Journal of Experimental Psychology:
Learning, Memory, and Cognition
2002, Vol. 28, No. 6, 1083–1094

0278-7393/02/$5.00 DOI: 10.1037//0278-7393.28.6.1083

Differential Effects of List Strength on Recollection and Familiarity

Kenneth A. Norman
University of Colorado at Boulder

Numerous studies have found a null list strength effect (LSE) for recognition sensitivity: Strengthening memory traces associated with some studied items does not impair recognition of nonstrengthened studied items. In Experiment 1, the author found a LSE using receiver operating characteristic-based measures of recognition sensitivity. To account for the discrepancy between this and prior research, the author (a) argues that a LSE occurs for recollection but not for discrimination based on familiarity, and (b) presents self-report data consistent with this hypothesis. Experiment 2 tested the dual-process hypothesis more directly, using switched-plurality (SP) lures to isolate the contribution of recollection. There was a significant LSE for comparisons involving SP lures; the LSE for discrimination of studied items and nominally unrelated lures (which can be supported by familiarity) was not significant.

One of the fundamental goals of memory research is to characterize how memory traces interfere with one another. Traditionally, this question has been addressed by lengthening the study list (i.e., adding items that do not match other studied items) to see how this affects memory. Increasing list length impairs performance on tests of recognition, free recall, and cued recall (e.g., Gillund & Shiffrin, 1984; but see Dennis & Humphreys, 2001, for a discussion of confounds that are frequently present in list length experiments). A related question is how increasing *strength*—strengthening the memory traces associated with some, but not all, list items—affects memory for nonstrengthened list items. Tulving and Hastie (1972) were the first to ask this question; they found that strengthening some items (by increasing the presentation frequency of those items) impaired free recall of nonstrengthened items.

After the Tulving and Hastie (1972) study, the list strength issue lay dormant until it was revisited by Ratcliff, Clark, and Shiffrin (1990). Ratcliff et al. (1990) found that increasing list strength (by increasing the presentation frequency or presentation duration of some items) impaired free recall and cued recall of nonstrengthened items but that list strength manipulations had no effect on recognition of nonstrengthened items (i.e., participants' ability to discriminate between nonstrengthened studied items and lures was unimpaired). This null list strength effect (LSE) for recognition sensitivity has since been replicated by several other researchers (Hirshman, 1995; Murnane & Shiffrin, 1991a, 1991b; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Shiffrin, Huber, & Marinelli, 1995; Yonelinas, Hockley, & Murdock, 1992). The LSE for cued recall (using pairs of unrelated words as stimuli) has since been replicated by Kahana and Rizzuto (2002).

In preparing this article, I set out to assess the generality of the null LSE for recognition sensitivity. Although this finding has been replicated several times, there are reasons to believe that—under proper circumstances—it should be possible to obtain a LSE for recognition sensitivity. In particular, the LSE for cued recall indicates that recollection of details from the study phase (i.e., which words were paired together) can be impaired by list strength. According to dual-process theories of recognition, this kind of recollection contributes to recognition performance (along with nonspecific feelings of familiarity; see, e.g., Hintzman & Curran, 1994; Jacoby, Yonelinas, & Jennings, 1997; Mandler, 1980; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). Putting these two claims together brings one to the following: If recognition sensitivity is driven, in part, by recollection and if recollection is impaired by list strength, then it should be possible to observe a LSE for recognition sensitivity when recollection is contributing to recognition. (See Norman & O'Reilly, in press, for a recently developed computational model of recognition memory that predicts a LSE for recollection but not familiarity.)

The paradigm I use in this article was designed to ensure that recollection would contribute to recognition. It also incorporates several other design features (described in the Shared Design Elements section) aimed at maximizing the odds of detecting a LSE. In Experiment 1, I show that it is possible to obtain a LSE for recognition sensitivity and present some suggestive evidence from self-report data that the LSE observed is attributable to recollection. In Experiment 2, I present stronger evidence that list strength impairs recollection by using a paradigm in which participants have to discriminate between studied items and related switched-plurality (SP) lures (e.g., Hintzman, Curran, & Oppy, 1992).

Shared Design Elements

Basic Paradigm

Practically all list strength studies have used the "mixed-pure" paradigm pioneered by Ratcliff et al. (1990). In this paradigm, there are three types of study lists: mixed lists, consisting of both

Kenneth A. Norman, Department of Psychology, University of Colorado at Boulder.

This research was supported by National Institutes of Health National Research Service Award Grant MH12582-01. I thank Tim Curran, Lew Harvey, Doug Hintzman, David Huber, and two anonymous reviewers for commenting on previous versions of this article, and I thank Mike Kahana and Daniel Schacter for their advice during early stages of this research. I am also grateful to My Nguyen and Mani Nadjmi for collecting data from participants in Experiments 1 and 2, respectively.

Correspondence concerning this article should be addressed to Kenneth A. Norman, who is now at the Department of Psychology, Princeton University, Princeton, New Jersey 08544. E-mail: knorman@princeton.edu

strengthened (*strong*) items and nonstrengthened (*weak*) items; pure weak lists; and pure strong lists. Strengthening is achieved either by using a longer study duration or additional study presentations. A typical experiment consists of multiple study–test blocks, alternating between mixed, pure weak, and pure strong lists. If there is a LSE for recognition, participants should be worse at recognizing nonstrengthened (weak) items in mixed lists than in pure weak lists. Likewise, they should be worse at recognizing strengthened (strong) items in pure strong lists than in mixed lists.

One major limitation of the mixed-pure paradigm is that researchers are not free to repeat strong items as many times as they see fit. Because strong items are tested, memory for those items has to be kept below ceiling; otherwise, it would be possible to explain away null LSEs for strong items in terms of ceiling effects. To get around this limitation, I did not use the full mixed-pure design in the experiments reported here. Rather, I used a simplified design with only two kinds of lists, weak interference (WI) and strong interference (SI). Both types of lists were comprised of target (to-be-tested) items, and nontested interference items. Target items were presented once in both conditions; list strength was manipulated by presenting interference items once in the WI condition versus multiple times in the SI condition. The effect of list strength could be measured by comparing memory for targets in the WI versus the SI conditions. A key facet of this design is that, because interference items were not tested, these items could be overlearned in the SI condition without any adverse consequences. Taking advantage of this fact, interference items were presented six times on SI lists in the experiments reported here.

In all of the experiments reported here, participants studied (and were tested on) one WI list and at least one SI list. Figure 1 shows the general structure of the SI and WI blocks. In both kinds of blocks, participants first studied target and interference items once (mixed together). For SI lists, participants then studied the list of interference items five more times. The interference items were presented in a different order each time the list of interference items was repeated.

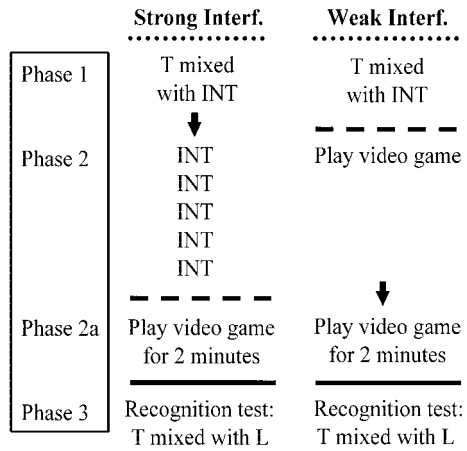

Figure 1. Diagram of the experimental procedure. Phase transitions marked with arrows were invisible to participants. Dotted horizontal bars indicate the beginning of the study phase. Dashed horizontal bars indicate the beginning of the video game phase. Solid horizontal bars indicate the beginning of the test phase. In all of the experiments, some extra words were presented at the beginning and end of Phase 1, serving as primacy and recency buffers, respectively. Interf. = interference; INT = interference items; L = lures; T = targets.

Participants played a video game immediately after each study list; WI study lists were followed by a long video game phase, and SI study lists were followed by a short video game phase, such that the average time between studying a target and being tested on that item was the same in the WI and SI conditions. The SI video game phase lasted 2 min and the WI video game phase lasted longer (the exact length was a function of the number of items and the study duration for that particular experiment). After the video game phase, participants were given a recognition test consisting of studied target items and nonstudied lure items.

Controlling Encoding

Floor effects on recollection can sabotage the LSE—if memory traces are too impoverished to support recollection in the WI condition, then it is not possible to observe a decrease in recollection in the SI condition. At the other end of the scale, if memory traces are too distinctive (such that overlap between traces is minimal), then interference effects cannot be obtained. To avoid these problems, one must select encoding parameters that force participants to do some elaboration—thereby ensuring that recollection is above floor—but also prevent participants from carrying out too much elaboration (insofar as this leads to overly distinctive memory traces).

In these experiments I used a size judgment encoding task. Words (concrete nouns) appeared on the screen and participants had to judge whether the thing denoted by that word would fit in a box with prespecified dimensions. Participants were given just over 1 s to make their size judgment. I selected this encoding duration because it gives participants just enough time to form a mental image of the item and decide whether it fits in the box, but—crucially—participants do not have enough time to generate idiosyncratic, distinctive elaborations (such as an image of a bicycle that is broken into pieces so that it fits in the box). Another useful aspect of the size judgment encoding task is that it induces some level of overlap between all memory traces formed at study, because participants think about all words with respect to the same reference box. Injecting a salient, shared element into all of the memory traces should bolster the extent to which memory traces interfere with one another.

Eliminating Rehearsal Confounds

As discussed by Ratcliff et al. (1990), participants' rehearsal of weak items at the expense of strong items in the SI (mixed list) condition can mask a LSE by artificially boosting memory for weak items. Furthermore, participants' rehearsal of strong items at the expense of weak items in the SI condition can result in a spurious LSE by artificially reducing memory for weak items. (This occurred in the study by Yonelinas et al., 1992.) The paradigm described above was designed to minimize rehearsal confounds. Stimuli were presented very briefly, and performing the encoding task took up almost the entire presentation interval; thus, participants had very little time left over for rehearsing previously presented stimuli. Furthermore, the study list was structured such that all of the target items were presented before any of the interference items were repeated. As such, there was no way to tell

which items were weak and which were strong during the part of the list where targets were presented; therefore, there was no way for participants to redistribute rehearsal according to strength.

Dependent Measures

I collected confidence rating data at test, using a 6-point confidence scale where the numbers from 1 to 6 were labeled (in order) *definitely new*, *probably new*, *guess new*, *guess old*, *probably old*, and *definitely old*. I then generated receiver operating characteristic (ROC) curves for individual participants by computing hits and false alarms on the basis of different confidence criteria (Macmillan & Creelman, 1991). To index sensitivity for each participant and for each interference strength condition, I used the RSCORE PLUS algorithm (Harvey, 2001) to fit a Gaussian model to that participant's or condition's confidence rating data[1]; I then used the parameters of the best-fitting Gaussian model to compute A_z, an estimate of the area under the ROC curve (Macmillan & Creelman, 1991).

Note that I am using a Gaussian model (which posits that recognition decisions are based on a single signal that is distributed normally for both studied items and lures) to estimate sensitivity for purely pragmatic reasons: (a) Gaussian models tend to provide a good fit to recognition memory ROC curves and (b) so long as a model provides a good fit, estimates of the area under the ROC derived using this model should be valid. For all of the experiments reported here, I provide evidence that, overall, Gaussian models provide a good fit to the data, which in turn validates the use of A_z to index sensitivity. For comparison purposes, I also used ROC data to compute the sensitivity measure d_a (Macmillan & Creelman, 1991). This measure assumes an underlying Gaussian model and provides an estimate of the distance between the studied-item and lure-item *memory signal* distributions. None of the conclusions I present here depend on the use of A_z as opposed to d_a; the LSE was significant for d_a if and only if it was significant for A_z.[2]

Next, as a preliminary means of getting at the idea that list strength affects recollection but not familiarity, I collected self-report measures of recollection in Experiment 1. Whenever a participant thought that an item was *old* (i.e., assigned the item a confidence rating > 3 on the 6-point scale), the participant was asked whether he or she remembered studying the item (i.e., recollected specific details) or whether the item just seemed familiar but no specific details came to mind (Gardiner, 1988; Rajaram, 1993; Tulving, 1985). The logic in collecting this data is quite straightforward: If recollection is contributing to recognition performance, "remember" responses should isolate this contribution more so than "old" responses.[3] Thus, under the assumption that list strength affects recollection, one would expect indices of discrimination (e.g., d') computed using remember responses to show a LSE, whereas indices of discrimination computed using old (confidence > 3) responses may not show a clear LSE (insofar as both familiarity and recollection can drive old responses). To test this prediction, I computed d'[4] on the basis of remember and old responses in Experiment 1; I refer to these measures as *d′(Remember)* and *d′(Old)*, respectively.[5] For both experiments, alpha was set to .05, two-tailed.

Experiment 1

Method

Participants. Thirty-six University of Colorado undergraduates (14 women and 22 men, mean age = 19.3 years) participated in the experiment. The experiment lasted approximately 1 hr, and participants received course credit.

Materials. Stimuli were 300 highly imageable, concrete, familiar medium-frequency nouns; imageability, concreteness, familiarity, and Kučera–Francis (K-F) frequency data were obtained from the MRC Psycholinguistic Database (Coltheart, 1981): mean imageability = 5.76 out of 7, range = 5.02–6.59; mean concreteness = 5.83 out of 7, range = 5.00–6.48; mean familiarity = 5.02, range = 4.00–6.16; mean K-F frequency = 15.8 occurrences per million, range = 0–99; mean word length = 5.54, range = 3–10. Because the purpose of this experiment was to examine LSEs using lures that were nominally unrelated to studied items, I took steps to ensure that none of the words were strongly (semantically) related to one another. Pairwise semantic relatedness assessments of 847 concrete nouns were generated using Latent Semantic Analysis (LSA; applied to the GenCOL corpus, which is meant to reflect what a person has read up to the first year of college; semantic representations were constrained to use 300 feature dimensions; Landauer, Foltz, & Laham, 1998); of these 847 words, 300 words were selected such that the maximum pairwise LSA cosine for the 300 words (larger cosines reflect higher semantic relatedness) was .42. A small number of near-synonymous words not caught by the LSA screening were removed by hand (e.g., *coffin* and *casket*). Also, I excluded some compound words because their constituent words were also included in the stimulus set, and I made an attempt to exclude ambiguous words (e.g., *ram*). In addition to the 300 words described above, 20 other words were used as primacy and recency buffers at study.

The 300 words not used as buffers were split into two 150-word groups, and each group was divided into three 50-word subgroups. Also, the following steps were taken to ensure that the six 50-word subgroups were matched, on average, for important word characteristics: (a) I matched

[1] RSCORE PLUS builds on the RSCORE maximum likelihood parameter estimation algorithm developed by Dorfman and Alf (1969; see also Dorfman, Beavers, & Saslow, 1973) by incorporating more robust nonlinear fitting techniques and other mathematical advances. The RSCORE PLUS software can be downloaded from http://psych.colorado.edu/~lharvey (from the main page, follow the link to "Software").

[2] For each participant and condition, we also used ROC data to compute the sensitivity measure A_g, which (like A_z) is an estimate of the area under the ROC. A_g (unlike A_z) does not assume a Gaussian model; instead, it estimates area by connecting the dots that comprise the ROC (Macmillan & Creelman, 1991). The results I obtained using A_g were qualitatively identical to the results I obtained using d_a and A_z; to conserve space, these results are not presented here.

[3] The claim that remember responses specifically index recollection, as opposed to confidence, is quite controversial. (For opposing perspectives on this debate, see, e.g., Donaldson, 1996; Hirshman & Master, 1997, vs. Gardiner & Gregg, 1997.)

[4] Because d' is undefined with extreme values (i.e., hits or false alarms = 0 or 1), I substituted $.5/N$ (where N = the number of items per condition) when hits or false alarms = 0, and I substituted $1 - (.5/N)$ when hits or false alarms were equal to 1 (Green & Swets, 1974; Macmillan & Creelman, 1991).

[5] I also computed sensitivity on the basis of remember and old responses using a different measure (A'; Donaldson, 1992; Macmillan & Creelman, 1991), and the results were qualitatively identical to the results that were obtained using d'; to conserve space, A' results are not reported here.

subgroups for K-F frequency, familiarity, concreteness, imageability, and length, and (b) I compiled remember–familiar and confidence ratings for individual items as part of a pilot experiment and made sure that subgroups were matched, on average, for item memorability (operationalized using these ratings).

Design. There were two study–test blocks: a WI block and an SI block. For half of the participants the WI block came first; the SI block was first for the other half of the participants. Assignment of words to conditions was balanced such that each word appeared equally often as a target, interference item, and lure. Also, each word appeared equally often in the first versus the second block and in the WI versus the SI block. This balancing was accomplished by having each of the two groups serve equally often in the SI and WI blocks and by having each of the three subgroups within each group serve equally often as targets, interference items, and lures. Combining this rotation of two groups and three subgroups through conditions with whether SI or WI came first ($2 \times 3 \times 2$) created 12 between-subjects counterbalancing conditions.

The general structure of the WI and SI blocks follows the description provided in the Shared Design Elements section. In SI blocks, participants studied 50 target items once and the 50 interference items six times. In WI blocks, participants studied 50 target items once and the 50 interference items once. Five primacy buffers were presented at the beginning of each study list, and five recency buffers were presented after all targets and interference items had been presented once (but before any interference items were repeated). The lengths of the study and video game phases were complementary, such that the time between studying and being tested on a target item was equivalent for SI and WI blocks.

Procedure. Testing was done on an Apple iMac computer running PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). Before the start of the experiment, participants were shown a banker's box (approximately 1 ft [30.5 cm] wide, 2 ft [71 cm] long, 1 ft [30.5 cm] deep) on the floor of the testing room.

During the study phase, words appeared onscreen for 1,150 ms (with 500 ms between words), and participants were instructed to respond "yes" if a typical instance of that item would fit in the banker's box and to respond "no" if a typical instance of that item would not fit in the box. If the participant entered a response during the 1,150 ms interval, the computer made a *beep* noise; if the participant failed to respond within the 1,150 ms interval, the computer made a *buzz* noise. Participants were informed at the beginning of the experiment that their memory would be tested for the words they studied. They were also warned that some items would be presented multiple times at study.

For the video game phase of the experiment, participants played a Macintosh game called *Skittles*. They were told that they should try their best to accumulate points and that the experimenter would tell them when they could stop playing the game. The video game phase lasted 2 min in the SI condition and 8 min, 53 s in the WI condition.

At test, words appeared one at a time on the computer monitor. For each item, participants rated their recognition confidence on a scale from 1 to 6, as described in the *Dependent Measures* section. Participants were encouraged to spread out their confidence ratings across the 6 points of the scale. Also, if participants gave a 4, 5, or 6 response (indicating that they thought the word was studied), they were asked to make a remember–familiar judgment. Specifically, participants were asked to press the "remember" key if they recollected specific details from when the word was presented at study and to press the "familiar" key if they responded old (4, 5, or 6) because the item seemed familiar (but they did not remember any specific details). Participants were given several examples of the kinds of things that would justify a remember response (e.g., if they remembered thinking about whether the stimulus would fit in the box, if they remembered forming a mental image of the stimulus, or if they remembered how the word looked when it appeared on screen at study). The memory test was self-paced, but participants were told not to dwell too long on any one item.

Participants were given a practice phase before the start of the actual experiment in which they studied a short list of words (some of which were presented multiple times), played Skittles for 2 min, and were tested on the words they studied. Participants were informed after the practice phase that they would be cycling through the three tasks (study, video game, and test) twice. Also, they were told that each test phase contained only (a) items from the immediately preceding study phase and (b) completely new items. Therefore, for example, they did not need to worry about items from the practice or from the first study phase showing up on the second memory test.

Results

Raw data from Experiment 1 are presented in Table 1, and derived sensitivity measures are presented in Table 2. The ROC curves (plotted on the basis of the pooled data in Table 1) for the WI and SI conditions are shown in Figure 2. From this figure, it is apparent that points from the two interference conditions lie on distinct ROC curves, with a larger area under the WI ROC than under the SI ROC (thereby indicating greater overall recognition sensitivity in the WI condition). In keeping with this claim, both A_z and d_a (computed on the basis of individual participants' data) were significantly larger in the WI condition: for A_z, $F(1, 35) = 11.082$, $MSE = 0.002$; for d_a, $F(1, 35) = 16.531$, $MSE = 0.120$.

The Gaussian model I used to compute A_z and d_a was a good fit, overall, to individual participants' data. Across the 72 model fits (36 participants $\times$ SI/WI), the average chi-squared ($df = 3$) value for the Gaussian model was 3.82 ($SD = 3.57$). Because A_z and d_a may not accurately reflect sensitivity when the Gaussian model is a poor fit, I needed to ensure that the observed LSE for recognition sensitivity did not depend on inclusion of values from participants or conditions for which the Gaussian model was a poor fit. To address this concern, I reperformed the analysis, excluding participants with large chi-squared values; I used a very liberal exclusion criterion (chi-squared $p < .05$ for one of the interference conditions) to maximize the odds that participants with poor fits would be excluded.[6] Eleven out of 36 participants were excluded according to this criterion, but the overall pattern of results was unchanged. Most important, the LSE for A_z and d_a was still significant in the reanalysis.

Remember–Familiar Data

As predicted, d' computed on the basis of remember hits and false alarms showed a significant LSE, $F(1, 35) = 5.42$, $MSE = 0.135$. When d' was computed on the basis of old responses (i.e., confidence > 3), the LSE was not significant, $F(1, 35) = 1.484$, $MSE = 0.204$, $p = .23$.

Discussion

The most important result from this experiment is the finding of a significant LSE for overall recognition sensitivity, indexed using d_a and A_z. However, as in previously published list

[6] For discussion of why an alpha of .05 is a liberal value for rejecting chi-squared model fits, see, for example, Press, Teukolsky, Vetterling, and Flannery (1992), chapter 15.

Table 1
Proportions of Old, Remember, and Familiar Responses in Experiment 1 as a Function of Interference Strength, Study Status, and Confidence Criterion

	> 1		> 2		> 3		> 4		> 5	
Condition	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
Weak interference										
Studied										
Old	.98	0.00	.95	0.01	.91	0.01	.83	0.02	.70	0.03
Remember					.73	0.03	.72	0.03	.67	0.03
Familiar					.18	0.03	.12	0.02	.04	0.02
Nonstudied										
Old	.71	0.03	.40	0.03	.22	0.02	.12	0.02	.05	0.01
Remember					.05	0.01	.05	0.01	.04	0.01
Familiar					.16	0.02	.07	0.01	.01	0.00
Strong interference										
Studied										
Old	.92	0.02	.84	0.02	.77	0.02	.66	0.03	.50	0.03
Remember					.55	0.03	.54	0.03	.47	0.03
Familiar					.23	0.02	.13	0.02	.03	0.01
Nonstudied										
Old	.55	0.04	.26	0.03	.11	0.02	.05	0.01	.02	0.01
Remember					.02	0.00	.02	0.00	.01	0.00
Familiar					.09	0.01	.03	0.01	.00	0.00

Note. 1 = definitely new; 2 = probably new; 3 = guess new; 4 = guess old; 5 = probably old.

strength studies (e.g., Hirshman, 1995), I also found a LSE for bias. Figure 2 shows clearly that responding was more conservative overall (fewer hits and false alarms) in the SI condition than in the WI condition (see Hirshman, 1995, for a review of data on how list strength affects bias and a discussion of possible mechanisms of this effect).

Whenever an independent variable simultaneously affects signal-detection measures of sensitivity and bias, there is always a concern that observed sensitivity differences may be an artifact of participants shifting their criteria—it is a well-known fact that some signal-detection measures of sensitivity are affected by criterion shifts (Macmillan & Creelman, 1991; Snodgrass & Corwin, 1988). However, there is no way to explain the observed changes in A_z and d_a in terms of criterion shifts applied to a Gaussian memory process (with no real change in sensitivity). If one assumes that the underlying signal is normally distributed and that confidence judgments are made by placing an escalating set of criteria along the memory signal continuum, then sensitivity measures such as A_z and d_a that depend on the shape of the normalized

Table 2
Derived Sensitivity Measures for Experiment 1

	Weak interference		Strong interference	
Measure	*M*	*SEM*	*M*	*SEM*
A_z	.93	0.01	.89*	0.01
d_a	2.19	0.08	1.86*	0.09
d′(Remember)	2.44	0.09	2.24*	0.08
d′(Old)	2.35	0.12	2.22	0.10

Note. A_z = estimate of the area under the receiver operating characteristic curve; d_a = estimate of the distance between the studied item and lure item memory signal distributions; *d′(Remember)* = index of discrimination based on "remember" responses; *d′(Old)* = index of discrimination based on "old" responses.
* $p < .05$.

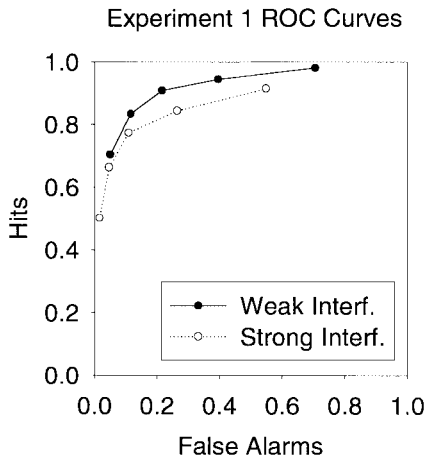

Figure 2. Experiment 1: Receiver operating characteristic (ROC) curves (plotting hits vs. false alarms for different confidence criteria) for the weak interference (WI) and strong interference (SI) conditions of Experiment 1. These ROC curves are plotted on the basis of the pooled data contained in Table 1. The area under the ROC is larger in the WI condition than the SI condition, indicating that sensitivity is higher in the WI condition. Interf. = interference.

ROC would be invariant as a function of criterion placement (Macmillan & Creelman, 1991).[7]

It is worth noting that, under different assumptions about how confidence judgments are generated, criterion placement can affect ROC parameters. Van Zandt (2000) demonstrated that if raw memory scores feed into a "horse race" decision process, in which evidence accumulates in parallel for whether an item is old or new (and confidence depends on the distance between these accumulators when the decision criterion is reached), adopting a more conservative criterion can reduce the slope and intercept of the normalized ROC and thus could affect sensitivity measures that are computed from this slope and intercept, such as d_a. However, there is no evidence that the LSE observed here is an artifact of decreased slope in the more conservative condition; the slope of the normalized ROC (estimated separately for individual participants using RSCORE PLUS) was numerically slightly higher in the condition for which responding was more conservative: slope = .69 (SEM = 0.07) in the SI condition and slope = .68 (SEM = 0.04) in the WI condition.

Having demonstrated a LSE for recognition sensitivity, the next step is to isolate the underlying causes of this effect. One possibility, mentioned earlier, is that list strength impairs recollection of specific details from the study phase but does not impair recognition discrimination based on familiarity. Self-report data collected in Experiment 1 provide some evidence in favor of this idea; when discrimination (d') was computed on the basis of remember responses (which should reflect the contribution of recollection), there was a significant LSE, but the LSE was not significant when d' was computed on the basis of old responses (which can be driven either by recollection or familiarity).

However, although these results provide converging evidence in favor of my dual-process hypothesis, they are far from definitive. Measures of sensitivity that are based on a single hit–false alarm pair (such as A' and d') can be strongly affected by criterion shifts (Donaldson, 1993; Macmillan & Creelman, 1991). Also, the very low rate of remember false alarms can distort d' values. Thus, there is no way to definitively rule out the possibility that the observed LSE for d'*(Remember)* and the difference in the size of the LSE for d'*(Remember)* versus d'*(Old)*, could both be artifacts of participants responding more conservatively in the SI condition.[8] To obtain stronger evidence in favor of my dual-process hypothesis, I needed to find some way of isolating the contribution of recollection that did not force reliance on single-point estimates of sensitivity such as d' (which, as discussed above, are hard to interpret when bias is changing and false alarms are low). This is what I set out to accomplish in Experiment 2.

Experiment 2

One way to increase the extent to which recollection (vs. familiarity) contributes to recognition performance is to increase target–lure similarity. In situations in which lures are highly similar to studied items, both studied items and lures will trigger strong feelings of familiarity, forcing participants to rely on recollection of specific, discriminative details in order to respond differentially to studied items and lures (for empirical evidence in support of this claim, see, e.g., Hintzman & Curran, 1994; Rotello, Macmillan, & Van Tassel, 2000). If use of lures that are highly similar to studied items (*related lures*) bolsters the relative contribution of recollection, and list strength affects recollection (but not familiarity), then one would expect to find a significant LSE when related lures are used at test. In contrast, the LSE may not be significant when lures are unrelated to studied items, insofar as discrimination between studied items and nominally unrelated lures can be supported by familiarity (as well as recollection).

To test this hypothesis, I used a variant of the plurals recognition paradigm introduced by Hintzman et al. (1992), in which participants studied singular and plural words. At test, there were two kinds of lures: related SP lures (e.g., study *scorpions*, test with *scorpion*) and unrelated lures (e.g., study *scorpions*, test with *banana*). Participants were instructed to say "old" if the test word exactly matched a studied word, and to say "new" otherwise. According to the above account, the ability to discriminate between studied words and related SP lures should depend on recollection. Thus, one should find a significant LSE for studied versus SP lure discrimination but not necessarily for studied versus unrelated lure discrimination.

Furthermore, one can also look at SP versus unrelated lure *pseudodiscrimination*—that is, how much more likely participants are to say "old" to related versus unrelated lures. Familiarity supports pseudodiscrimination (insofar as SP lures are more familiar than unrelated lures), but recollection of plurality information lowers pseudodiscrimination by allowing participants to confidently reject SP lures (i.e., participants can confidently reject the word *scorpion* if they recollect having studied *scorpions*; for evidence that this *recall-to-reject* process contributes to performance on plurality recognition tests, see Rotello et al., 2000). If increasing list strength reduces recollection, but has no effect (or a positive effect) on discrimination based on familiarity, the net effect should be an increase in pseudodiscrimination. Hence, I predicted a negative LSE for pseudodiscrimination (i.e., it should be higher in the SI condition than in the WI condition).

Method

Participants. Eighty University of Colorado undergraduates and graduate students (49 women and 31 men, mean age = 20.3 years) volunteered to participate in the experiment. The experiment lasted approximately 1 hr and participants were either paid $10 or given course credit.

Materials. Stimuli were 250 highly imageable, concrete, familiar medium-frequency nouns; for all of these words, the plural form of the

[7] Hintzman (2001; see also Wickelgren, 1968) pointed out that criterion variability within a particular condition can reduce ROC-based estimates of sensitivity. If criterion placement happened to be more variable in the SI (vs. WI) condition, this could result in an artifactual LSE for A_z and d_a. However, there is no reason to think that criterion placement would be more variable in the SI condition.

[8] I should point out that although criterion shifts can affect d', there is no way to explain the observed LSE for d'*(Remember)* in terms of the mere fact that responding was more conservative in the SI condition. When the slope of the normalized ROC is less than 1 (as was the case in Experiment 1), adopting a more conservative criterion leads to an increase in d' (Donaldson, 1993; Macmillan & Creelman, 1991). Thus, I cannot explain the observed decrease in d' purely in terms of a criterion shift. The low overall rate of remember false alarms in the SI condition, which forced me to use the correction for zero false alarms more frequently in this condition than in the WI condition, is a more serious problem with respect to interpreting the LSE for d'*(Remember)*.

word was generated by adding *s* to the singular form of the word. Practically all of these words were also used as stimuli (in their singular form) in Experiment 1; therefore, the overall characteristics of the stimuli used in this experiment were practically identical to the characteristics of the stimuli used in Experiment 1. In addition to the aforementioned 250 words, 20 other words were used as primacy and recency buffers at study. The 250 words not used as buffers were split into 10 groups of 25 words. These groups were matched, on average, for important word characteristics such as word frequency, as well as memorability (see the *Method* section of Experiment 1 for more details).

Design. Apart from the plurality manipulation, the design of this experiment was very similar to the design of the prior experiment. There were two study–test blocks: a WI block and a SI block. For half of the participants the WI block came first; the SI block was first for the other half of the participants. Half of the items on the study list were studied in their singular form, and half were studied in their plural form. In situations in which an item was presented repeatedly at study, the item was always presented either as a singular word or as a plural word—in no case was an item studied in both its singular form and its plural form. The 10 word groups were rotated across the 10 conditions shown in Table 3 to ensure that words from each group served equally often in each condition.

Finally, each item was studied equally often (across participants) in its singular and plural form. Combining all of these factors together created 40 between-subjects counterbalancing conditions: (rotate 10 word groups across 10 conditions) × (study each item as both a singular and a plural word) × (Weak block first vs. Strong block first).

The overall structure of the WI and SI study lists was the same as in Experiment 1: Each list contained 50 targets and 50 interference items; interference items were studied one time in the WI list versus six times in the SI block (see the Shared Design Elements section for more details). A minor difference between Experiment 2 and the preceding experiment is that stimuli were presented in a random order in this experiment (subject to the constraints outlined in Figure 1), whereas the preceding experiment presented stimuli in a fixed order for a given counterbalancing condition.

A different video game (Gem Master) was used in this experiment. As in Experiment 1, the length of the video game phase was complementary to the length of the study phase, such that the total time elapsed between studying a target item and being tested on that item was the same in the WI and SI conditions. The video game lasted 2 min in the SI condition and 8 min, 53 s in the WI condition.

The recognition test comprised 25 studied target items (items presented in the same plurality at study and test), 25 SP lures (target items that were presented in a different plurality at study vs. at test), and 25 unrelated lures (items that were not presented in either plurality at study). The 75 test items were presented in a random order, with the constraint that each miniblock of 15 items consisted of 5 studied words, 5 SP lures, and 5 unrelated lures.

Table 3
Counterbalancing Conditions for Experiment 2

Block	Study as	Test
First	Target	Same plurality
First	Target	Switched plurality
First	Interference	
First	Interference	
First		Unrelated lure
Second	Target	Same plurality
Second	Target	Switched plurality
Second	Interference	
Second	Interference	
Second		Unrelated lure

Note. Word groups were rotated across these 10 conditions such that each group served equally often in each of the conditions.

After the aforementioned 75 items were presented, participants were given 15 extra test items: 5 studied interference items, 5 lures generated by switching the plurality of studied interference items, and 5 more unrelated lures; these different groups were randomly mixed together. I did not score these extra 15 test trials; the purpose of testing interference items was to reinforce the idea that participants should pay attention to interference items at study.

Procedure. Testing was done on a Dell Dimension computer running E-Prime software (E-Prime, 2002). The study procedure was very similar to the procedure used in Experiment 1: Words appeared onscreen for 1,150 ms (with 500 ms between words). The main difference is that, in this experiment, participants were asked to pay close attention to the plurality of studied items, and the encoding task was modified to force participants to attend to plurality. Specifically, participants were told that if the word was plural, they should picture more than one of that object and say whether multiple (i.e., at least two) copies of that object would fit in the box and that if the word was singular, they should picture only one of that object and say whether that single object would fit in the box. The instructions repeatedly emphasized that—to have good plurality memory—participants had to actively try to picture multiple objects for plural words and single objects for singular words. If participants failed to enter a response ("no" or "yes") within the 1,150 ms interval in which an item was onscreen, the experiment was temporarily suspended and a message appeared onscreen telling them to respond more quickly; participants had to press the space bar to continue.

At test, participants had to make a studied–nonstudied judgment for each item. Participants were told to respond "studied" if the test word exactly matched a word that was studied during the size judgment task (i.e., they studied this word in this plurality), and they were told to respond "nonstudied" if the test word did not exactly match a studied word. Participants were also told to be very particular about the plurality of test words (i.e., if *scorpion* is presented at study but its plural form, *scorpions*, is presented at test, the correct answer would be nonstudied). For each item, after participants made their studied–nonstudied response, they were asked to rate their confidence on a 3-point scale (1 = *guess*; 2 = *probably right*; 3 = *sure*). Participants were encouraged to spread out their confidence ratings across the entire scale. When I analyzed the data, I converted confidence ratings to a 6-point scale that matches the scale used in Experiment 1 (1 = *definitely new*; 2 = *probably new*; 3 = *guess new*; 4 = *guess old*; 5 = *probably old*; 6 = *definitely old*). The test was self-paced, but participants were told not to dwell too long on any one item. As in the preceding experiments, participants were given a short practice study and test phase before the start of the actual experiment.

Results

Raw data are presented in Table 4 and derived sensitivity measures are presented in Table 5. I was interested in three different kinds of recognition discrimination: studied versus unrelated lure discrimination, studied versus SP lure discrimination, and SP versus unrelated lure pseudodiscrimination.

To compute sensitivity, I generated six ROC curves for each participant (SI and WI × the three different types of discrimination: studied versus unrelated, studied versus SP, and SP versus unrelated pseudodiscrimination). A Gaussian model was fit separately to each of these curves, and I used the parameters of the best-fitting Gaussian to compute A_z and d_a for each curve. Figure 3 provides an overview of the data; it plots (using the pooled data in Table 4) ROC curves for the three types of discrimination as a function of interference strength. As with Experiment 1, I am using Gaussian models to compute sensitivity for a purely pragmatic reason—they provide a good fit to the data—not because I have

Table 4
Proportions of Old Responses in Experiment 2 as a Function of Interference Strength, Study Status, and Confidence Criterion

	> 1		> 2		> 3		> 4		> 5	
Condition	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
Weak interference										
Studied	.92	0.01	.82	0.01	.77	0.01	.73	0.02	.55	0.02
Switched plurality	.72	0.02	.52	0.02	.45	0.02	.40	0.02	.24	0.02
Unrelated	.65	0.03	.33	0.02	.18	0.01	.13	0.01	.05	0.01
Strong interference										
Studied	.82	0.02	.70	0.02	.64	0.02	.60	0.02	.38	0.02
Switched plurality	.64	0.02	.43	0.02	.36	0.02	.30	0.02	.14	0.01
Unrelated	.41	0.03	.19	0.02	.09	0.01	.06	0.01	.02	0.00

Note. 1 = definitely new; 2 = probably new; 3 = guess new; 4 = guess old; 5 = probably old.

any commitment to the idea that the underlying distributions are Gaussian.

As predicted, there was a significant LSE for studied versus SP discrimination, indexed using both A_z and d_a: for A_z, $F(1, 79) = 6.076$, $MSE = 0.008$; for d_a, $F(1, 79) = 7.125$, $MSE = 0.139$. The LSE for studied versus unrelated discrimination was numerically smaller, and not significant: for A_z, $F(1, 79) = 1.755$, $MSE = 0.005$, $p = .19$; for d_a, $F(1, 79) = 0.704$, $MSE = 0.201$, $p = .40$. Finally, there was a significant negative LSE for SP versus unrelated pseudodiscrimination: for A_z, $F(1, 79) = 21.024$, $MSE = 0.011$; for d_a, $F(1, 79) = 18.690$, $MSE = 0.221$.

The Gaussian model I used to compute A_z and d_a was a very good fit to individual participants' data. Across the 480 model fits (80 participants × SI and WI × the three types of discrimination), the average chi-squared ($df = 3$) value for the Gaussian model was 2.69 ($SD = 2.48$). To address the concern that A_z and d_a may not accurately reflect sensitivity when the Gaussian model is a poor fit, I reperformed the analysis, excluding participants with high chi-squared values. As in Experiment 1, I used a very liberal exclusion criterion to maximize the odds of rejecting participants with poor fits (chi-squared p value $< .05$ for at least one of the conditions; 16 out of 80 participants were excluded according to this criterion). The results of this reanalysis were qualitatively identical to the original results: There was a significant LSE for studied versus SP discrimination, a significant negative LSE for SP versus unrelated pseudodiscrimination, and the LSE was not significant for studied versus unrelated discrimination.

Table 5
Derived Sensitivity Measures for Experiment 2

	A_z		d_a	
Interference strength	*M*	*SEM*	*M*	*SEM*
Studied versus switched plurality				
Weak	0.72	0.01	0.87	0.06
Strong	0.68*	0.01	0.71*	0.06
Studied versus unrelated				
Weak	0.85	0.01	1.55	0.06
Strong	0.83	0.01	1.49	0.07
Switched plurality versus unrelated				
Weak	0.61	0.02	0.44	0.07
Strong	0.69*	0.02	0.76*	0.07

Note. A_z = estimate of the area under the receiver operating characteristic curve; d_a = estimate of the distance between studied-item and lure-item memory signal distributions.
* $p < .05$.

Discussion

The hypothesis that list strength reduces recollection (but not familiarity-based discrimination) led to two major predictions for this experiment. First, there should be a LSE for studied versus SP lure discrimination, insofar as this kind of discrimination depends on recollection. Second, there should be a negative LSE for SP versus unrelated lure pseudodiscrimination: Recollection of plurality information reduces pseudodiscrimination by allowing participants to confidently reject SP lures; thus, if list strength reduces recollection, this should improve pseudodiscrimination. Both of these predictions were confirmed by the data.

The SP versus unrelated ROC curves in Figure 3 specifically indicate that increasing list strength reduces participants' ability to confidently reject SP items. The height of the rightmost point on the ROC corresponds to the proportion of SP items that are given a confidence rating greater than 1. Thus, a decrease in the number of definitely new (confidence = 1) responses to SP items should show up as an increase in the height of the rightmost ROC point. This is exactly what we found; for SP versus unrelated pseudodiscrimination, the rightmost point of the SI ROC is clearly "bumped up" relative to the WI ROC.

Some, but not all, of the results of this experiment are consistent with a "bias shift-only" explanation. For studied versus SP discrimination, the slope and intercept of the normalized ROC were lower in the SI condition, for which responding was more conservative. Van Zandt (2000) showed how the observed changes in ROC parameters (decreased slope and intercept) can arise purely as a function of participants adopting a more conservative criterion, without any real change in sensitivity. However, there is no way to explain the LSE for SP versus unrelated pseudodiscrimination purely in terms of participants shifting their criteria. The Van Zandt theory predicts that adopting a more conservative

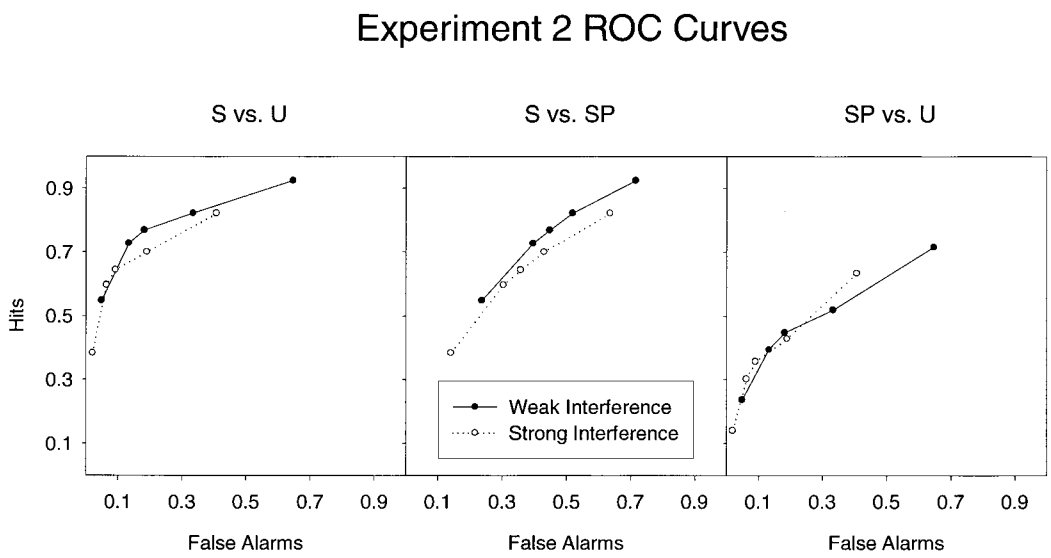

Figure 3. Experiment 2: Receiver operating characteristic (ROC) curves, looking at studied versus unrelated lure (S vs. U) discrimination, studied versus switched-plurality lure (S vs. SP) discrimination, and switched-plurality versus unrelated lure (SP vs. U) pseudodiscrimination, as a function of interference strength. These ROC curves are plotted on the basis of the pooled data contained in Table 4.

criterion would result in a monotonic decrease in z-ROC slope, but for SP versus unrelated pseudodiscrimination, z-ROC slope was significantly higher in the more conservative (SI) condition: slope = .93 (*SEM* = 0.07) in the SI condition and slope = .66 (*SEM* = .04) in the WI condition, $F(1, 79) = 12.975$, *MSE* = 0.223. Thus, bias shift-only accounts do not provide a coherent account of the entire pattern of data (whereas the hypothesis that list strength impairs recollection does provide a coherent account of these data).

The LSE for studied versus unrelated lure discrimination (indexed using A_z and d_a) was smaller in Experiment 2 than in Experiment 1. In Experiment 2, the mean LSE for A_z was 0.015 (*SEM* = 0.011); in Experiment 1, the mean LSE for A_z was 0.036 (*SEM* = 0.011); the LSE was significant in Experiment 1 but not in Experiment 2. The dual-process hypothesis implies that the size of the LSE for studied versus unrelated lure discrimination is a function of exactly how much recollection (relative to familiarity) is contributing to recollection. Thus, one may be able to explain the reduced size of the LSE in terms of the idea that recollection was contributing less to studied versus unrelated lure discrimination in this experiment than in Experiment 1. There are several reasons why this may have been the case. First, recollection of item information (i.e., did I study this word, regardless of plurality?) is diagnostic in Experiment 1, but item information alone is not diagnostic in this experiment—if one does not remember plurality, one cannot be sure that one studied that exact word. It therefore stands to reason that participants would weigh item recollection less heavily in Experiment 2 than in Experiment 1. Also, I collected remember–familiar data in Experiment 1 but not in Experiment 2; use of remember–familiar testing may induce participants to pay more attention to recollection (in situations in which familiarity also discriminates) than they would otherwise.

General Discussion

In Experiment 1, I obtained a significant LSE for recognition sensitivity, indexed using A_z and d_a. This result is important insofar as every published list strength study prior to this one has reported a null LSE for recognition sensitivity (e.g., Ratcliff et al., 1990). To account for this novel result, I argue that increasing list strength impairs recollection of specific studied details but that it does not affect recognition discrimination driven by familiarity. This implies that a LSE for recognition sensitivity should be present when the contribution of recollection is large relative to the contribution of familiarity. Conversely, the LSE for recognition sensitivity should be null (or negative) to the extent that participants are relying on familiarity. I obtained some suggestive evidence in favor of this dual-process hypothesis in Experiment 1 through the use of self-report measures of recollection (remember responses). Discrimination computed on the basis of remember responses showed a significant LSE, but the LSE was not significant when discrimination was computed on the basis of old responses (which can be driven by either familiarity or recollection). In Experiment 2, I obtained more definitive evidence in favor of the dual-process hypothesis. The logic of this experiment centered on the idea that recollection is especially important when related lures (lures that are highly similar to specific studied items) are used at test. Thus, recognition tests with related lures should show a LSE, but discrimination of studied items versus unrelated lures may not show a LSE insofar as this kind of discrimination can be supported by familiarity (as well as recollection). In Experiment 2, I used a plurality discrimination paradigm in which participants were given related SP lures as well as unrelated lures at test. The results were exactly as I had predicted: There was a significant LSE for studied versus SP lure discrimination (indexed using A_z and d_a), and there was a significant negative LSE for SP

versus unrelated lure pseudodiscrimination, but the LSE for studied versus unrelated lure discrimination was not significant.

Relation to Other List Strength Studies

In the experiments reported here, I replicated the null LSE that other studies have found when single-point estimates of recognition sensitivity (e.g., d') are computed on the basis of whether participants think the item is old or new and when lures are not strongly related to studied items. More specifically, in Experiment 1 and Experiment 2 (unrelated lure condition), the LSE was not significant when d' or A' was computed on the basis of the probability of calling an item *old* (i.e., assigning a confidence rating > 3).[9] The fact that I replicated the null LSE for old–new recognition sensitivity despite the use of a paradigm that differs in several ways from the paradigm used in other list strength studies, attests to the robustness of this finding. The results of Experiment 1 also show that d' (computed on the basis of old responses) may fail to detect sensitivity differences that are revealed by other, multipoint measures of sensitivity; whereas the LSE for *d′(Old)* was not significant, the LSE was highly significant for A_z and d_a.

This study is the first list strength study to use lures that were highly similar to studied items (i.e., the SP lures used in Experiment 2). Thus, the fact that I found LSEs for discrimination using highly similar lures does not directly contradict extant results. Other studies have used related lures (e.g., Shiffrin et al., 1995, used nonstudied category exemplars from studied categories as lures, as did Ratcliff et al., 1994, Experiment 6), but lures in these studies were not nearly so similar (to studied items) as the lures used here. As such, participants may have been able to rely on familiarity (which, by hypothesis, is unaffected by list strength) in these experiments.

The results presented by Ratcliff et al. (1994) provide an interesting counterpoint to our results. Just as I did in Experiment 1, Ratcliff et al. (1994, in their Experiments 1–5) used nominally unrelated lures, manipulated list strength, collected confidence ratings at test, and plotted ROC curves on the basis of confidence data. However, in contrast to my Experiment 1 results, the results presented by Ratcliff et al. (1994) do not show any clear evidence of a LSE for recognition sensitivity.

At this point, one can only speculate as to why my Experiment 1 found a LSE for recognition sensitivity but Ratcliff et al.'s (1994) experiments did not. In my experiments, I took several steps to boost the size of the LSE. My experiments used a powerful strength manipulation: Interference items were presented six times in the SI condition versus once in the WI condition. Also, I used an encoding task that was designed to yield a moderate level of memory trace distinctiveness (such that traces would be rich and distinctive enough to support recollection but overlap enough to yield interference). In contrast, Ratcliff et al. (1994) used a less extensive strength manipulation, and their experiments did not use an encoding task (apart from "learn these words"). Furthermore, some (but not all) of the Ratcliff et al. (1994) experiments that failed to obtain a LSE used short study durations (on the order of 50–100 ms), which may have led to floor effects on recollection (see Gardiner & Gregg, 1997, for evidence that recollection is very poor following shallow encoding and brief study presentations). Another factor to consider is test instructions. As I mentioned earlier, it is possible that the remember–familiar instructions that I used in Experiment 1 led participants to pay more attention to recollection than they would have otherwise, thereby boosting the size of the LSE. Additional research is needed to determine which, if any, of the aforementioned factors are responsible for the observed differences in the size of the LSE for recognition sensitivity.

Finally, I should note that my dual-process hypothesis clearly predicts a LSE for tests of cued recall, insofar as cued recall involves recollection of specific details from the study phase. However, some studies have failed to find a significant LSE for cued recall with unrelated word-pair stimuli (e.g., Ratcliff et al., 1990, Experiment 3). Furthermore, in other studies that have found a significant LSE for cued recall, the size of this effect was quite small (e.g., Ratcliff et al., 1990, Experiment 6). I think that the small size of the LSE for cued recall in published studies may be attributable to these studies' failure to tightly control encoding processes and to their use of a less-than-maximally powerful list strength manipulation. This hypothesis needs to be tested directly. One prediction is that use of an encoding task (like the task used here) designed to minimize floor effects on recollection (while ensuring some degree of trace overlap) should bolster the LSE for cued recall relative to a condition in which participants are told only to learn the words. Another prediction is that increasing the amount of strengthening that occurs at study (i.e., the number of interference item repetitions) should bolster the LSE for cued recall.

Implications for Extant Mathematical Models of the LSE

Up to this point, almost all theoretical work on the LSE for recognition has been conducted within the framework of single-process, *global matching* mathematical models (for a review of global matching models, see Clark & Gronlund, 1996). According to these models, recognition decisions are based (in their entirety) on a scalar signal that indexes how well the test probe matches each of the items stored in memory.

Ratcliff et al.'s (1990) finding of a null LSE for recognition sensitivity was a watershed event in the development of mathematical models of recognition memory. Most global matching models in the literature circa 1990 predicted that increasing list strength should impair recognition discrimination (see Shiffrin et al., 1990, for a discussion of the issue). Search of Associative Memory (SAM; Gillund & Shiffrin, 1984) is typical of these models; in SAM, increasing list strength increases the mean global match signal triggered by both targets and lures, and the variance of the global match signal (intuitively, the consequences of test probe X spuriously matching memory trace Y are larger when memory trace Y is strong than when memory trace Y is weak). This increase in variance leads to decreased discriminability. Researchers have been working since 1990 to modify global matching models so that they predict the null LSE for recognition obtained by Ratcliff et al. (1990).

Just as the Ratcliff et al. (1990) results pose problems for recognition models that predict that LSEs will always be obtained,

[9] In Experiment 2, *d′(Old)* for studied versus unrelated discrimination was 1.82 ($SEM = 0.08$) in the WI condition and 1.90 ($SEM = 0.07$) in the SI condition, $F(1, 79) = 0.974$, $MSE = 0.250$, $p = .33$.

the results reported here pose problems for recognition models that predict LSEs will never be obtained for item recognition sensitivity (e.g., Dennis & Humphreys, 2001; Murdock & Kahana, 1993). Murdock and Kahana (1993) argued that items from the current study list make a negligibly small contribution to the variance of the global match signal relative to the contribution of all the other items that have been studied (over the person's lifetime); thus, strengthening items from the current list would not boost variance enough to hurt recognition. Dennis and Humphreys (2001) argued that, in recognition tests that use single words as stimuli, the primary source of noise when a word is presented at test is exposure to that word outside of the experimental context (*context noise*). According to this model, other words from the study list do not affect the memory signal triggered by a word at test (i.e., there is no *item noise*); as such, strengthening some list items should have no effect on memory for other nonstrengthened list items. These two models, in their present form, cannot accommodate the significant LSEs for recognition sensitivity reported here, except by arguing that list strength is confounded with some other factor. For example, Dennis and Humphreys (2001) argued that—in retroactive interference designs—list length and strength effects could arise spuriously if participants mentally focus on the latter part of the SI list (which does not contain target items) when making recognition judgments at test.

Another approach to modeling the null LSE for recognition sensitivity is to posit that *differentiation* occurs as a consequence of strengthening (Shiffrin et al., 1990); the gist of differentiation is that, as participants acquire experience with an item, the item's representation becomes increasingly refined and increasingly distinct from the representations of other items. In models in which differentiation occurs, strengthening a memory trace decreases the odds that it will (spuriously) match a lure at test. Therefore, increasing list strength may actually reduce variability by reducing the number of spurious matches to interference items (both the Retrieving Effectively from Memory [REM] model presented by Shiffrin & Steyvers, 1997, and the model presented by McClelland & Chappell, 1998, have this property). Recognition models like REM can accommodate both the null LSE for recognition sensitivity reported by Ratcliff et al. (1990), and the significant LSE reported in this study, depending on parameter settings. However, it remains to be seen whether REM can accommodate the specific pattern of results reported here—for example, the finding in Experiment 2 of a LSE for studied versus SP discrimination, coupled with a negative LSE for SP versus unrelated pseudodiscrimination.

At present, the only model of recognition that has been shown to fit my data is the dual-process Complementary Learning Systems (CLS) neural network model (Norman & O'Reilly, in press). Unlike single-process models such as REM, this model incorporates distinct components that are responsible for familiarity and recollection, and the operating characteristics of these components are constrained by data regarding the neural substrates of familiarity and recollection. Motivating why the CLS model predicts differential effects of list strength for recollection and familiarity is beyond the scope of this article—see Norman & O'Reilly (in press) for a detailed discussion of the model's list strength predictions.

Conclusion

The experiments reported here show that LSEs sometimes are obtained for recognition sensitivity. Furthermore, the hypothesis that list strength impairs recollection but not familiarity appears to be a useful guide as to whether LSEs are obtained. Future research in my laboratory will explore the boundary conditions of the LSEs reported here (i.e., why they were obtained here but not in other studies, such as Ratcliff et al., 1994). Also, although the results reported here are certainly consistent with the dual-process hypothesis, more work is needed to assess whether this dual-process account provides a better account of these results than extant single-process models (e.g., REM; Shiffrin & Steyvers, 1997).

References

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3,* 37–60.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, & Computers, 25,* 257–271.

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

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

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General, 121,* 275–277.

Donaldson, W. (1993). Accuracy of d' and A' as estimates of sensitivity. *Bulletin of the Psychonomic Society, 31,* 271–274.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523–533.

Dorfman, D. D., & Alf, E., Jr. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data. *Journal of Mathematical Psychology, 6,* 487–496.

Dorfman, D. D., Beavers, L. L., & Saslow, C. (1973). Estimation of signal detection theory parameters from rating-method data: A comparison of the method of scoring and direct search. *Bulletin of the Psychonomic Society, 1,* 207–208.

E-Prime (Version 1.0) [Computer software]. (2002). Pittsburgh, PA: Psychology Software Tools.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition, 16,* 309–313.

Gardiner, J. M., & Gregg, V. H. (1997). Recognition memory with little or no remembering: Implications for a detection model. *Psychonomic Bulletin & Review, 4,* 474–479.

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

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics.* Huntington, NY: Krieger.

Harvey, L. O., Jr. (2001). *Parameter estimation of signal detection models: RSCORE PLUS user's manual* [Computer software manual]. Boulder, CO: Author.

Hintzman, D. L. (2001). Judgments of frequency and recency: How they relate to reports of subjective awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1347–1358.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33,* 1–18.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and

repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 667–680.
Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302–313.
Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember–know paradigm. *Memory & Cognition, 25,* 345–351.
Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen & J. W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 13–47). Mahwah, NJ: Erlbaum.
Kahana, M. J., & Rizzuto, D. (2002). *An analysis of the recognition–recall relation in four distributed memory models.* Manuscript submitted for publication.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.
Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.
Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 8,* 252–271.
McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724–760.
Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 689–697.
Murnane, K., & Shiffrin, R. M. (1991a). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 855–874.
Murnane, K., & Shiffrin, R. M. (1991b). Word repetitions in sentence recognition. *Memory & Cognition, 19,* 119–130.
Norman, K. A., & O'Reilly, R. C. (in press). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review.*
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing* (2nd ed.). New York: Cambridge University Press.
Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition, 21,* 89–102.
Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178.
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,* 763–785.
Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535.
Rotello, C., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43,* 67–88.
Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 267–287.
Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 179–195.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166.
Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50.
Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26,* 1–12.
Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology, 92,* 297–304.
Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600.
Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology, 5,* 102–122.
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 and Cognition, 5,* 418–441.
Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 345–355.

Received October 22, 2001
Revision received April 27, 2002
Accepted May 21, 2002 ■