Positive and negative remember judgments and ROCs in the plurals paradigm: Evidence for alternative decision strategies

AYCAN KAPUCU, NEIL A. MACMILLAN, AND CAREN M. ROTELLO
University of Massachusetts, Amherst, Massachusetts

Using old–new ratings and remember–know judgments we explored the plurals paradigm, in which studied words must be distinguished from plurality-changed lures. The paradigm allowed us to investigate *negative remembering*—that is, the remembering of a plural-altered study item; capacity for this judgment was found to be poorer than or equivalent to the conventional positive remembering. A response-bias manipulation affected positive but not negative remembering. The ratings were used to construct ROC curves and test the prediction of the most common dual-process theory of recognition memory (Yonelinas, 2001) that the amount of recollection can be independently estimated from ROC curves and from remember judgments. By fitting the individual data with pure signal detection theory (SDT) models and dual-process models that combined SDT and high-threshold components (HTSDT), we identified two types of subjects. For those who were better described by HTSDT, the predicted convergence of remember–know and ROC measures was observed. For those who were better described by SDT, the ROC intercept could not predict the remember rate. The data are consistent with the idea that all subjects rely on the same representation but base their decisions on different partitions of a decision space.

Recognition memory is thought by many to be served by two processes: recollection, in which a test item evokes the studied event in context; and familiarity, in which the mere strength of the test item leads to the inference that it was studied. One class of models for recognition experiments, naturally referred to as “dual process,” incorporates these two types of analysis; specific models within this class differ in the details of assumed representation and response strategies (Jacoby, 1991; Wixted, 2007; Yonelinas, 1994). Dual-process models correspond to some common intuitions—for example, that an item (a face, say) can seem familiar even if it appears in a new context that undermines recollection. The most popular dual-process model (Yonelinas, 1994, 1997) assumes that recognition on a given test trial results from either a high-threshold recollection process (in which false alarms do not occur) or a continuous, signal-detection-based familiarity process. We term this model HTSDT.¹

An alternative class of models imagines that a single-strength variable underlies recognition judgments. To account for the dual-process intuition, some models of this type assume that the judgment dimension sums the outputs of continuous familiarity and retrieval processes on each trial (Wixted & Stretch, 2004). The most common single-process model, which we call simply SDT, assumes a Gaussian signal detection representation and response rule. This model can be viewed as a special case of HTSDT in which the high-threshold recollective process has zero sensitivity.

The threshold and SDT assumptions lead to different predictions about the ROC curves generated in a recognition experiment in which subjects provide confidence ratings rather than simple binary responses. Both models yield nonlinear, convex ROCs, but HTSDT predicts a nonzero *y*-intercept, whereas SDT does not. An appealing aspect of the HTSDT approach is that the *y*-intercept (on probability coordinates) provides a quantitative estimate of the proportion of trials on which recollection occurs.

This calculated recollection rate would be most compelling if it could be shown to correspond to values of recollection accuracy obtained by some other means, and the remember–know paradigm (Gardiner & Richardson-Klavehn, 2000; Tulving, 1985) has often been used to obtain such estimates. In this paradigm, subjects report for each item identified as “old” the subjective basis for their decision: Do they “remember” the item from the study trial—that is, have a specific memory of experiencing it—or do they just “know” that it was presented on the basis of its strength? Remember–know judgments are seductive because they promise direct access to the dual-process intuition. Agreement between recollection rates estimated from ROCs and from remember–know judgments would be valuable support for HTSDT.

Yonelinas (2001) measured ROC intercepts and remember rates in separate experiments, and found good agreement between them. Few researchers have examined both measures in the same experiment, but Rotello, Mac-

millan, Reeder, and Wong (2005) showed that remember rates could be strongly influenced by instructions that left the ROCs unchanged. Their most conservative instructions (argued by HTSDT theorists to be the most appropriate) showed the closest agreement between the remember rates and the ROC intercept; the other conditions showed substantial discrepancies. Clearly an experimental comparison of remember-know and ratings-based estimates of recollection must take biasing instructions into account.

The Plurals Paradigm

In most recognition experiments with words, the lures used at test are selected randomly from a large set. We chose instead to use lures that differed from targets only in plurality: If *frog* and *computers* were studied, then *frogs* and *computer* were to be treated as new rather than old items. The appropriate response to a test item depended on the presence or absence of the final "s." There are several reasons for using this paradigm to investigate the agreement between the ROCs and remember-know judgments.

First, no one has collected remember-know judgments in the plurals paradigm. When a "remember" judgment is offered in a conventional random-lure experiment, a wide range of context information may contribute to the response, but in the plurals paradigm it can only be that single piece of information that distinguishes singular and plural words. This reduction in variability might be expected to produce less noisy data, but in any case a comparison with otherwise similar standard experiments allows us to compare remember-know performance across different sets of lures.

Second, the plurals paradigm allows us to study what might be called "negative remembering." Typically, remember-know judgments are solicited only following "old" responses, but we asked for them following "new" judgments as well. A "remember" response to an item thought to be a lure is justified if the subject remembers that the same word with different plurality was studied. For example, if *frog* was studied and *frogs* is presented as a lure, a "new, remember" response indicates a strong memory for *frog*, implying that *frogs* must not have been studied. To our knowledge, soliciting negative remembering judgments in an item recognition experiment is relatively novel, having been used previously by Migo, Montaldi, Norman, Quamme, and Mayes (2009) and by Jones and Atchley (2006) in studying compound words. An obvious question about this design is whether positive and negative remember rates are of the same magnitude, as would be expected if subjects are equally able to use negative and positive information about plurality. Using confidence ratings and a multinomial (high-threshold) model, Rotello (2000) found that recall-to-accept processing was about twice as likely as recall-to-reject processing for plurality-changed stimuli, so we expected more positive than negative remembering under HTSDT assumptions.

A final motivation for using the plurals paradigm is that it has been seen to provide an ideal test bed for model comparison. Parks and Yonelinas (2007, p. 191) argued that "relational recognition tasks, including source, as-

sociative, and plurality-reversed recognition tests, provide a perfect arena for pitting these two models against one another because they call for the retrieval of some specific detail about a test item." One special aspect of the paradigm is that the ROCs obtained are unusual. Rotello, Macmillan, and Van Tassel (2000) plotted the hit rate for targets versus the false alarm rate for plural-altered lures and found linear functions; ROCs for targets versus new lures, like those from other experiments with random lures, were curvilinear and consistent with a Gaussian model (see also Rotello, 2000). Heathcote, Raymond, and Dunn (2006) reported curvilinear old-similar ROCs, a result they viewed as a failure to replicate Rotello et al. (2000), but Figure 1 shows that the discrepancy is just a matter of degree. Rotello et al.'s (2000) ROCs differ slightly but systematically from linearity; those of Heathcote et al. show more curvature, but not as much as is obtained in experiments with random lures. We call ROCs of this intermediate shape "flattened," to stress their relation to those found in standard recognition studies. To provide a theoretical account of the ROC data, we used a modeling approach, to be described later.

The Present Experiments

We conducted two experiments that differed in their instructions. Proponents of the HTSDT model (Yonelinas, 2001) have argued that "remember" responses can be expected to reflect recollection only if subjects are instructed to limit them to cases in which they remember specific details about the item's presentation. In view of Rotello et al.'s (2005) results, we used standard remember-know instructions in one experiment and strongly restrictive instructions in the other. We expected to replicate Rotello

Figure 1. Old-similar ROCs from Rotello, Macmillan, and Van Tassel (2000, Figures 3B and 5B) and Heathcote, Raymond, and Dunn (2006, Figure 6).

et al.'s (2005) finding that remember rates agreed more closely with the HTSDT recollection parameter in the latter case. Subjects studied lists of words, and at test had to distinguish those words from similar (plural-altered) words and from entirely new words. They used a rating scale to make this old–new judgment. On all trials, they also contributed a “remember–know” response; a “remember” following an “old” response meant they remembered studying the word and a “remember” following a “new” response that they remembered studying the alternate-plurality version of the word.

METHOD

Subjects

Fifty-six undergraduate students at the University of Massachusetts participated (27 in Experiment 1, 29 in Experiment 2) in exchange for a small payment or extra credit in their psychology courses. All were native English speakers.

Stimuli

Three hundred common nouns (mean frequency, 70.6 per million; Kučera & Francis, 1967) and their plurals (e.g., *frog–frogs*) were selected from the MRC Psycholinguistic Database (Coltheart, 1981). Nouns were 3–12 characters in length and were chosen only if their plural form could be created by adding an “s.”

Design

One word from each of 144 singular–plural pairs was randomly selected, independently for each subject, to form a set of 72 singular and 72 plural words. The selected words were evenly divided into three study lists of 48 critical words each. Four additional words from the word pool were used to fill primacy and recency buffers on each list. Therefore, each study list was composed of 52 words in total. Presentation order was randomized for each list and each subject.

For each study list, the test list was composed of three types of items: Twenty-four randomly chosen studied words served as targets (12 singular, 12 plural); another 24 words were presented in the altered plurality (we call these *similar lures*); and a final 24 words, selected from the remaining 144 singular–plural pairs, were completely unstudied and served as new lures. Test order was randomized for each subject.

In addition, a practice test list comprised six items: two old words, two similar lures, and two new lures. Two of the four primacy words were chosen to serve as old items, the unstudied altered form of the remaining two served as similar lures, and two unstudied words served as new items.

Procedure

Subjects studied three lists of 52 words each, at a presentation duration of 3 sec. They were told to pay attention to the plurality of each word in preparation for an unspecified memory test. Immediately following each list, subjects received a recognition test composed of words that had been studied previously (*targets*), plural/singular altered forms of the studied words (*similar lures*), and completely new words (*new lures*). For each test item, subjects gave old/new confidence ratings on a 6-point scale (1 = *sure new*, 2 = *probably new*, 3 = *maybe new*, 4 = *maybe old*, 5 = *probably old*, 6 = *sure old*). Following each old–new judgment, subjects were asked to make a binary remember–know judgment for the same test item. They were informed that words were presented in either singular or plural form, never both.

The standard remember–know instructions of Rajaram (1993) were modified to require a remember–know decision on all trials. If the test item had been judged new, a “remember” response indicated the recollection of specific information inconsistent with the test item. Subjects were given the example that if they could recall study-

ing *computers*—if they for instance remembered imagining a computer lab full of computers while studying the word *computers*—they could not have studied *computer* and should give *computer* a “new” response (at some confidence level) and a “remember” judgment. In Experiment 2 (but not Experiment 1), subjects were in addition told to give a “remember” judgment only if they could recall specific plurality information about the test item.

RESULTS

Before analyzing the data, we eliminated subjects whose accuracy on target and new-lure trials was below 60%. This restriction reduced the number of usable subjects from 27 to 20 in Experiment 1 and from 29 to 25 in Experiment 2. The ability of the remaining subjects to distinguish targets from lures (using just the old–new judgment, not the ratings) is summarized in the first three lines of Table 1. Results of the two experiments were quite similar, with an average hit rate of .66, false alarms to .35 of similar items, and false alarms to .21 of new lures. Little effect of the instructional manipulation (which referred directly to remember–know, not old–new, judgments) is in evidence.

The remember–know data are also summarized in Table 1. Positive remember rates are reliably higher than negative ones in Experiment 1 [.41 vs. .27; $t(19) = 2.44$, $p = .025$] and numerically higher in Experiment 2 (.33 vs. .29). Apparently subjects were able to report negative remembering, but at least some of them found the “new, remember” response hard to use or understand. The numerically lower positive remember rate in Experiment 2 indicates responsiveness to the more conservative instructions in that experiment; the lack of such an effect for negative remembers may be because using this response already required a rather conservative decision strategy. The overall remember rates are similar to those typically obtained with randomly chosen rather than plural-altered lures. In Dunn's (2004) database of 400 remember–know studies, the average remember hit rate was .42; in our experiments, the value for positive remembering is .37. We conclude that the presence of plural-altered lures in the test list did not affect the use of positive remember–know judgments by our subjects.

Table 1
Hits, False Alarms, and Remember Rates

	Experiment 1		Experiment 2		Average
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	
Probability of saying “old”					
Targets	.67	.026	.65	.023	.66
Similar lures	.37	.044	.34	.030	.35
New lures	.17	.021	.24	.029	.21
Positive remember rate					
Targets	.41	.047	.33	.042	.37
Similar lures	.12	.028	.09	.016	.10
New lures	.02	.007	.02	.008	.02
Negative remember rate					
Targets	.03	.010	.05	.010	.04
Similar lures	.27	.053	.29	.042	.28
New lures	.12	.027	.06	.017	.09

Figure 2 shows ROCs based on data pooled across subjects for both experiments. The form of the curves is consistent with past work: The old–new ROCs display curvature of the sort usually found with such random lures, whereas the old–similar ROCs are flattened. Although it has been common to argue from ROC shape to an underlying model, such inferences face a number of difficulties, and we instead resort to explicit models of the sort outlined earlier. We now describe this modeling effort.

Model-Fitting Analysis

All the models we evaluated are submodels of a general dual-process model that combines categorical recollection and SDT-based familiarity. Our experiments required recognition judgments for three stimulus classes—targets, lures, and plural-altered lures—rather than two. The signal

detection process therefore comprises three distributions corresponding to these classes. Recollection is assumed to occur only for targets and similar items. The most general model, which we call *HTSDT unequal variance*, has 11 parameters: R_O and R_N (the proportions of target and similar trials on which correct positive and negative recollection occurs), d_T and d_L (the means of the target and similar distributions), s_T and s_L (the standard deviations of the target and similar distributions), and five criterion locations. The new-lure distribution has a mean of 0 and a standard deviation of 1. There are 15 independent data points ($5 = 6 - 1$ ratings $\times 3$ stimulus types), so the HTSDT unequal-variance model has 4 degrees of freedom.

We also considered five models nested within HTSDT unequal variance, as shown in Table 2. The *HTSDT equal-variance* model sets s_T and s_L equal to 1 (the standard deviation of the new-lure distribution). The *HTSDT similar = lure* model is nested within the HTSDT equal-variance model, with the added restriction that the means of the similar and new-lure distributions are equal. The *HTSDT similar = target* model is also nested within the HTSDT equal-variance model, with the added restriction that the means of the similar and target distributions are equal. The final two models dispense with the recollection process, and are therefore single-process SDT. Either all three variances can be unequal (the *SDT three-variance* model) or s_T and s_L are set equal (*SDT two-variance*). Equations expressing data in terms of model parameters are provided in the Appendix.²

Quantitative models of memory experiments have almost always been evaluated against group data of the sort shown in Figure 2. The risks in this kind of averaging have long been understood in general and in the context of memory modeling (Cohen, Sanborn, & Shiffrin, 2008; Estes & Maddox, 2005). When two ROC curves are averaged, the shape of the resulting group function may not resemble that of either component. More generally, if Model A provides the best fit for 2 subjects, Model B may outperform it with the average data. Cohen, Rotello, and Macmillan (2008) observed that in remember–know experiments in which old–new judgments were made prior to remember–know decisions, a version of the dual-process model was often found to fit the group data best even though the individual subjects (or simulated subjects) were best described by an SDT model. Because one interpretation of our models is as distinct, optional decision strategies, we expected that not all subjects would be fit by the same model; in such a situation, analysis of data separately for each subject is essential.

Each of the submodels listed in Table 2 was fit to the data of all subjects whose accuracy on target and new-lure trials was at least 60%. To choose the best model of the ratings data, we used the AIC and BIC statistics (Akaike, 1973; Schwarz, 1978), which help to adjust for differing degrees of freedom when comparing nonnested models. Table 3 shows the breakdown of best-fitting models by subject. Across Experiments 1 and 2, AIC found an SDT model to be best in 22 cases and an HTSDT model to be best in 23 cases. According to BIC, these numbers were 26 and 19. The two partitions are very similar, but they

Figure 2. Pooled old–similar and old–new ROCs: (A) Experiment 1, (B) Experiment 2.

Table 2
Models Tested

Model	Parameters ^a		Restriction on Full HTSDT	#pars ^b	df ^c	Nested in
	Recollection	Familiarity				
1. HTSDT Unequal variance	R_O, R_N	d_T, d_L, s_T, s_L	—	11	4	—
2. HTSDT Equal variance	R_O, R_N	d_T, d_L	$s_T = s_L = 1$	9	6	#1
3. HTSDT Similar = lure means	R_O, R_N	d_T	$s_T = s_L = 1$ $d_L = 0$	8	7	#2
4. HTSDT Similar = target means	R_O, R_N	d_T, d_L	$s_T = s_L = 1$ $d_T = d_L$	8	7	#2
5. SDT Three-variance	—	d_T, d_L, s_T, s_L	$R_O = R_N = 0$	9	6	#1
6. SDT Two-variance	—	d_T, d_L, s_T	$R_O = R_N = 0$ $s_T = s_L$	8	7	#5

^aEquations relating parameters to response rates are given in the Appendix. ^bNumber of parameters, including five criterion locations. ^cThere are 15 independent data cells, so the *df* is obtained by subtracting the number of parameters from 15.

are not identical. Which provides the better picture of the distribution of models across subjects?

An issue that is not dealt with by the information criteria is model complexity. More complex models can account for a wider variety of empirical outcomes, even when they have the same number of parameters as a less complex model (Pitt, Myung, & Zhang, 2002). Because AIC and BIC only penalize the models for their parameters, not their complexity, they provide imperfect measures for model selection. Cohen, Rotello, and Macmillan (2008) studied an experimental paradigm that is closely related to the present one: Subjects made old–new ratings and remember–know judgments, although the latter were omitted on trials for which the test item was called “new.” They concluded that some versions of the HTSDT model are more complex than the SDT models and that conclusions based on AIC in that paradigm are biased in favor of the HTSDT model. An optimal criterion is one that takes account of model complexity as well as number of parameters, and for the models studied by Cohen, Rotello, and Macmillan, BIC is closer to optimal than AIC. Thus, to the extent that we can generalize from those models to the present set of similar but not identical models, we should put greater faith in the implications of BIC. To be conservative, however, we describe our results both ways. The parameter estimates themselves differ according to which criterion is used, but the overall pattern is very similar.

It is instructive to examine the ROC curves generated by individuals. Figure 3 shows old–similar and old–new ROCs for 4 subjects best described by HTSDT and 4 others best described by SDT (according to both AIC and BIC). These particular ROCs provide a by-eye confirmation of the modeling diagnosis (some of the HTSDT ROCs appear to be linear and to have nonzero *y*-intercepts), although for some other subjects the conclusion is not so intuitively clear. The individual curves in Figure 3 vary widely in shape, spread, and of course performance level. The best performer in either experiment, S18, in the fig-

ure, is the only subject whose accuracy exceeded 90%. A complete set of ROC data for all individual subjects is available from the authors.

The nesting structure of models (shown in Table 2) can be used to test statistical hypotheses about their parameters. Our focus on individual subjects limits the interest in such hypothesis testing, but Table 3 allows some interesting conclusions nonetheless. First, models that permitted all variances to be unequal (Models 1 and 5) received very little support according to both AIC and BIC. A reasonable conclusion is that variances are equal for at least the target and similar distributions. Comparison of the HTSDT models reveals that the HTSDT similar = target model (Model 4) is the most successful. This finding is compatible with an interpretation of these models in which the signal detection process reflects familiarity, which is the same for target and plural-altered items. Within the two

Table 3
Number of Subjects Best Fit by Models 1–6 in Experiments 1 and 2, According to AIC and BIC

Model	Experiment 1		Experiment 2	
	AIC	BIC	AIC	BIC
1. HTSDT Unequal variance	0	0	4	1
2. HTSDT Equal variance	2	0	2	0
3. HTSDT Similar = lure	1	1	1	1
4. HTSDT Similar = target	7	9	6	7
5. SDT Three-variance	2	0	2	0
6. SDT Two-variance	8	10	10	16
All HTSDT (1, 2, 3, & 4)	10	10	13	9
All SDT (5 & 6)	10	10	12	16

Note—For details about the models, see Table 2.

Figure 3. Old-similar ROCs for individual subjects in Experiments 1 and 2. Left column, HTSDT provides best fit; right column, SDT provides best fit (according to both AIC and BIC). Solid lines refer to old-similar data, dashed lines to old-new. Superimposed functions merely connect the points and do not reflect any model.

SDT models, it is also clearly better to force targets and similar items to have equal variance.

Determining the average parameter values for the two classes of subjects allows us to describe them in more detail. To obtain meaningful averages, we analyzed all subjects according to the best-fitting model for their group: the similar = target model for HTSDT subjects and the two-variance model for SDT subjects. As Table 3 shows, these two models were in fact best fitting for 31 of the 45 subjects by the AIC standard and 42 of the 45 according to BIC.

Average values of the model parameters for the two subgroups are given in Table 4. In Experiment 1, AIC and BIC always agree on whether a subject should be assigned to the HTSDT or SDT groups, but in Experiment 2 the classification depends on which information criterion is used in four cases. In that experiment, therefore, we consider the data separately from both points of view. As is evident

in Table 4, the outcomes are similar for both experiments and both information criteria.

HTSDT subjects. Adherents of HTSDT recollected the appropriate study item about 34% of the time, approximately equally for positive and negative recollection. The signal detection process was of modest accuracy for targets ($d_T = 0.65$ to 0.85). These values can be interpreted as distances from the mean of the new-lure distribution along a decision axis, so the signal detection process is somewhat successful in distinguishing targets and similar lures from new lures. Recall that, according to the model, it cannot distinguish targets from similar lures at all, as is to be expected for a process based on “familiarity.” The more conservative instructions in Experiment 2 numerically increased the negative but not the positive recollection rates. Table 4 reports the old-new criterion (c_3) and the most conservative one (c_5); the latter would be expected to be most responsive to the biasing manipu-

Table 4
Average Parameter Values of Appropriate Models

HTSDT Subjects											
Experiment & Criterion	<i>N</i>	d_T		R_O		R_N		c_3 (old vs. new)		c_5 ("sure old")	
		<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
1	10	0.85	0.07	.32	.05	.23	.06	0.98	0.15	2.60	0.58
2–AIC	13	0.65	0.09	.35	.05	.36	.05	0.66	0.11	2.58	0.61
2–BIC	9	0.72	0.10	.34	.07	.33	.08	0.68	0.15	3.04	0.84

SDT Subjects											
Experiment & Criterion	<i>N</i>	d_T		d_L		s_T		c_3 (old vs. new)		c_5 ("sure old")	
		<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
1	10	1.99	0.40	0.09	0.49	1.66	0.20	1.09	0.13	2.01	0.14
2–AIC	12	1.57	0.20	0.07	0.20	2.05	0.19	0.88	0.15	2.30	0.21
2–BIC	16	1.65	0.17	–0.19	0.20	2.19	0.17	0.83	0.12	2.24	0.18

Note—For subjects better fit by HTSDT, the parameters are for the HTSDT similar = target model. For those better fit by SDT, the values are for the SDT two-variance model. In Experiment 1, AIC and BIC agree on the classification of subjects; in Experiment 2 they differ, and the results are given for both breakdowns.

lation. Neither these nor the other criteria were reliably affected. The lack of change in recollection is consistent with the threshold nature of that construct in the model, and the failure of the criteria in the signal detection component to respond to this instruction supports the claim that “remember” responses are driven by the recollection component for HTSDT subjects.

SDT subjects. For subjects operating under SDT principles, the average data imply that targets are well-discriminated from both similar and new lures, which are poorly discriminated from each other. However, the individual data reveal a division among these subjects. For some (7 of 10 in Experiment 1, 7 of 12 in Experiment 2 [AIC], and 7 of 16 in Experiment 2 [BIC]), the similar lures differ from new lures in the same direction as targets, with d_L averaging 0.82 and 0.46 in the two experiments. For the others (3 in Experiment 1, 5 in Experiment 2 [AIC], and 9 in Experiment 2 [BIC]), the similar lures differ from the new lures in the *opposite* direction as targets, with d_L averaging -1.63 , -0.48 , and -0.69 , respectively. One might interpret the decision axis for the first subset as familiarity, with similar lures having values that are lower than targets but still positive, whereas for the second group the decision axis is better understood as strength of plural-appropriateness. We consider how both of these representations might occur in the Discussion.

The middle criterion (c_3 in Table 4) was lower under conservative instructions for SDT subjects, but because these instructions referred to the “remember” response, the key criterion is the highest one (c_5). The location of c_5 is, as expected, slightly higher in Experiment 2 (2.27 vs. 2.01), but the difference is not reliable.

Finally, the standard deviations of the target and similar-lures distributions (which are set equal) are greater than those of the new-lure distribution by factors of 1.66 and 2.12 in the two experiments. This finding is consistent with the asymmetric old–new ROCs almost always found in recognition memory experiments.

“Remember” Responses

How do the models interpret “remember” responses? For HTSDT subjects, the remember hit rate and the remember correct rejection rate are estimates of plurality-information recollection rates. They should therefore agree with the parameters obtained in HTSDT model fits.

The proportion of “remember” responses after an “old” judgment was computed for the highest confidence “old” response (rating 6) and the proportion of “remember” responses after a “new” judgment was computed for the highest confidence “new” response (rating 1).³ The remember rates were “corrected” by subtracting the appropriate remember false alarm rate, a procedure that allows the comparison of remember hit rates when false alarm rates differ. This adjustment was recommended by Yonelinas, Dobbins, Szymanski, Dhaliwal, and King (1996, p. 431), and is justified if the HTSDT model is correct, but only then (Rotello, Masson, & Verde, 2008).

A comparison between remember rates and estimated recollection rates for the HTSDT subjects is provided in Table 5. Both mean values and correlations are relevant, and positive and negative remember and recollection rates can be considered separately or together. No matter how the data are examined, there is good support for the HTSDT model among these (HTSDT) subjects. Averaging across experiments, the positive and negative remember rates are .28 and .25 and the recollection rates estimated by the model are very similar, .33 and .29. Thus, remember rates slightly but consistently underestimate recollection rates. The correlation between the two measures is at least .93 in both experiments. Subsets of the data show the same effects. A curious aspect of the data is that the more stringent remember instructions in Experiment 2 led to an *increased* rate of both positive and negative remembering for this subgroup.

What about the subjects who behave more in accordance with SDT? Clearly the remember rates for our SDT subjects cannot be predicted from R_O and R_N , which are

Table 5
Comparison of Recollection Rates in HTSDT Model
With Corrected Remember Rates for HTSDT Subjects

Experiment & Criterion	<i>N</i>	<i>R_O</i>		Positive Remember Rate		Correlation	<i>R_N</i>		Negative Remember Rate		Correlation
		<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>		<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	
1	10	.32	.05	.28	.05	.95	.23	.06	.21	.06	.99
2-AIC	13	.35	.05	.30	.06	.93	.36	.05	.32	.06	.94
2-BIC	9	.34	.07	.28	.09	.94	.33	.08	.28	.09	.94

zero for these individuals. (The covariance between two variables, one of which is a constant [in this case, zero] is zero.) We attempted a partial test by examining the fits of the full HTSDT model (Model 1) and correlating the recollection estimates from that model with the remember rates. The resulting correlation was .36 in both experiments (averaging across positive and negative remembering); if a few outliers were omitted, the values were .29 and .09 in Experiments 1 and 2. The relation between recollection and remembering for these subjects is clearly weak, certainly weaker than for HTSDT subjects; this result is, of course, to be expected if our division of subjects into groups was appropriate.

The uncorrected remember hit rates for both groups, displayed in the top half of Table 6, are comparable: .38 versus .37 for positive remembering and .28 versus .26 for negative remembering (averaging across the two experiments). For SDT (but not HTSDT) subjects, the instruction manipulation showed a trend toward working as intended for positive remembering: The nonreliable .16 change in the positive remember rate is comparable to the .20 difference reported by Rotello et al. (2005) for a similar manipulation. The trend in the same direction for negative remembering is much smaller (.05). For completeness,

the bottom half of Table 6 contains remember false alarm rates for both subject groups in both experiments.

The SDT models assert that the “remember” response simply indicates high strength (Donaldson, 1996). Unlike the HTSDT models that include recollection, they do not predict an equivalence between remember rates and any aspect of the rating data—for example, the remember criterion need not be the same as any of the ratings criteria (Wixted & Stretch, 2004). We cannot use the remember-know data to directly test this model without making specific assumptions (Dougal & Rotello, 2007; Kapucu, Rotello, Ready, & Seidl, 2008; Rotello & Macmillan, 2006; Rotello, Macmillan, Hicks, & Hautus, 2006).⁴

DISCUSSION

By systematic model fitting, we have identified two classes of subjects, one relying entirely on a continuous signal detection process and the other supplementing such a process with threshold-based recollection. This finding accounts for the difficulty people have encountered in distinguishing these classes of models and may help to explain the variety of ROC shapes obtained in this and related paradigms. For example, the support that is some-

Table 6
Positive and Negative Remember Hit
and False Alarm Rates (Uncorrected) for HTSDT and SDT Subjects

Subject Class	Type of Remembering	Experiment 2								
		Experiment 1			AIC Criterion			BIC Criterion		
		Proportion			Proportion			Proportion		
		<i>M</i>	<i>SE</i>	<i>N</i>	<i>M</i>	<i>SE</i>	<i>N</i>	<i>M</i>	<i>SE</i>	<i>N</i>
Remember Hit Rates										
HTSDT	Positive	.38	.05	10	.39	.06	13	.36	.09	9
	Negative	.24	.06		.36	.03		.30	.09	
SDT	Positive	.45	.08	10	.27	.05	12	.32	.04	16
	Negative	.29	.09		.20	.04		.28	.05	
All	Positive	.41	.05	20	.33	.04	25	.33	.04	25
	Negative	.27	.05		.29	.04		.29	.04	
Remember False Alarm Rates										
HTSDT	Positive	.09	.03	10	.09	.02	13	.08	.02	9
	Negative	.03	.01		.04	.01		.02	.01	
SDT	Positive	.13	.05	10	.10	.03	12	.10	.02	16
	Negative	.04	.01		.05	.01		.07	.01	
All	Positive	.11	.03	20	.09	.02	25	.09	.02	25
	Negative	.04	.01		.05	.01		.05	.01	

Note—Positive “remember” responses are for old items, negative for new items.

times found for mixture models interpreted in terms of attention might actually result from a mixture of different subject strategies. Similar issues arise in associative and source memory, so the result is of broad interest.

We now address three questions: (1) Are there really two subsets, or is model identification just a noisy business? (2) How might multiple models derive in different subjects from the same information? And (3) What implications do these results have for conventional analysis of data from this and related paradigms?

Subject Classification or Model Mimicry?

Identifying the model that best describes data is always an inferential process. In the past, models in the class we have considered have usually been compared by examining specific aspects of ROC curves (most often, type and degree of curvature). Model fitting, unlike reliance on such “signature predictions,” takes all of the data into account. The noisiness of data guarantees that conclusions will be tentative—that is, that the incorrect model may have been identified. We have taken several approaches to reassure ourselves that our model-fitting process has not seriously misled us.

Akaike weights and evidence ratios. First, we asked how convincing the statistical process is by which one model is declared the winner for an individual subject. The magnitude of the differences in AIC and BIC varies from substantial to rather small. One way to assess the solidity of our procedure is to use *Akaike weights* for AIC values and corresponding weights for BIC (Wagenmakers & Farrell,

2004). These weights compare the differences in information criterion value among the models. The weights can in turn be used to calculate *evidence ratios*, which are the odds favoring the best model over each of the others.

Table 7 shows both statistics for both experiments. The Akaike weights compare the best model for a particular subset of subjects with the other models for those same subjects. In Experiment 1, the two dominant models (SDT two-variance and HTSDT similar = target) have weights of .44, which can be interpreted as the likelihood that these models are correct for those subjects. In Experiment 2, the weights are higher (and again equal) for both sets of subjects, at .53. The evidence ratios comparing the dominant model with its closest competitor (usually the other dominant model) are between 1.5 and 2.7; these odds ratios correspond to probabilities of .66 to .73.

The Bayesian criterion values are stronger: Weights for the dominant models are .65 or .66 in Experiment 1 and .73 or .74 in Experiment 2. The evidence ratios for the winning models range from 3 to over 8, which correspond to probabilities of .76 to .89. According to BIC, therefore, we can be fairly confident that our testing procedure has chosen the correct model, whereas AIC comparisons are less decisive. Recall that BIC tends to favor simpler models (those with fewer parameters). The two dominant models have fewer parameters than some of their competitors (see Table 2), but because they each have eight parameters, neither has an advantage in computing evidence ratios.

Simulations. Another way to assess the confusability of these models was employed by Wixted (2007). He

Table 7
AIC and BIC Weights and Evidence Ratios for Both Experiments

Model	AIC			BIC		
	Number Subjects	Weight	Evidence Ratio	Number Subjects	Weight	Evidence Ratio
Experiment 1						
2. HTSDT						
Equal variance	2	0.46	1.51			
3. HTSDT						
Similar = lure	1	0.48	1.93	1	0.62	1.93
4. HTSDT						
Similar = target	7	0.44	1.82	9	0.66	3.15
5. SDT						
Three-variance	2	0.48	1.42			
6. SDT						
Two-variance	8	0.44	1.54	10	0.65	5.78
Total	20			20		
Experiment 2						
1. HTSDT						
Unequal variance	4	0.66	4.47	1	0.70	4.83
2. HTSDT						
Equal variance	2	0.40	1.22			
3. HTSDT						
Similar = lure	1	0.56	2.69	1	0.74	3.43
4. HTSDT						
Similar = target	6	0.53	2.12	7	0.77	8.08
5. SDT						
Three-variance	2	0.60	2.77			
6. SDT						
Two-variance	10	0.53	2.25	16	0.73	7.50
Total	25			25		

examined a (slightly different) implementation of the HTSDT and SDT models by simulating each and fitting the resulting data to both the model of origin and its competitor. The result was that in 20 of 30 cases the model that generated the data was better fit, and this proportion was the same for both models. According to this result, we may expect that one third of the subjects were misclassified, but there is no bias; that is, the proportion allocated to the two models is approximately correct. A similar result was obtained by Cohen, Rotello, and Macmillan (2008), who compared remember-know extensions of HTSDT and SDT models. The HTSDT model was found to be more flexible—that is, it could account for data generated by SDT assumptions more often than the reverse. To the degree our designs are similar, the present analyses may underestimate the number of SDT subjects.

Past attempts to classify subjects. The possibility that subjects differ in the models to which they adhere is supported by past recognition studies in our laboratory. In several experiments, we have fit HTSDT and SDT models to ratings or ratings-plus-remember-know data. Table 8, which displays the number of subjects whose data were best described by each of these models, tells a consistent story: Both models are supported by some subjects, with the SDT model gaining more adherents than HTSDT. In none of these past studies does the proportion of HTSDT subjects reach the levels observed in the present data; it may be that the plurals paradigm, in providing something concrete to recollect (namely, an “s”), is more amenable to an HTSDT strategy.

Two other paradigms, associative recognition and source recognition, bear both formal and empirical similarities to the plurals paradigm. Formally, in all cases discrimination depends on a single, well-identified piece of information, whereas in a typical recognition memory test with random lures, many features distinguish test words. Empirically, all three of these designs lead to flattened ROCs, and these ROCs have been interpreted as providing support for dual-process theories (Parks & Yonelinas, 2007), models in which inattention mixes with attention (DeCarlo, 2002),

or models with nonlinear decision bounds (Hautus, Macmillan, & Rotello, 2008).

To our knowledge, three articles have compared competing model fits for individual subjects in these designs. In their study of associative recognition, Kelley and Wixted (2001, Experiment 4) found that 14 out of 18 subjects were better fit by SDT than by a threshold model; the two models provided essentially equivalent fits for 6 other subjects. They attributed this division to the overflexibility of the threshold model. Healy, Light, and Chung (2005) fit versions of many models, including SDT, HTSDT, and Kelley and Wixted’s SON model, to individual subject data in associative recognition. They found SDT and SON to be favored over HTSDT and the others. Slotnick and Dodson (2005) fit versions of continuous SDT and high-threshold models to source memory data from individual subjects’ data and found that 11 of 15 subjects were better fit by an unequal-variance SDT model than by a high-threshold model. It appears that unanimity in the best-fitting model is not found when tested for, but that SDT models perform as well as or better than other models in these competitions.

A Single Representation for All, but a Variety of Decision Rules?

Does our conclusion—that subjects presented with the same experimental problem can be described by different models—make sense? Must we conclude that there are two kinds of people, with distinct memory representations? To the contrary, we believe the data are best interpreted to mean that different response strategies are available to apply to a common representation. We have previously proposed such a representation for conventional recognition experiments (Macmillan & Rotello, 2006), and extend it here to apply to the plurals paradigm.

Two kinds of information are available; following common usage, we term these *familiarity* and *recollection*. A two-dimensional space defined by these dimensions (both of which are assumed to be continuous) is shown in Figure 4. Bivariate distributions of information for targets, similar lures, and new lures are located in the space, rep-

Table 8
Number of Subjects in Previous Remember-Know Experiments
Best Described by HTSDT and SDT

Experimental Condition	Reference	HTSDT or Process-Pure	SDT
Emotional words	Dougal & Rotello (2007)	8	22
Emotional words—younger subjects	Kapucu et al. (2008)	3	19
Emotional words—older subjects	Kapucu et al. (2008)	7	16
Remember-first paradigm	Rotello & Macmillan (2006)	2	13
Trinary paradigm	Rotello & Macmillan (2006)	2	35
30% old items	Rotello et al. (2006)	3	19
70% old items	Rotello et al. (2006)	5	19
Conservative bias	Rotello et al. (2006)	2	17
Liberal bias	Rotello et al. (2006)	3	15
Total		35	175

Note—An additional 23 subjects were best fit by STREAK (Rotello et al., 2004). However, 18 of these fits were for the alternative remember-know designs studied by Rotello et al. (2006), and one of those (the trinary paradigm) has been shown to be particularly poor at distinguishing models (Cohen, Rotello, & Macmillan, 2008).

resented by circles. Targets naturally have higher values on both dimensions than either type of lure. We conjecture that the similar items are intermediate in familiarity between targets and lures, but differ from the lures on recollection in the opposite direction from targets. As is illustrated in Figure 4A, HTSDT subjects respond “old” for observations above a criterion on the recollection axis;⁵ for lower values of recollection, they rely on familiarity. The dashed lines in the figure are decision boundaries: Observations in the lower left corner of the decision space are labeled “new,” the rest “old.”

SDT subjects instead base their ratings judgments (and remember-know decisions) on a single axis. The most nearly optimal such dimension is a sum of the two

strengths, as proposed by Wixted and Stretch (2004). A decision bound that accomplishes this rule is shown in Figure 4B. Two additional rules are also illustrated: A subject relying on pure familiarity bases decisions on the horizontal axis (Figure 4C), whereas someone relying on pure (continuous) recollection bases decisions on the vertical axis (Figure 4D). The projected location of the similar items differs for these last two rules. For subjects relying on familiarity, similar items are intermediate in strength between targets and lures; for subjects relying on recollection, such items are more discriminable from lures than are targets. Recall that some of our SDT subjects were best described by each of these inferred arrangements.

Figure 4. An interpretation of the HTSDT and SDT models in a common decision space. The F - and R -axes quantify familiarity and recollection, and bivariate normal distributions for new and old items are indicated by constant-likelihood circles. Dashed lines are decision boundaries dividing “old” and “new” responses. (A) A “process-pure” model that resembles HTSDT; an “old” response is made if either dimension exceeds a criterion. (B) A one-dimensional model in which decisions are based on the sum of F and R . (C) A one-dimensional model in which decisions are based on F . (D) A one-dimensional model in which decisions are based on R .

The models we have tested can thus be viewed as strategies available to all subjects, the choice being influenced by experimental factors including instructions. Banks (2000) demonstrated exactly such an effect in the context of source memory. This interpretation implies that an individual might be encouraged to shift between strategies with the proper incentives or instructions, but there are no overall differences in accuracy between the subject groups, and it is not clear what kinds of incentives might accomplish such a shift.

Implications for Conventional Data Analysis

Most recognition experiments, including many of ours, base their conclusions on group data. What would we have concluded had we analyzed the present data in this conventional way? We evaluated the group data (displayed in Figure 2) by adding log likelihoods for all subjects for each model and then examining AIC and BIC as before (see Heathcote et al., 2006, for details of this procedure). In Experiment 1, AIC chose the HTSDT equal-variance model whereas BIC selected the SDT two-variance model. The positive and negative remember rates were .24 and .20, close to the R_O and R_N estimates (for the AIC best model) of .23 and .16. In Experiment 2, SDT two-variance was selected by both criteria, and of course the remember rates (.22 and .21) were not close to the model recollection values of 0. If the true situation is as we claim—different models applying to different subjects—then this analysis is quite meaningless.

The take-home message is that this paradigm cannot be used in any single simple way to extract different memory components, as has sometimes been done in the past. For example, Curran (2000) proposed that particular ERP components are identified with familiarity and recollection processes based on contrasts between different conditions of the plurals paradigm—that is, the contrast between targets and similar lures identifies recollection, whereas the contrast between similar lures and unrelated lures identifies familiarity. Our analyses suggest that at the very least this may be true for only a subset of subjects, whereas for another subset, the same contrasts imply different aspects of memory.⁶

The discrepancy between individual and group data analysis has been noted many times (e.g., Cohen, Sanborn, & Shiffrin, 2008; Estes & Maddox, 2005; Lee & Webb, 2005). The plurals paradigm may be particularly vulnerable to strategy differences. Subjects can adopt different methods for remembering plurality information, be more or less willing to engage in recall-to-reject reasoning, and have different interpretations of the “negative remembering” response. However, a variety of decision rules are available in many recognition designs—the remember-know literature, for instance, abounds in conflicting proposals about how information is used to generate responses. Whenever there are options, it is risky to assume that all participants will choose to participate in the same way.

AUTHOR NOTE

This research was supported by an NIMH grant to C.M.R. and N.A.M. (R01 MH60274) and is the basis of a master's thesis by A.K. A pre-

liminary version of this article was presented at the 2006 meeting of the Psychonomic Society in Houston, Texas. We thank John Dunn and two anonymous reviewers for their helpful comments. Correspondence should be addressed to N. A. Macmillan, Department of Psychology, University of Massachusetts, Amherst, MA 01003 (e-mail: nam@psych.umass.edu).

REFERENCES

- AKAIKE, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akadémiai Kiadó.
- BANKS, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science*, **11**, 267-273. doi:10.1111/1467-9280.00254
- COHEN, A. L., ROTELLO, C. M., & MACMILLAN, N. A. (2008). Evaluating models of remember-know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review*, **15**, 906-926. doi:10.3758/PBR.15.5.906
- COHEN, A. L., SANBORN, A. N., & SHIFFRIN, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, **15**, 692-712. doi:10.3758/PBR.15.4.692
- COLTHEART, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.
- CURRAN, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition*, **28**, 923-938.
- DECARLO, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, **109**, 710-721. doi:10.1037/0033-295X.109.4.710
- DONALDSON, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, **24**, 523-533.
- DOUGAL, S., & ROTELLO, C. M. (2007). “Remembering” emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, **14**, 423-429.
- DUNN, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, **111**, 524-542. doi:10.1037/0033-295X.111.2.524
- ESTES, W. K., & MADDOX, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, **12**, 403-408.
- GARDINER, J. M., & RICHARDSON-KLAVEHN, A. (2000). Remembering and knowing. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 229-244). Oxford: Oxford University Press.
- HAUTUS, M. J., MACMILLAN, N. A., & ROTELLO, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review*, **15**, 889-905. doi:10.3758/PBR.15.5.889
- HEALY, M. R., LIGHT, L. L., & CHUNG, C. (2005). Dual-process models of associative recognition in young and older adults: Evidence from receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 768-788. doi:10.1037/0278-7393.31.4.768
- HEATHCOTE, A., RAYMOND, F., & DUNN, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory & Language*, **55**, 495-514. doi:10.1016/j.jml.2006.07.001
- JACOBY, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, **30**, 513-541. doi:10.1016/0749-596X(91)90025-F
- JONES, T. C., & ATCHLEY, P. (2006). Conjunction errors, recollection-based rejections, and forgetting in a continuous recognition task. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **32**, 70-78. doi:10.1037/0278-7393.32.1.70
- KAPUCU, A., ROTELLO, C. M., READY, R. E., & SEIDL, K. N. (2008). Response bias in “remembering” emotional stimuli: A new perspective on age differences. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 703-711. doi:10.1037/0278-7393.34.3.703
- KELLEY, R., & WIXTED, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 701-722. doi:10.1037/0278-7393.27.3.701
- KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of*

- present-day American English. Providence, RI: Brown University Press.
- LEE, M. D., & WEBB, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, **12**, 605-621.
- MACMILLAN, N. A., & ROTELLO, C. M. (2006). Deciding about decision models of remember and know judgments: A reply to Murdock (2006). *Psychological Review*, **113**, 657-665. doi:10.1037/0033-295X.113.3.657
- MICKES, L., WAIS, P. E., & WIXTED, J. T. (2009). Recollection is a continuous process: Implications for dual-process theories of recognition memory. *Psychological Science*, **20**, 509-515. doi:10.1111/j.1467-9280.2009.02324.x
- MIGO, E., MONTALDI, D., NORMAN, K. A., QUAMME, J., & MAYES, A. (2009). The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology*, **62**, 1198-1215. doi:10.1080/17470210802391599
- MURDOCK, B. (2006). Decision-making models of remember-know judgments: Comment on Rotello, Macmillan, and Reeder (2004). *Psychological Review*, **113**, 648-656. doi:10.1037/0033-295X.113.3.648
- PARKS, C. M., & YONELINAS, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted (2007). *Psychological Review*, **114**, 188-202. doi:10.1037/0033-295X.114.1.188
- PITT, M. A., MYUNG, I. J., & ZHANG, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491. doi:10.1037/0033-295X.109.3.472
- RAJARAM, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition*, **21**, 89-102.
- ROTELLO, C. M. (2000). Recall processes in recognition memory. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 40, pp. 183-221). San Diego: Academic Press.
- ROTELLO, C. M., & MACMILLAN, N. A. (2006). Remember-know models as decision strategies in two experimental paradigms. *Journal of Memory & Language*, **55**, 479-494. doi:10.1016/j.jml.2006.08.002
- ROTELLO, C. M., MACMILLAN, N. A., HICKS, J. L., & HAUTUS, M. J. (2006). Interpreting the effects of response bias on remember-know judgments using signal detection and threshold models. *Memory & Cognition*, **34**, 1598-1614.
- 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. doi:10.1037/0033-295X.111.3.588
- 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 from ROC curves. *Journal of Memory & Language*, **43**, 67-88. doi:10.1006/jmla.1999.2701
- ROTELLO, C. M., MASSON, M. E. J., & VERDE, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, **70**, 389-401. doi:10.3758/PP.70.2.389
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.
- SLOTNICK, S. D., & DODSON, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition*, **33**, 151-170.
- TULVING, E. (1985). Memory and consciousness. *Canadian Psychology*, **26**, 1-12. doi:10.1037/h0080017
- WAGENMAKERS, E.-J., & FARRELL, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, **11**, 192-196.
- WIXTED, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, **114**, 152-176. doi:10.1037/0033-295X.114.1.152
- WIXTED, J. T., & STRETCH, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, **11**, 616-641.
- YONELINAS, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1341-1354. doi:10.1037/0278-7393.20.6.1341
- YONELINAS, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, **25**, 747-763.
- YONELINAS, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General*, **130**, 361-379. doi:10.1037/0096-3445.130.3.361
- YONELINAS, A. P., DOBBINS, I., SZYMANSKI, M. D., DHALIWAL, H. S., & KING, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness & Cognition*, **5**, 418-441. doi:10.1006/ccog.1996.0026

NOTES

1. Yonelinas and colleagues generally refer to this model as "dual process," and when it was first formulated that term served to distinguish it from other models. Since then, a number of other models postulating two processes have been developed (Kelley & Wixted, 2001; Murdock, 2006; Rotello, Macmillan, & Reeder, 2004), and we use the term "HTSDT" to be explicit about the two processes in question.
2. Another class of models assumes that performance is a mixture of attended and less attended trials (DeCarlo, 2002). We developed a model of this type for our design, but it provided a poor fit for almost all subjects relative to the models in the HTSDT class listed in Table 2.
3. Alternatively, one might include "remember" responses at all correct ratings—that is, 4, 5, and 6 for old items and 1, 2, and 3 for new items. We adopted the method preferred by HTSDT proponents. A few subjects who did not use the extreme ratings were dropped from the analysis.
4. In some interpretations, SDT predicts a correlation between remember hits and remember false alarms and HTSDT does not. In Experiment 1, these correlations are .09 (for similar false alarms) and -.12 (for new false alarms); in Experiment 2, the values are .34 and .28. This is not a revealing pattern.
5. The figure implies that recollection is continuous, as has been argued recently (Mickes, Wais, & Wixted, 2009), whereas HTSDT posits that it is a threshold process. The figure is most consistent with the process-pure model of Murdock (2006), but does illustrate the possibility of partitioning the decision space in different ways. The linear decision bounds imply SDT-like ROCs and would therefore require modification in an actual implementation of the model. Likelihood-ratio decision bounds, as employed in Hautus et al. (2008), can produce appropriately flattened ROCs.
6. We thank John Dunn for this point.

APPENDIX

Equations for the Models

The parent model is a version of HTSDT in which the variances are unconstrained. In all conditions, recollection leads to a “sure–yes” response for targets with probability R_O and a “sure–no” response for similar lures with probability R_N . When recollection fails, a signal detection process occurs with the following parameters: Distribution means 0 for new lures, d_L for similar lures, and d_T for the targets; corresponding standard deviations 1, s_L , and s_T ; and criteria c_1 , c_2 , c_3 , c_4 , and c_5 , in increasing order. Submodels are created by fixing the values of parameters (see Table 2). When recollection does not occur ($R_O = R_N = 0$), a pure signal detection process results.

Targets

$$P(6) = R_O + (1 - R_O) \left[1 - \Phi \left(\frac{c_5 - d_T}{s_T} \right) \right]$$

$$P(i) = (1 - R_O) \left[\Phi \left(\frac{c_i - d_T}{s_T} \right) - \Phi \left(\frac{c_{i-1} - d_T}{s_T} \right) \right], i = 2 \dots 5$$

$$P(1) = (1 - R_O) \Phi \left(\frac{c_1 - d_T}{s_T} \right)$$

Lures (Plurality Changed)

$$P(6) = (1 - R_N) \left[1 - \Phi \left(\frac{c_5 - d_L}{s_L} \right) \right]$$

$$P(i) = (1 - R_N) \left[\Phi \left(\frac{c_i - d_L}{s_L} \right) - \Phi \left(\frac{c_{i-1} - d_L}{s_L} \right) \right], i = 2 \dots 5$$

$$P(1) = R_N + (1 - R_N) \Phi \left(\frac{c_1 - d_L}{s_L} \right)$$

Lures (New)

$$P(6) = (1 - R_N) [1 - \Phi(c_5)]$$

$$P(i) = (1 - R_N) [\Phi(c_i) - \Phi(c_{i-1})], i = 2 \dots 5$$

$$P(1) = R_N + (1 - R_N) \Phi(c_1)$$

(Manuscript received May 27, 2009;
revision accepted for publication December 24, 2009.)