrimination respectively to 10% and 22% for corresponding no-new experiments. Comparison of DP-su recollection estimates for list discrimination also revealed a reduction from 24% to 9%, although for plural discrimination, recollection estimates increased from 14% to 19%. Overall, these results are more consistent with increased accuracy in no-new experiments being due to reduced evidence noise, as predicted by the cue matching model, rather than to an increase in recollection. In addition, there was also no evidence that explicit recollect-to-reject instructions in experiment four increased the probability of recollection for similar non-targets compared to experiment three in which participants were not given such instructions.

## General Discussion

We examined two issues fundamental issues for recognition memory, whether decisions are based on discrete recollection, continuous evidence, or both, and whether continuous evidence is restricted to familiarity, providing only information about the recency and duration of study, or whether continuous evidence can also convey specific information about details of a study episode. Dual process theory assumes that both recollection and continuous evidence contribute to recognition decisions, with recollection providing information about details of the study episode and evidence restricted to the generalized strength or familiarity of a test item (Yonelinas, 2002). Single process theories assume that recognition decisions are based only on continuous evidence containing details of the study episode derived from task

dependent cue matching processes, which provides what we have called *specific strength*.

We examined these issues using Receivers Operating Characteristic (ROC) curves, parametric plots of the probability of accepting test items of different types as a function of decision confidence. ROC analysis has been a particular focus of Yonelinas's (1994) dual process signal detection theory. This theory assumes that decisions are based on recollection when it occurs, leading to high confidence responses. When recollection fails decisions are based on a familiarity, which is continuous and distributed with equal variance for all types of test items, leading to graded confidence responses depending on the magnitude of familiarity. We compared the predictions of the dual process model with the predictions of single process models where all decisions are based on a signal detection decision process applied to a continuous evidence dimension. In the single process models different types of test items can have differing evidence variance, and evidence can consist of familiarity and/or specific strength, depending on task demands.

Single and dual process accounts were tested with data from a list discrimination paradigm (Yonelinas, 1994) and a plural discrimination paradigm (Rotello et al., 2000). Importantly, these paradigms attempted to minimize differences in familiarity between targets and non-targets that are very similar to targets (i.e., test items studied in a non-target list or with the opposite plurality to a studied item). When familiarity is equated the dual process model predicts linear target versus similar non-target ROCs because decisions can only be based on discrete information provided by recollection. The single process model, in contrast, predicts concave ROCs in these paradigms. While our experiments replicated most of the findings of Yonelinas (1994) and Rotello et al., particularly an unusual "dip effect", wherein the ROC curve for similar non-targets versus new non-targets (i.e., items not studied in either list or plurality) deviates below the main diagonal to the right, we did not replicate Rotello et al.'s finding of a linear target versus similar non-target ROC curve. Instead, we found pronounced concave ROC curves. We also replicated these findings in a further experiment that did not test new non-targets.

We proposed and tested two modified dual process models. One, the DP-s model, allowed

specific strength as well as familiarity to contribute to continuous evidence. The other, the DP-su model, further allowed unequal evidence variance. Like the single process model, the modified dual process models permits targets and similar non-target with equal familiarity to be discriminated on the basis of specific strength consistent with the concave target versus similar non-target ROC curves which were observed. We also proposed a single process model based on the idea that specific strength is derived from the difference between matches to target and non-target cues. This cue matching model predicts that evidence variance for targets and similar non-targets should be equal and that greater discrimination between targets and similar non-targets should be found when new non-targets are not included in testing.

Overall, our results are inconsistent with Yonelinas's (1994) original dual process theory in both list discrimination and plural discrimination paradigms, both with and without new test items. In contrast, our results are consistent with the single process cue matching model in all four experiments. It can be concluded that, at least in these experiments, continuous evidence is not restricted to familiarity. Instead, evidence can also convey specific information about details of the study episode, such as list membership and the plurality in which an item was studied. Yonelinas (1999) speculated that continuous evidence might support source discrimination "in conditions in which the item and source information are more closely integrated, such as may be the case when two individuals are holding a conversation" (p.1416). Our results indicate that continuous evidence can support discriminations based on source in more impoverished situations where source and item information are not closely integrated, as was the case for our list discrimination experiments where the two sources were only distinguished by a small pause between otherwise homogenous study lists.

Our results do not necessarily reject the idea of a discrete recollection processes in recognition memory, but only when continuous evidence is allowed to convey information about details and to have unequal variance, at least for new items relative to other types of test items. When allowed these extensions, dual process models assuming discrete

recollection provided a very accurate account of the data. However, recollection was estimated to be the basis of only a minority of correct responses, about 23% on average, and the improvement of fit afforded by recollection was, in all but one of eight cases, insufficient to warrant the increase in model complexity which it entails. It would be of interest to know if this is also the case in other paradigms or for other groups of participants who, for whatever reason, place greater strategic emphasis on recollection. For example, increased reliance on a discrete recollection process would enable the extended dual process model which we have proposed to account for the few linear ROCs that have been reported in similar paradigms to those we considered here (Rotello et al., 2000; Yonelinas, 1997, 1999). However, a strong reliance on recollection appears to be the exception rather than the rule. Further, we did not find that reliance on recollection was influenced by strategic factors, such as giving explicit recollect-to-reject instructions, as was the case in our plural discrimination experiments but not our list discrimination experiments.

An alternative possibility, and one that has received increasing support recently (e.g., Kelley & Wixted, 2001; Rotello, Macmillan, Reeder & Wong, 2005; Wixted & Stretch, 2004), is that recollection, like cue matching, can produce continuous, or at least graded, evidence, especially in situations where a rich array of details about the study context and study item are available. The experimental paradigms we have examined have usually been assumed not to have this richness. However, it is possible that in some circumstances recollection acts like a discrete process and in others like a graded or even continuous process, depending on the strategies which participants adopt to encode study items. In this view, what we have called specific strength might be thought to be, at least in part, the result of a more finely graded recollective process, and the few exceptional cases where linear ROC curves have been observed may be attributed to particularly impoverished encoding.

Our results provided clear support, in terms of a balance between goodness-of-fit and simplicity, for the single process cue matching model which we proposed (see Appendix B for details). We also confirmed two predictions made by this model; that targets and similar non-targets should have the same evidence variance and that discrimination of targets and

similar non-targets is reduced when new non-targets are included in testing. The latter prediction is not unique; it could also be made by dual process models if excluding new non-targets encourages recollection. However, this possibility was not confirmed by recollection parameter estimates derived from fits of dual process models to our data. Instead, these fits generally supported decreased recollection and increased discrimination on the basis of continuous evidence when new non-targets were not tested. Further investigation of this effect is warranted as other factors may have also played a part, such as differential encoding of study items, which might have occurred because participants were aware of the makeup of test lists before they commenced study.

The cue matching theory demonstrates how a single process memory theory is able to provide a plausible explanation of the pattern of parameters estimated by fitting the normal unequal variance signal detection model of choice. Although this theory does not specify the processes by which matches are obtained nor the representations used for cues and memory traces, Clark (1997) implemented a closely related model for two alternative forced choices between targets and similar non-targets using Hintzman's (19988) MINERVA theory of recognition memory, which does make these aspects explicit. Clark assumed that forced choice decisions are based on a difference between matches to target and non-target test alternatives, just as our cue matching theory assumes that yes/no choices are based on the difference between matches to target and non-target cues. Given the success of the cue matching model, future research might test more detailed implementations using MINERVA, other global matching models, or theories which assume that evidence is based on a likelihood transformation of cue match strength (e.g., Dennis & Humphreys, 2001; Shiffrin & Steyvers, 1997). An alternative possibility is provided by theories that combine two sources of continuous evidence as assumed by in two-dimensional signal detection (e.g., Banks, 2000; Rotello, Macmillan & Reeder, 2004). Although space constraints preclude provision of any details, we have found that two-dimensional signal detection models can provide an accurate account of our ROC data with the same economy of estimated parameters as the cue-matching model.

# Appendix A: ROC Models

According to dual process theory, two distinct sources of information can potentially support recognition judgments. The probability of correctly identifying a target depends upon the probability that it is recollected and, if recollection fails, on the probability that its familiarity exceeds a given criterion value. Formally, this is represented by the following equation:

$$P_T(c) = r_T + (1 - r_T)(1 - \Phi(c - d'_T)) \quad (1)$$

$P_T(c)$ is the probability of correctly identifying a target at the level of confidence corresponding to the decision criterion, *c*; $r_T$ is the probability of recollecting the target; $d'_T$ is the mean level of familiarity of the targets; and $\Phi(.)$ is the normal cumulative distribution function. In the case of non-targets, the probability of incorrect identification as a target depends upon the probability that recollection fails and that familiarity exceeds the specified criterion value. This probability is given by the following equation:

$$P_N(c) = (1 - r_N)(1 - \Phi(c - d'_N)) \quad (2)$$

$P_N(c)$ is the probability of incorrectly identifying a non-target as a target at the level of confidence corresponding to *c*, $r_N$ is the probability of recollection and $d'_N$ is the mean level of general familiarity of non-targets. It is assumed that if details relevant to discriminating targets and non-targets are recollected a correct judgment is always made (Yonelinas, 1999a). For targets, this represents a *recollect-to-accept* strategy, while for non-targets, it represents a *recollect-to-reject* strategy.

Given Equations (1) and (2), it is possible to derive the form of the ROC for targets vs. non-targets when familiarity is the same for both; that is, when $d'_T = d'_N$. In this case, after rearranging the terms in the two earlier equations, we find the following equation for $P_T(c)$ as a function of $P_N(c)$:

$$P_T(c) = r_T + \frac{(1 - r_T)}{(1 - r_N)} P_N(c) \quad (3)$$

Equation (3) depends solely on recollection and is unaffected by familiarity. It describes a straight line with a y-intercept (i.e. the value of $P_T(c)$ when $P_N(c) = 0$) equal to $r_T$ and a

slope equal to the ratio of $1-r_T$ to $1-r_N$. Thus, under this model, a linear ROC is indicative of equal levels of familiarity for targets and non-targets.

In single process theories, discrimination between targets and non-targets is based on a single strength-of-evidence dimension. Hence:

$$\begin{aligned} P_T(c) &= 1-\Phi\left((c-d'_T)/s_T\right) \\ P_N(c) &= 1-\Phi\left((c-d'_N)/s_N\right) \end{aligned} \tag{4}$$

Here, $s_T$ and $s_N$ are the standard deviations of the strength-of-evidence distributions of targets and non-targets, respectively. If $d'_T > d'_N$ this equation results in a typically concave ROC curve that intersects the y-axis at the origin. Its shape depends upon the ratio of the two standard deviations. If $s_T > s_N$ then the curve is asymmetrical with the equal likelihood point (the point at which the slope of the ROC curve is equal to one) shifted relatively to the left, as is frequently observed in item recognition tasks in which non-targets are unstudied items. If $s_T = s_N$ the curve is symmetrical about the anti-diagonal. If strength-of-evidence is based only on familiarity, which is the same for targets and non-targets (i.e. $d'_T = d'_N$ and $s_T = s_N$), then the ROC curve reduces to a straight line given by,

$$P_T(c) = P_N(c) \tag{5}$$

Thus, if familiarity in dual process models and strength-of-evidence in single process models corresponds to generalized strength, in the sense proposed by Chalmers and Humphreys (1998), then in situations where targets and non-targets are equally familiar, the resulting ROC curve is necessarily a straight line. The crucial difference between the models is that the dual process model predicts that the linear ROC can fall above the main diagonal because of recollection. In contrast, if strength-of-evidence in single process models, at least in part, consists of specific strength able to support the discrimination required by the task, the resulting ROC curve will be concave, even if targets and non-targets have equal familiarity.

*The generalized dual process signal detection model*

We start with a generalized dual-process signal detection model and derive specific cases from it. Let $C = (c_1, c_2, c_3, c_4, c_5)$ be a non-decreasing sequence of values on a familiarity

dimension such that $c_1 \geq c_2 \geq c_3 \geq c_4 \geq c_5$. In experiments one and two, there were three types of item presented at test; *target*, *similar non-target* and *new non-target*. In experiments three and four, there were two types of item; *target* and *similar non-target*. A set of five models was fit to the data from experiments one and two (with-new experiments) and experiments three and four (no-new experiments). These models can each be viewed as a constrained version of a *generalized dual process signal detection model*, named the DP-su model in Table 1. The DP-su model has a different form in the with-new and no-new experiments.

The generalized dual process signal detection model for the with-new experiments is defined as follows. Let $P_T(c_i)$ be the probability of responding "yes" to a target for some $c_i$ in *C*. Similarly, let $P_{SN}(c_i)$ and $P_{NN}(c_i)$ be the probabilities of responding "yes" to a similar non-target and a new non-target, respectively. Then,

$$\begin{aligned} P_T(c_i) &= r_T + (1-r_T)\left(1-\Phi\left(c_i - d'_T\right)/s_T\right) \\ P_{SN}(c_i) &= (1-r_{SN})\left(1-\Phi\left(c_i - d'_{SN}\right)/s_{SN}\right) \\ P_{NN}(c_i) &= 1-\Phi(c_i) \end{aligned} \tag{6}$$

Here, $r_T$ is the probability of recollecting an old item ("recollect-to-accept"), $r_{SN}$ is the probability of recollecting a similar non-target ("recollect-to-reject"), $d'_T$ is the mean evidence for target items, $d'_{SN}$ is the mean evidence of similar non-targets, $s_T$ is the standard deviation of target evidence, and $s_{SN}$ is the standard deviation of similar non-target evidence. Evidence is scaled with reference to the new non-target distribution whereby the mean evidence of new non-targets is zero (i.e., $d'_{NN} = 0$) and the corresponding standard deviation is one (i.e., $s_{NN} = 1$). The function, $\Phi(.)$ is the normal cumulative distribution function.

The generalized dual process signal detection model for the no-new experiments is defined as follows:

$$\begin{aligned} P_T(c_i) &= r_T + (1-r_T)\left(1-\Phi\left(c_i - d'_T\right)/s_T\right) \\ P_{SN}(c_i) &= (1-r_{SN})\left(1-\Phi(c_i)\right) \end{aligned} \tag{7}$$

In this case, evidence is scaled with reference to the similar non-target distribution. Thus, the mean evidence of similar non-targets is zero (i.e., $d'_{SN} = 0$) and the corresponding standard deviation is one (i.e., $s_{SN} = 1$).

*Guessing*

Equations (6) and (7) may also be modified to incorporate guessing on some proportion of trials, *g*. In this case, participants are assumed to select a response category at random with an equal probability (across of categories) of 1/6. Let *J* be the set of conditions in an experiment and let $j \in J$ . Let $P_j(c_i, g)$ be the probability of an "yes" response for some $c_i$ under condition *j* and for some probability of guessing, *g*. Then,

$$P_j(c_i, g) = gQ(c_i) + (1-g)P_j(c_i) \tag{8}$$

Where $Q(c_i) = \sum_r^i q_r$ is the cumulative probability of selecting each response category *r* for $r = 1 \ldots i$ and $q_r = 1/6$ for all $r = 1 \ldots 6$.

*Model generation*

In Table 1 the generalized dual process signal detection model is indicated by the acronym DP-su. Each of the remaining models listed in Table 1 can be generated from the DP-su model by applying each of three constraints corresponding to the "no" entries in Table 1. These constraints are,

*Zero recollection*: $r_T = r_{SN} = 0$ (9)

*Equal variance*: $s_T = s_{SN} = 1$ (10)

*Familiarity evidence*: $d'_T = d'_{SN}$, $s_T = s_{SN}$ (11)

To illustrate the procedure, the DP model is derived from the DP-su model by applying both the equal variance and familiarity evidence constraints (indicated by the "no" entries in the corresponding row of Table 1). This leads to the following equations for the DP model for the with-new experiments, derived from Equation (6),

$$\begin{aligned} P_T(c_i) &= r_T + (1-r_T)(1-\Phi(c_i - d')) \\ P_{SN}(c_i) &= (1-r_{SN})(1-\Phi(c_i - d')) \\ P_{NN}(c_i) &= 1-\Phi(c_i) \end{aligned} \tag{12}$$

For the no-new experiments the DP model is derived from Equation (7),

$$\begin{aligned} P_T(c_i) &= r_T + (1-r_T)(1-\Phi(c_i)) \\ P_{SN}(c_i) &= (1-r_{SN})(1-\Phi(c_i)) \end{aligned} \tag{13}$$

It should be noted that the signal detection part of Equation (13) may be interpreted in one of two different ways. First, it may be interpreted as implementing a genuine signal detection process based on identical familiarity distributions for both the targets and similar non-targets. Second, it may be equally well interpreted as implementing a biased guessing process in which the probability of guessing each response category is parameterized in terms of a decision criterion, *c*. Although, for reasons of consistency and comparability between models, this process is given this parameterization, it is formally equivalent to guessing each response category *i* with some probability, $g_i$, under the constraint that $\sum_i g_i = 1$.

The SP model is obtained by applying the zero recollection constraint. The SP-c model is obtained by applying the further constraint that $s_{SN}=s_T$. In the no-new experiments the latter constraint is equivalent to the equal variance constraint, whereas in the with-new experiments the SP-c model allows the standard deviation for new non-targets to differ from the standard deviation of targets and similar non-targets.

# Appendix B: Cue Matching Model

Let *I* be a test item and let *T*(*I*) and *SN*(*I*) be the match strength of *I* to target and non-target cues, respectively. For example, in the plurals discrimination paradigm, *T*(*I*) is the strength of the match between *I* and a cue corresponding to the same plural form at study, while *SN*(*I*) is the strength of the match between *I* and a cue corresponding to the alternative plural form at study. The specific strength that *I* is a target, $s_T(I)$ is given by the difference between these two match strengths: $s_T(I)=T(I)–SN(I)$.

Let *T* be a target item, let *SN* be a similar non-target item, and let *NN* be a new non-target item. Then, $s_T(T)=T(T)–SN(T)$, $s_T(SN)=T(SN)–SN(SN)$, and $s_T(NN)=T(NN)–SN(NN)$. We further assume that $T(T)=SN(SN)>T(SN)=SN(T)>T(NN)=ST(NN)$ yielding three distinct match strengths, denoted by *a*, *b*, and *c*. We assume that each of these is normally distributed with means, $\mu_a > \mu_b > \mu_c$ and standard deviations, $\sigma_a > \sigma_b > \sigma_c$. Hence, the specific strength of each item, *I*, is also normally distributed. Let $\mu_I$ and $\sigma_I$ be the mean and standard deviation of the distribution of $s_T(I)$. Then,

$$\mu_T = \mu_a - \mu_b \qquad \sigma_T^2 = \sigma_a^2 + \sigma_b^2$$
$$\mu_{NS} = \mu_b - \mu_a \qquad \sigma_{NS}^2 = \sigma_b^2 + \sigma_a^2$$
$$\mu_{NN} = \mu_c - \mu_c \qquad \sigma_{NN}^2 = \sigma_c^2 + \sigma_c^2$$

It follows that $\mu_T = -\mu_{NS}$ and $\mu_{NN} = 0$, and $\sigma_T = \sigma_{NS} > \sigma_{NN}$.

In the "no-new" condition, no new items are presented at test and we assume that decisions are based solely on the specific strength, $s_T(I)$. Hence, relative to the non-target condition, $d'_{T/SN} = (\mu_T - \mu_{SN})/\sigma_T = 2\mu_T/\sigma_T$. Similarly, $d'_{SN/SN} = (\mu_{SN} - \mu_{SN})/\sigma_{SN} = 0$.

In the "with-new" condition new items are presented at test and we assume that decisions are based on the sum of specific strength, $s_T(I)$, and familiarity, $F(I)$. We further assume that $F(I)$ is normally distributed with mean, $\mu_{F(I)}$, and standard deviation, $\sigma_{F(I)}$, and that $\mu_{F(T)} = \mu_{F(SN)} > \mu_{F(NN)}$ and $\sigma_{F(T)} = \sigma_{F(SN)} > \sigma_{F(NN)}$. Let $\varepsilon(I) = s_T(I) + F(I)$. It follows that,

$$\mu_{\varepsilon(T)} = \mu_T + \mu_{F(T)} \qquad \sigma_{\varepsilon(T)}^2 = \sigma_T^2 + \sigma_{F(T)}^2$$
$$\mu_{\varepsilon(SN)} = -\mu_T + \mu_{F(T)} \qquad \sigma_{\varepsilon(SN)}^2 = \sigma_{\varepsilon(T)}^2$$
$$\mu_{\varepsilon(NN)} = \mu_{F(NN)} \qquad \sigma_{\varepsilon(NN)}^2 = \sigma_{NN}^2 + \sigma_{F(NN)}^2$$

Hence, d' values for the target condition relative to the new non-target condition (*T*/*NN*),

$d'_{T/NN} = \left(\mu_{\varepsilon(T)} - \mu_{\varepsilon(NN)}\right)\big/\sigma_{\varepsilon(NN)} = \left(\mu_T + \mu_{F(T)} - \mu_{F(NN)}\right)\big/\sigma_{\varepsilon(NN)}$. For the similar non-target condition relative to the new non-target condition, $d'_{SN/NN} = \left(-\mu_T + \mu_{F(T)} - \mu_{F(NN)}\right)/\sigma_{\varepsilon(NN)}$ and $d'_{NN/NN} = \left(\mu_{F(NN)} - \mu_{F(NN)}\right)/\sigma_{\varepsilon(NN)} = 0$.

Let $d'_n$ above be the relative discriminability of targets and non-targets in the in the "no-new" condition (given by $d'_{T/SN}$ above) and let $d'_w$ be the corresponding discriminability in the "with-new" condition. Then,

$$\begin{aligned} d'_w &= \left(d'_T - d'_{SN}\right)\left(\sigma_{\varepsilon(NN)}/\sigma_{\varepsilon(T)}\right) \\ &= 2\mu_T / \sqrt{\sigma_T^2 + \sigma_{F(T)}^2} \\ &= d'_n\left(\sigma_T / \sqrt{\sigma_T^2 + \sigma_{F(T)}^2}\right) \end{aligned}$$

Thus, the relative discriminability of targets and similar non-targets in the "with-new" condition is reduced from the "no-new" condition by a factor that depends upon the relative variance of familiarity compared to the variance of episode specific strength. For example, if the respective variances are equal such that $\sigma_{F(T)} = \sigma_T$, then $d'_w = d'_n/\sqrt{2} = 0.707 d'_n$. More generally, if $\sigma_{F(T)}^2 = k\sigma_T^2$, then $d'_w = d'_n/\sqrt{1+k}$ and, solving for *k*, we have $k = \left(d'_n/d'_w\right)^2 - 1$.

Consistent with these predictions, we found that the ratio of $d'_T$ in the no-new experiment to $d'_T$-$d'_{SN}$ in the with-new experiments was greater than one. For list discrimination this ratio was 1.51 and for plural discrimination it was 1.60. According to the cue matching theory, the size of these ratios depends on the relative standard deviations of specific strength and familiarity. The ratio for the list discrimination experiment indicates that the standard deviation of specific strength was 22% greater on average than the standard deviation of familiarity. In the plural discrimination experiments this value is slightly larger, at 36%.

# Acknowledgements

Thanks to Julie Johnston for running the list discrimination experiments and to reviewers for constructive suggestions. Thanks also to the Department of Psychology, University of Western Australia, for making the MRC Psycholinguistic Database available on the web (http://www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm), and to the Computational Memory Lab (http://memory.psych.upenn.edu/wordpools.php) at the University of Pennsylvania for making the Toronto word pool available on the web.

# References

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

Chalmers, K. A. & Humphreys, M. S. (1998). Role of generalize and episode specific memories in the word frequency effect in recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*, 610-632.

Clark, S. E. (1997). A familiarity-based account of confidence-accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(1), 232-238.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497-505.

Dennis, S. & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*, 452-478.

Friendly, M., Franklin, P.E., Hoffman, D., & Rubin, D.C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods and Instrumentation, 14(4),* 375-399.

Gillund, G. & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1-67.

Glanzer, M., Hilford, A., Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*(6), 1176-1195.

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*, New Your: Robert E. Kreiger Publishing.

Healy, M.R., Light, L.L., & Chung, C. (2005). Dual-process models of associative recognition in younger 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), 1210-1230.

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple - trace memory model. *Psychological Review, 95*, 528-551.

Humphreys, M. S., Bain, J. D., & Burt, J. S. (1989). Episodically unique and generalized memories: Applications to human and animal amnesics. In S. Lewandowsky, J. C. Dunn, & K. Kirsner *(Eds.), Implicit memory: Theoretical issues (pp.* 139-156). Hillsdale, NJ: Erlbaum.

Humphreys, M. S., Pike, R., Bain, J. D. & Tehan, G. (1989). Global matching: A comparison of SAM, Minerva II, Matrix and TODAM models. *Journal of Mathematical Psychology*, *33*, 36-67.

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), 701-722.

Kučera, H. & Francis, W.N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

MacMillan, N.A., & Creelman, C.D. (1991). *Detection Theory: A users guide*. New York: Cambridge University Press.

Myung, I. J. & Pitt, M.A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79-95.

Pitt, M.A., Myung, I. J. & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472-491.

Qin, J., Raye, C.L., Johnson, M.K., & Mitchell, K.J. (2001). Source ROCs are (typically)

curvilinear: Comment on Yonelinas (1999). *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*, 1110-1115.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*(4), 763-785.

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

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,* 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*, 865-873.

Rotello, C. M., Macmillan, N.A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence form ROC curves. *Journal of Memory and Language, 43*, 67-88.

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., 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.

Wixted, J. T. & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*, 616-641.

Verde, M. F., & Rotello, C. M. (2004). Strong memories obscure weak memories in associative recognition. *Psychonomic Bulletin & Review, 11*, 1062-1066.

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), 1341-1354.

Yonelinas, A. P. (1997). Recognition memory ROC's for item and associative information: The contribution of recollection and familiarity. *Memory and Cognition, 25*, 747-763.

Yonelinas, A. P. (1999a). 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), 1415-1434.

Yonelinas, A. P. (1999b). Recognition memory ROCs and the dual-process signal detection model: Comment on Glanzer, Kim, Hilford and Adams (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 514-521.

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

# Figure Captions

Figure 1. Yonelinas's (1994), experiment one, participant average list discrimination ROCs (measured from his Figure 4 and averaged over short, 10 item, and long, 30 item, study list conditions) and participant average ROCs for Rotello et al.'s (2000) experiment one (measured from their Figure 3).

Figure 2. Distribution and cumulative distribution functions for an unequal variance normal signal detection model fit to Rotello et al.'s (2000) experiment one data. Vertical dotted lines indicate the estimated criterion for the new-old decision (middle line) and for confidence ratings.

Figure 3. Experiment one participant average ROC data (circles) and fits of the Dual Process (DP) and Single Process cue (SP-c) models averaged over participants for experiment one. Upper symbols and lines are target versus new non-target. Lower symbols and lines are similar non-target versus new non-target.

Figure 4. Experiment two participant average ROC data (circles) and fits of the Dual Process (DP) and Single Process cue (SP-c) models averaged over participants for experiment two. Upper symbols and lines are target versus new non-target. Lower symbols and lines are similar non-target versus new non-target.

Figure 5. Participant average target versus similar non-target ROC data and fits of the Single Process cue (SP-c) and Dual Process (DP) models averaged over participants for experiment one (With New) and experiment three (No New).

Figure 6. Participant average target versus similar non-target ROC data and fits of the Single Process cue (SP-c) and Dual Process (DP) models averaged over participants for experiment two (With New) and experiment four (No New).

# Tables

**Table 1**

Definition of models according to the presence or absence of each of three assumptions. DP=dual process signal detection assuming only familiarity based evidence with equal variance, SP=single process signal detection, *s*=evidence with a specific (cue-dependent) component, *u*=unequal evidence variance.

| | Assumption | | |
|---|---|---|---|
| Model Acronym | Recollection | Unequal Variance | Specific Evidence |
| DP-su | Yes | Yes | Yes |
| DP-s | Yes | No | Yes |
| DP | Yes | No | No |
| SP | No | Yes | Yes |
| SP-c | No | New Only | Yes |

**Table 2**

Summary of models (summed fit and criterion measures and mean parameter estimates over participants) for Experiment 1. See Table 1 and Appendix for model definitions. *G*'=*G*/97.68, *AIC*'=*AIC*-22963.5 and *BIC*'=BIC-23109.5. Estimated parameter values and minimum *G*', *AIC*' and *BIC*' values are in bold type. Degrees of freedom (*df*) are given for *G* values. The *r* parameters are recollection probability estimates, the *d*' and *s* parameters are evidence mean and standard deviation estimates and the subscripts T and N refer to target and similar non-targets respectively. The five decision criterion parameters for each model are not shown.

| | Summed Model Fit | | | | Participant Average Parameter Estimates | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | *G'* | *AIC'* | *BIC'* | *df* | $r_T$ | $r_N$ | $s_T$ | $s_N$ | $d'_T$ | $d'_N$ |
| DP-su | **1** | **0** | 19 | 24 | **.27** | **.21** | **1.56** | **1.53** | **1.99** | **0.39** |
| DP-s | 2.58 | 131 | 119 | 36 | **.35** | **.32** | 1 | | **1.43** | **0.77** |
| DP | 6.45 | 496 | 470 | 42 | **.39** | **.39** | 1 | | **1.16** | |
| SP | 1.45 | 20 | 9 | 36 | 0 | | **1.97** | **1.84** | **2.85** | **-0.31** |
| SP-c | 1.64 | 26 | **0** | 42 | 0 | | **1.88** | | **2.79** | **-0.33** |

**Table 3**

Summary of model fits (summed fit and criterion measures and mean parameter estimates over participants) for Experiment 2. See Table 1 and Appendix A for model definitions. *G*'=*G*/112.33, *AIC*'=*AIC*-13292.0 and *BIC*'=*BIC*-13847.1. Estimated parameter values and minimum *G*', *AIC*' and *BIC*' values are in bold type. Degrees of freedom (*df*) are given for *G* values. The *r* parameters are recollection probability estimates, the *d*' and *s* parameters are evidence mean and standard deviation estimates and the subscripts T and N refer to target and similar non-targets respectively. The five decision criterion parameters for each model are not shown.

| | Model fit | | | | Participant Average Parameter Estimates | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | *G'* | *AIC'* | *BIC'* | *df* | $r_T$ | $r_N$ | $s_T$ | $s_N$ | $d'_T$ | $d'_N$ |
| DP-su | **1** | 27 | 235 | 76 | **.14** | **.14** | **1.57** | **1.51** | **1.63** | **0.18** |
| DP-s | 1.90 | 52 | 121 | 114 | **.29** | **.28** | 1 | | **0.63** | **0.95** |
| DP | 2.94 | 131 | 131 | 133 | **.31** | **.27** | 1 | | **0.77** | |
| SP | 1.55 | 12 | 82 | 114 | 0 | | **1.87** | **1.88** | **2.09** | **-0.34** |
| SP-c | 1.77 | **0** | **0** | 133 | 0 | | **1.87** | | **2.13** | **-0.33** |

**Table 4**

Summary of model fits (summed fit and criterion measures and mean parameter estimates over participants) for Experiment 3. See Table 1 and Appendix A for model definitions. *G*' = *G*/39.67, *AIC*'=*AIC*-14703.6 and *BIC*'=*BIC*-14827.4. Estimated parameter values and minimum *G*', *AIC*' and *BIC*' values are in bold type. Degrees of freedom (*df*) are given for *G* values. The *r* parameters are recollection probability estimates, the *d*' and *s* parameters are evidence mean and standard deviation estimates and the subscripts T and N refer to target and similar non-targets respectively. The five decision criterion parameters for each model are not shown.

| | Model fit | | | | Participant Average Parameter Estimates | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | *G'* | *AIC'* | *BIC'* | *df* | $r_T$ | $r_N$ | $s_T$ | $s_N$ | $d'_T$ | $d'_N$ |
| DP-su | 1 | 18 | 80 | 6 | **.11** | **.07** | **1.11** | 1 | **2.03** | 0 |
| DP-s | 1.28 | 17 | 59 | 12 | **.13** | **.07** | 1 | | **1.89** | 0 |
| DP | 14.08 | 513 | 534 | 18 | **.54** | **.51** | 1 | | 0 | |
| SP | 1.15 | 1 | 21 | 18 | 0 | | **1.13** | 1 | **2.22** | 0 |
| SP-c | 1.44 | **0** | **0** | 24 | 0 | | 1 | | **2.07** | 0 |

**Table 5**

Summary of model fits (summed fit and criterion measures and mean parameter estimates over participants) for Experiment 4. See Table 1 and Appendix A for model definitions. *G*'=*G*/95.14, *AIC*'=*AIC*-16152.9 and *BIC*'=*BIC*-16627.3. Estimated parameter values and minimum *G*', *AIC*' and *BIC*' values are in bold type. Degrees of freedom (*df*) are given for *G* values. The *r* parameters are recollection probability estimates, the *d*' and *s* parameters are evidence mean and standard deviation estimates and the subscripts T and N refer to target and similar non-targets respectively. The five decision criterion parameters for each model are not shown.

| | Model fit | | | | Participant Average Parameter Estimates | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | *G'* | *AIC'* | *BIC'* | *df* | $r_T$ | $r_N$ | $s_T$ | $s_N$ | $d'_T$ | $d'_N$ |
| DP-su | **1** | 53 | 290 | 23 | **.18** | **.20** | **1.16** | 1 | **1.12** | 0 |
| DP-s | 1.22 | 28 | 186 | 46 | **.25** | **.19** | 1 | | **0.98** | 0 |
| DP | 6.24 | 459 | 539 | 69 | **.38** | **0.34** | 1 | | 0 | |
| SP | 1.59 | 17 | 96 | 69 | 0 | | **1.09** | 1 | **1.64** | 0 |
| SP-c | 1.89 | **0** | **0** | 92 | 0 | | 1 | | **1.54** | 0 |

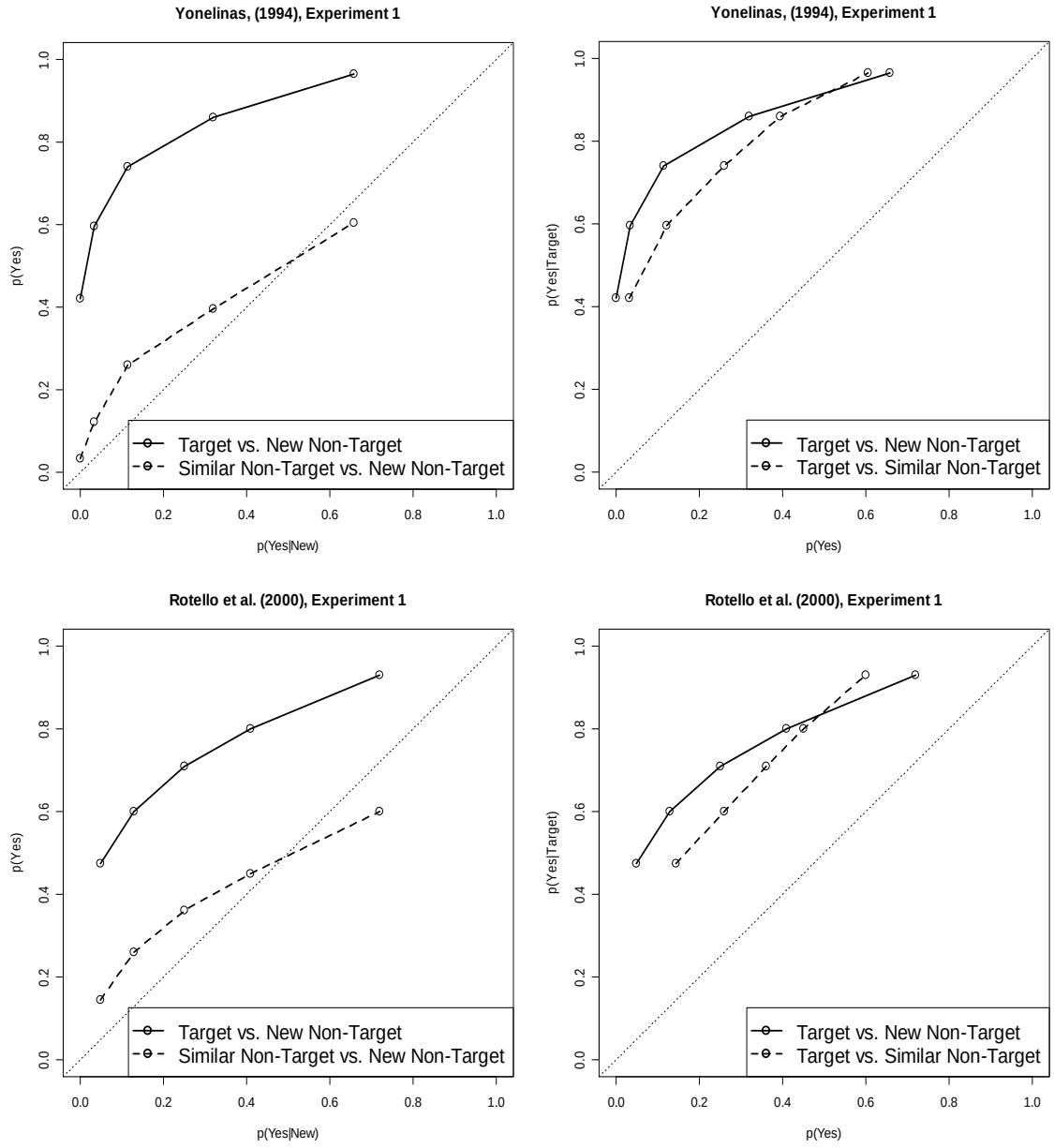
Yonelinas, (1994), Experiment 1
p(Yes)
p(Yes|New)
Target vs. New Non-Target
Similar Non-Target vs. New Non-Target
Yonelinas, (1994), Experiment 1
p(Yes|Target)
p(Yes)
Target vs. New Non-Target
Target vs. Similar Non-Target
Rotello et al. (2000), Experiment 1
p(Yes)
p(Yes|New)
Target vs. New Non-Target
Similar Non-Target vs. New Non-Target
Rotello et al. (2000), Experiment 1
p(Yes|Target)
p(Yes)
Target vs. New Non-Target
Target vs. Similar Non-Target

*Figure 1*

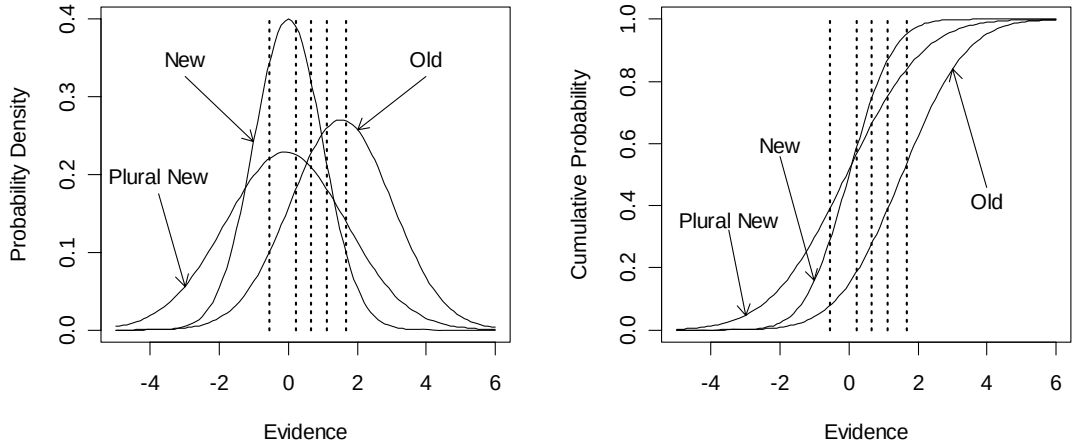
Probability Density
0.0
0.1
0.2
0.3
0.4
New
Old
Plural New
-4
-2
0
2
4
6
Evidence
Cumulative Probability
0.0
0.2
0.4
0.6
0.8
1.0
New
Plural New
Old
-4
-2
0
2
4
6
Evidence

*Figure 2*

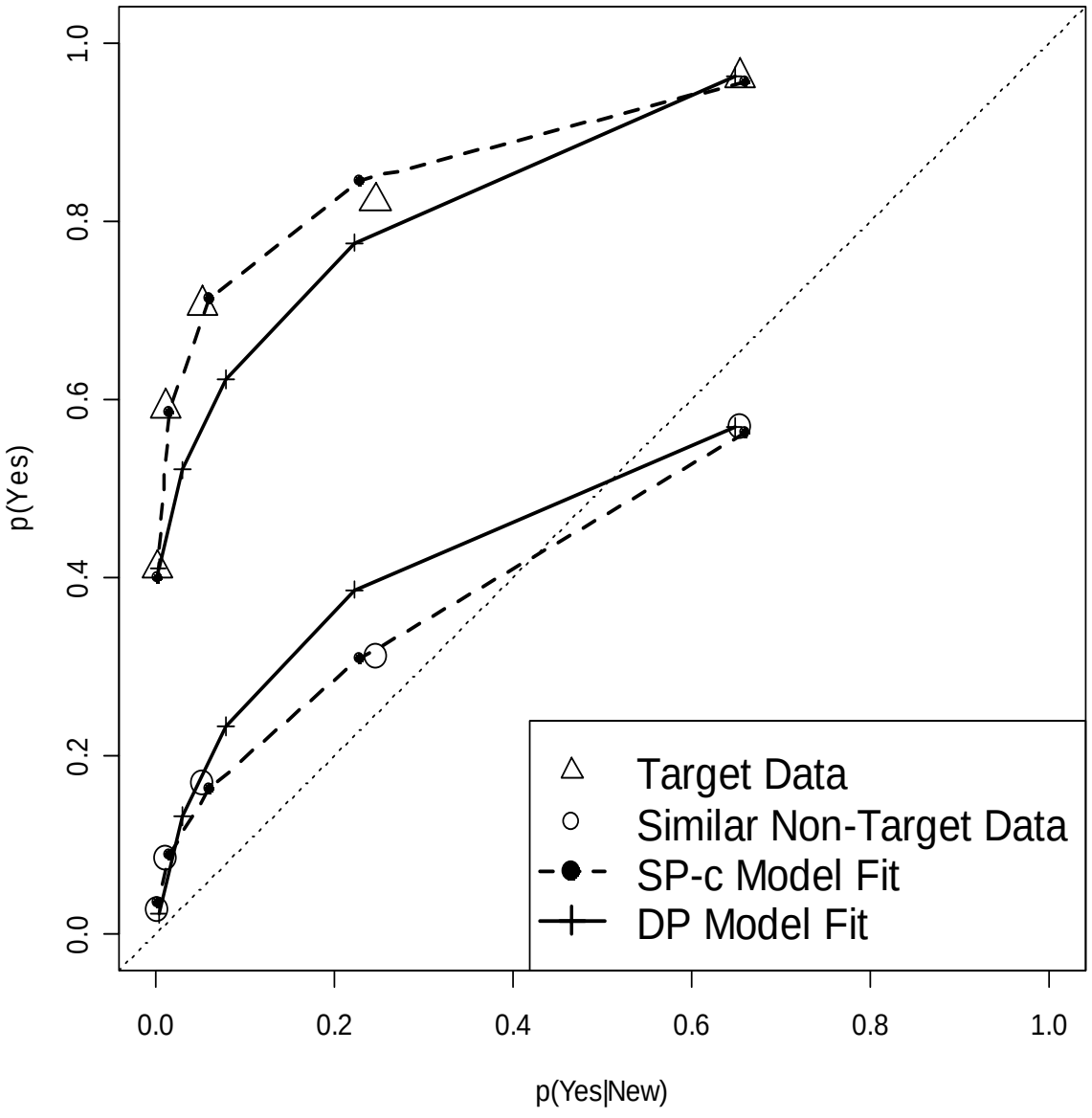
p(Yes)
0.0
0.2
0.4
0.6
0.8
1.0
p(Yes|New)
0.0
0.2
0.4
0.6
0.8
1.0
Target Data
Similar Non-Target Data
SP-c Model Fit
DP Model Fit

*Figure 3*

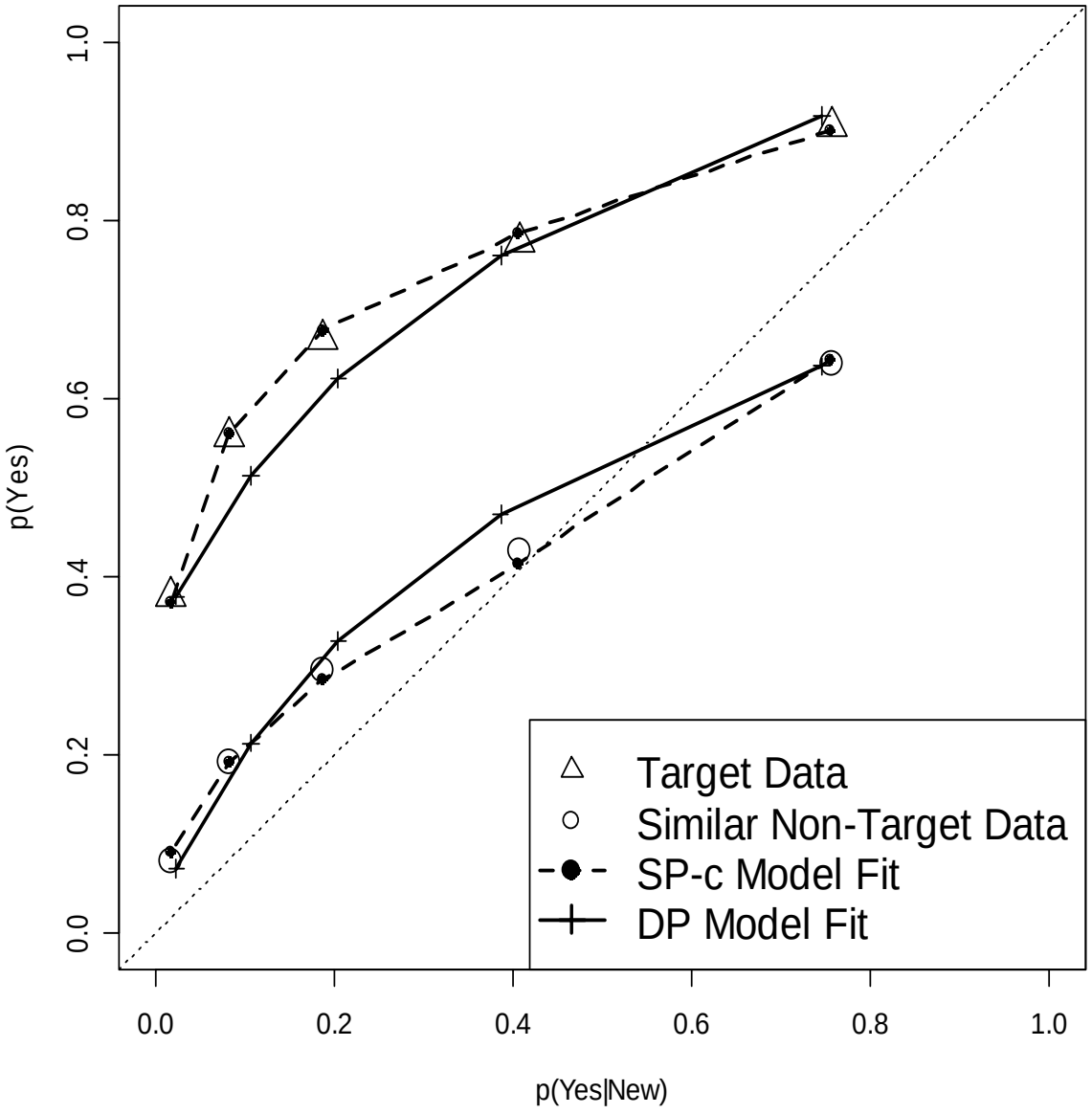
p(Yes)
0.0
0.2
0.4
0.6
0.8
1.0
p(Yes|New)
0.0
0.2
0.4
0.6
0.8
1.0
Target Data
Similar Non-Target Data
SP-c Model Fit
DP Model Fit

*Figure 4*

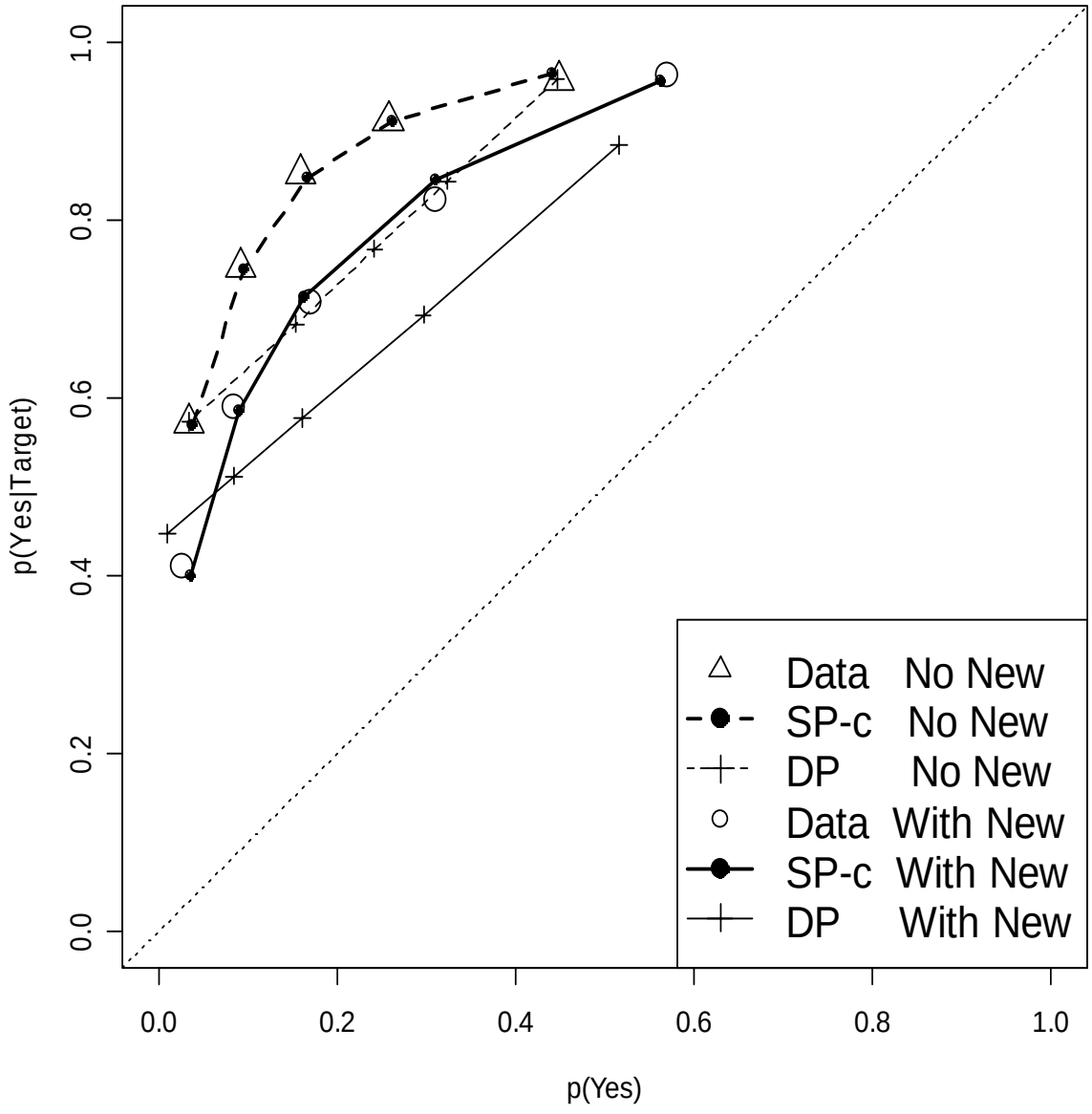
p(Yes|Target)
p(Yes)
0.0
0.2
0.4
0.6
0.8
1.0
Data No New
SP-c No New
DP No New
Data With New
SP-c With New
DP With New

*Figure 5*

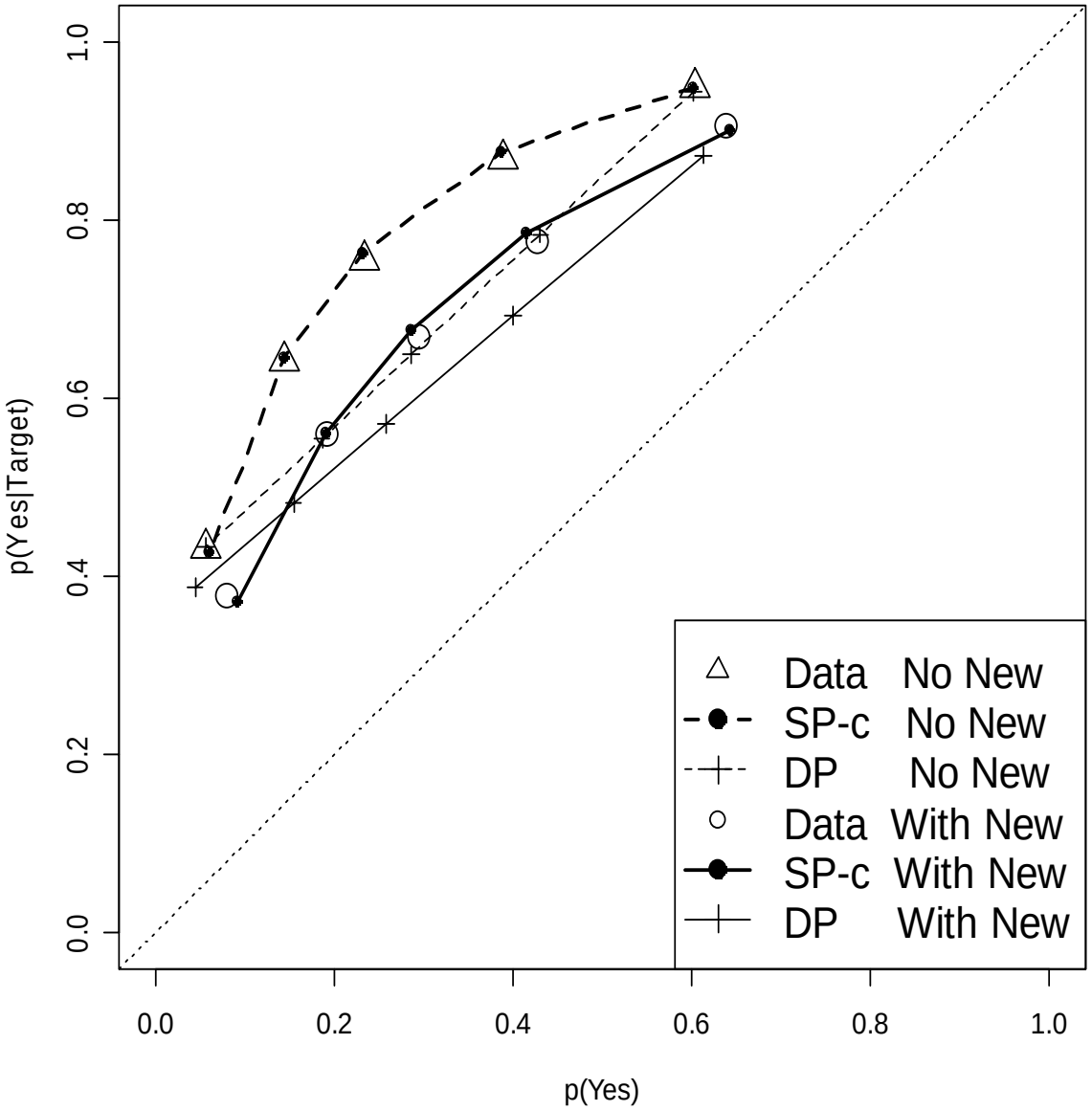
p(Yes|Target)
p(Yes)
0.0
0.2
0.4
0.6
0.8
1.0
Data No New
SP-c No New
DP No New
Data With New
SP-c With New
DP With New

*Figure 6*

# Footnotes

[1]Equal familiarity, on average, seems more likely in the list discrimination paradigm, where the first and second studied lists were equally often designated as the target in testing, than in the plural discrimination paradigm, where features related to the target's plurality would be more familiar. We compared estimates of familiarity between these two paradigms to examine this issue further

[2] A function is called concave if it has an inverted U-shape or, more strictly, if the y-value at the midpoint of the line segment connecting any two points on the function is less than the corresponding y-value on the function.

[3]Asymmetry is often assessed by examining the z-ROC curve, which plots the inverse cumulative normal (z) transformation of the probabilities constituting the original ROC curve. If the underlying distributions are normal, the z-ROC is a straight line with a slope equal to the standard deviation of the non-target distribution divided by the standard deviation of the target distribution. In item recognition experiments z-ROC slopes are usually found to be reliably less than one, at least when accuracy is better than chance, consistent with normally distributed evidence which is more variable for targets than non-targets.

[4]Although the cue-matching model is consistent with the non-target evidence distribution parameters estimated from Rotello et al's. (2000) participant average data, it is inconsistent with the standard deviation estimate for target (old) items (1.48), as it is less than the standard deviation estimate for similar non-targets (1.75). Given the effects of averaging this finding does not necessarily reject the cue-matching model, which predicts that these estimates should be close to equal. The same problem was not evident in fits to Yonelinas's (1994) participant average data, where similar non-targets have a slightly smaller standard deviation estimate (1.56) than targets (1.65). Note that in this study there were separate short and long list conditions, results from which were averaged in Figure 1 for clarity. Standard deviation estimates for the short condition were 1.7 and 1.46 and for the long condition 1.34 and 1.67 for similar non-targets and targets respectively.

[5] A convex curve is essentially the opposite of a concave curve. Where a concave curve has an inverted U-shape, a convex curve has a upright U-shape.

[6]We also examined a further extension that can be applied to all five models (see Appendix A for details), allowing responses to based on unbiased guessing on a proportion of trails. In contrast to the convex z-ROC curves produced by the addition of recollection to signal detection, the addition of guessing can produce

concave z-ROC curves (Ratcliff, McKoon & Tindall, 1994). Heathcote (2003) found greater evidence for concave than convex z-ROC curves in item recognition, although most curves were close to linear. Guessing was found to play only a minor role in the analyses reported here, and its inclusion did not change any of the conclusions based on models that omitted guessing, so we do not discuss it further.

[7]Criteria accounting for functional form have not yet been developed for the models considered here, although they could in principle be developed using the methods described by Pitt, Myung and Zhang (2002).

[8] Note that we define $n=k(m-1)$, as the only $m$-1 response categories contribute independent data.

[9] This is not the case when these ROCs are fit by linear regression, as used by Rotello et al. (2000). Because linear regression does not respect the bounded nature of probability, it can estimate a line of best fit through the middle of the curve, which gives an apparently better fit.