Differential Effects of List Strength on Recollection and Familiarity
Kenneth A. Norman
University of Colorado, Boulder

in press, Journal of Experimental Psychology: Learning, Memory, & Cognition

The author can be contacted by email at knorman@princeton.edu
Numerous studies have found a null list strength effect (LSE) for recognition sensitivity: Strengthening memory traces associated with some, but not all, studied items does not impair recognition of non-strengthened studied items. In Experiment 1, we set out to test the generality of this result; contrary to prior research, we found a LSE using ROC-based measures of recognition sensitivity (A\text{z} and d\text{a}). To account for the discrepancy between our result and prior results, we argue that a LSE is present for recollection but not for discrimination based on familiarity. Thus, the size of the LSE will depend on the extent to which recollection is driving recognition. In Experiment 1, we obtained suggestive evidence in favor of this hypothesis using self-report measures of recollection. In Experiment 2, we set out to increase the contribution of recollection by using switched-plurality (SP) lures that were highly similar to studied items. As predicted, there was a significant LSE for comparisons involving SP lures. In contrast, the LSE for discrimination of studied items and nominally unrelated lures was not significant.
Differential Effects of List Strength on Recollection and Familiarity

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), and seeing 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 discussion of confounds that are frequently present in list length experiments). A related question is how increasing list strength -- strengthening the memory traces associated with some, but not all, list items -- affects memory for non-strengthened 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 non-strengthened items.

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

In this paper, we 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., Mandler, 1980; Hintzman & Curran, 1994; Jacoby, Yonelinas, & Jennings, 1997; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). Putting these two claims together: If recognition sensitivity is driven, in part, by recollection, and recollection is impaired by list strength, then it should be possible to observe a LSE for recognition sensitivity when recollection is contributing (see Norman & O'Reilly, under revision, for a recently developed computational model of recognition memory that predicts a LSE for recollection but not familiarity).

The paradigm that we use in this paper 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, we show that it is possible to obtain a LSE for recognition sensitivity, and we present some suggestive evidence (from self-report data) that the LSE we observe is attributable to recollection. In Experiment 2, we present stronger evidence that list strength impairs recollection, using a paradigm where
participants have to discriminate between studied items and related, switched-plurality 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 “strong” items and “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 non-strengthened (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 list strength effects for strong items in terms of ceiling effects). To get around this limitation, the full “mixed-pure” design was not used in the experiments reported here. Rather, a simplified design was used with only two kinds of lists, Weak Interference and Strong Interference. Both types of lists were comprised of Target (to-be-tested) items, and non-tested Interference items. Target items were presented once in both conditions; list strength was manipulated by presenting Interference items once in the Weak Interference condition, vs. multiple times in the Strong Interference condition. The effect of list strength could be measured by comparing memory for Targets in the Weak vs. Strong Interference conditions. A key facet of this design is that, because Interference items were not tested, these items could be overlearned in the Strong Interference condition without any adverse consequences. Taking advantage of this fact, Interference items were presented six times on Strong Interference lists in the experiments reported here.

In all of the experiments reported here, participants studied (and were tested on) one Weak Interference list and at least one Strong Interference list. Figure 1 shows the general structure of the Strong and Weak Interference blocks.

\begin{figure}
\centering
\caption{The general structure of the Strong and Weak Interference blocks.}
\end{figure}

In both kinds of blocks, participants first studied Target and Interference items once (mixed together). For Strong Interference lists, participants then studied the list of Interference items five more times. The Interference items were presented in a different order each time that the list of Interference items was repeated.

Participants played a video game immediately after each study list; Weak Interference study lists were followed by a long video game phase, and Strong Interference 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 Weak Interference and Strong Interference conditions. The Strong Interference video game phase lasted two minutes and the Weak Interference 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 Weak Interference condition, then it will not be possible to observe a decrease in recollection in the Strong Interference condition. At the other end of the scale, if memory traces are too distinctive (such that overlap between traces is minimal), then interference effects will not 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 we 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 pre-specified dimensions. Participants were given just over a second to make their size judgment. This encoding duration was selected because it gives participants just enough time to form a mental image of the item, and see 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 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 another.

**Eliminating rehearsal confounds**

As discussed by Ratcliff et al. (1990), if participants rehearse weak items at the expense of strong items in the Strong Interference (“mixed list”) condition, this can mask a LSE by artificially boosting memory for weak items. Furthermore, if participants rehearse strong items at the expense of weak items in the Strong Interference condition, this can result in a spurious LSE by artificially reducing weak-item memory (this occurred in 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, so 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, so there was no way for participants to redistribute rehearsal according to strength.

**Dependent Measures**

We 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”. We then generated ROC curves for individual participants by computing hits and false alarms based on different confidence criteria (Macmillan & Creelman, 1991). To index sensitivity, for each participant and for each Interference Strength condition we used the RSCORE PLUS algorithm (Harvey, 2001) to fit a Gaussian model to that participant/condition's
confident rating data\textsuperscript{1}; we 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 we are 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 purely for pragmatic reasons: Gaussian models tend to provide a good fit to recognition memory ROC curves -- 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, we provide evidence that, overall, Gaussian models provide a good fit to the data, which in turn validates our use of $A_z$ to index sensitivity. For comparison purposes, we also used ROC data to compute the sensitivity measure $d_z$ (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 that we present here depend on our use of $A_z$ as opposed to $d_z$; the LSE was significant for $d_z$ if and only if it was significant for $A_z$\textsuperscript{2}.

Next, as a preliminary means of getting at the idea that list strength affects recollection, but not familiarity, we collected self-report measures of recollection in Experiment 1: Whenever a participant thought that an item was "old" (i.e., they assigned the item a confidence rating > 3 on the 6-point scale), we asked them whether they "remembered" studying the item (i.e., they recollected specific details) or whether the item just seemed "familiar", but no specific details came to mind (Tulving, 1985; Gardiner, 1988; Rajaram, 1993). Our logic in collecting this data was quite straightforward: If recollection is contributing to recognition performance, "remember" responses should isolate this contribution more so than "old" responses.\textsuperscript{3} Thus, under the assumption that list strength affects recollection, we 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, we

\textsuperscript{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”).

\textsuperscript{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 that we obtained using $A_g$ were qualitatively identical to the results that we obtained using $d_z$ and $A_z$; to save space, these results are not presented here.

\textsuperscript{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) and Hirshman & Master (1997) vs. Gardiner & Gregg (1997).
computed \( d' \) based on "remember" and "old" responses in Experiment 1; we will refer to these measures as \( d'(\text{Remember}) \) and \( d'(\text{Old}) \), respectively. 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 hour and participants received course credit.

**Materials.** Stimuli were 300 highly imageable, concrete, familiar medium-frequency nouns; imageability, concreteness, familiarity, and Kucera-Francis frequency data were obtained from the MRC Psycholinguistic Database (Coltheart, 1981): mean imageability = 5.76 out of 7, range = 5.02 to 6.59; mean concreteness = 5.83 out of 7, range = 5.00 to 6.48; mean familiarity = 5.02, range = 4.00 to 6.16; mean K-F frequency = 15.8 occurrences per million, range = 0 to 99; mean word length = 5.54, range = 3 - 10. Because the purpose of this experiment was to examine list strength effects using lures that were nominally unrelated to studied items, we 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 (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-casket). Also, some compound words were excluded because their constituent words were also included in the stimulus set, and an attempt was made 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 (Group One and Group Two), and each group was divided into three 50-word subgroups (Subgroups A, B, and C). Also, we took steps to ensure that the six 50-word subgroups were matched, on average, for important word characteristics.

---

4 Because \( d' \) is undefined with extreme values (i.e., hits or FA = 0 or 1), we substituted \( 0.5/N \) (where \( N \) is the number of items per condition) when hits or FAs = 0, and we substituted \( 1 - (0.5/N) \) when hits or FAs = 1 (Green & Swets, 1974; Macmillan & Creelman, 1991).

5 We also computed sensitivity based on “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 save space, \( A' \) results are not reported here.

6 Subgroups were matched for K-F frequency, familiarity, concreteness, imageability, and length. Also, we compiled remember-familiar and confidence ratings for individual items
**Design.** There were two study-test blocks: A Weak Interference block and a Strong Interference block. For half of the participants the Weak Interference block came first; vice-versa 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 vs. second block, and in the Weak Interference vs. Strong Interference block. This balancing was accomplished by having each of the two word Groups serve equally often in the Strong and Weak Interference blocks, and by having each of the three word Subgroups within each group serve equally often as Targets, Interference Items, and Lures. Combining this rotation of 2 Groups and 3 Subgroups through conditions with whether Strong or Weak Interference came first, there were $2 \times 3 \times 2 = 12$ between-subjects counterbalancing conditions.

The general structure of the Weak and Strong Interference blocks follows the description provided in the Shared Design Elements section. In Strong Interference blocks, participants studied 50 Target items once and the 50 Interference items six times. In Weak Interference 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 Strong Interference and Weak Interference blocks.

**Procedure.** Testing was done on an Apple iMac computer running PsyScope. Before the start of the experiment, participants were shown a “banker's box” (approximately 1 foot wide, 2 feet long, 1 foot deep) on the floor of the testing room. During the study phase, words appeared onscreen for 1150 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 1150 ms interval, the computer made a “beep” noise; if they failed to respond within the 1150 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 minutes in the Strong Interference condition, and 8 minutes, 53 seconds in the Weak Interference 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 1-6 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 a the box, or 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, in which they studied a short list of words (some of which were presented multiple times), played “Skittles” for 2 minutes, and were tested on the words they studied, before the start of the actual experiment. 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 only contained 1) items from the immediately preceding study phase and 2) completely new items, so, 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 based on the pooled data in Table 1) for the Weak Interference and Strong Interference 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 Weak Interference ROC than under the Strong Interference ROC (thereby indicating greater overall recognition sensitivity in the Weak Interference condition). In keeping with this claim, both $A_z$ and $d_a$ (computed based on individual participants’ data) were significantly larger in the Weak Interference condition; for $A_z$, $F(1,35) = 11.082$, $MSE = .002$; for $d_a$, $F(1,35) = 16.531$, $MSE = .120$.

The Gaussian model used to compute $A_z$ and $d_a$ was a good fit, overall, to individual participants’ data. Across the 72 model fits (36 participants X Strong/Weak Interference), the average chi-squared (3 df) 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, we need to ensure that the observed LSE for recognition sensitivity does not depend on inclusion of values from participants/conditions where the Gaussian model was a poor fit. To address this concern, we re-ran the analysis, excluding participants with large chi-squared values; we used a very liberal exclusion criterion (chi-squared p value < .05 for one of the interference conditions) to maximize the odds that participants with poor fits would be excluded. 7 11 out of 36 participants were excluded

---

7 For discussion of why alpha = .05 is a liberal value for rejecting chi-squared model fits, see, e.g., Press, Teukolsky, Vetterling, & Flannery (1992), Chapter 15.
according to this criterion, but the overall pattern of results was unchanged. Most importantly, the LSE for $A_Z$ and $d_a$ was still significant in the re-analysis.  

**Remember-familiar data**

As predicted, $d'$ computed based on "remember" hits and false alarms showed a significant LSE, $F(1,35) = 5.42$, MSE = .135. When $d'$ was computed based on "old" responses (i.e., confidence > 3), the LSE was not significant, $F(1,35) = 1.484$, MSE = .204, $p = .23$.

**Discussion**

The most important result from this experiment is our finding of a significant LSE for overall recognition sensitivity, indexed using $d_a$ and $A_Z$. However, as in other list strength studies, we also found a LSE for bias. Figure 2 shows clearly that responding was more conservative overall (fewer hits and false alarms) in the Strong Interference condition than in the Weak Interference condition (see Hirshman, 1995, for a review of data on how list strength affects bias, and 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 artefact of participants shifting their criteria -- it is a well-known fact that some signal-detection measures of sensitivity are affected by criterion shifts (Snodgrass & Corwin, 1988; Macmillan & Creelman, 1991). 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 we assume 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 like $A_Z$ and $d_a$ that depend on the shape of the normalized ROC will be invariant as a function of criterion placement (Macmillan & Creelman, 1991).8

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, where 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, like $d_a$. However, there is no evidence that the LSE observed here is an artefact 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 where responding was more conservative: slope = .69 (SEM = .07) in the Strong Interference condition and slope = .68 (SEM = .04) in the Weak Interference condition.

---

8 Hintzman (2001; see also Wickelgren, 1968) points 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 Strong (vs. Weak) Interference condition, this could result in an artefactual LSE for $A_Z$ and $d_a$. However, there is no reason to think that criterion placement would be more variable in the Strong Interference 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 that we collected in Experiment 1 provide some evidence in favor of this idea; when discrimination (d') was computed based on "remember" responses (which should reflect the contribution of recollection), there was a significant LSE; but the LSE was not significant when d' was computed based on "old" responses (which can be driven either by recollection or familiarity).

However, while these results provide converging evidence in favor of our 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 (Macmillan & Creelman, 1991; Donaldson, 1993). 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) vs. d'(Old), could both be artefacts of participants responding more conservatively in the Strong Interference condition. In order to obtain stronger evidence in favor of our dual-process hypothesis, we need to find some way of isolating the contribution of recollection that does not force us to rely on "single-point" estimates of sensitivity like d' (which, as discussed above, are hard to interpret when bias is changing and false alarms are low). This is what we 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: Intuitively, in situations where 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, and 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 we 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

9 We should point out that, while 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 Strong Interference 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' (Macmillan & Creelman, 1991; Donaldson, 1993). Thus, we can not explain the observed decrease in d' purely in terms of a criterion shift. The low overall rate of "remember" false alarms in the Strong Interference condition, which forced us to use the correction for zero false alarms more frequently in this condition than in the Weak Interference condition, is a more serious problem with respect to interpreting the LSE for d'(Remember).
studied items and nominally unrelated lures can be supported by familiarity (as well as recollection).

To test this hypothesis, we used a variant of the plurals recognition paradigm introduced by Hintzman, Curran, & Oppy (1992). Participants studied singular and plural words. At test, there were two kinds of lures: related switched-plurality (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, we should find a significant list strength effect for studied vs. SP lure discrimination, but not necessarily for studied vs. unrelated lure discrimination.

Furthermore, we can also look at SP vs. unrelated lure pseudodiscrimination, i.e., how much more likely are participants to say old to related vs. unrelated lures. Familiarity supports pseudodiscrimination (insofar as SP lures will be 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 “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 will be an increase in pseudodiscrimination. Hence, we predict a negative LSE for pseudodiscrimination (i.e., it should be higher in the Strong Interference condition than the Weak Interference 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 hour 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 word is 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, so 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 Experiment 1 Methods 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 Weak Interference block and a Strong Interference block. For half of the participants the Weak Interference block came first; vice-versa 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 where an item was presented repeatedly at study, the item was always presented as a singular word, or always presented 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 2 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, there were 40 between-subjects counterbalancing conditions: (rotate 10 word groups across conditions) X (study each item as both a singular and a plural word) X (Weak block first vs. Strong block first).

The overall structure of the Weak and Strong Interference study lists was the same as in Experiment 1; each list contained 50 Targets and 50 Interference items; Interference items were studied 1X in the Weak Interference list vs. 6X in the Strong Interference block (see the Shared Design Elements section for more details). A minor difference between this experiment and the preceding experiment is that stimuli were presented in a random order in Experiment 2 (subject to the constraints outlined in Figure 1), whereas the preceding experiment presented stimuli in a fixed order for a given counterbalancing condition.

We used a different video game (Gem Master) 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 Weak and Strong Interference conditions. The video game lasted 2 minutes in the Strong Interference condition, and 8 minutes, 53 seconds in the Weak Interference condition.

The recognition test was comprised of 25 studied target items (items presented in the same plurality at study and test), 25 switched-plurality 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 switched-plurality lures, and 5 unrelated lures. 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. We 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 an Dell Dimension computer running E-Prime software (Psychology Software Tools, Pittsburgh, PA). The study procedure was very similar to the procedure used in Experiment 1: Words appeared on screen for 1150 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: If the word is 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; if the word is 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 -- in order 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
List Strength Effects

(“no” or “yes”) within the 1150 ms interval in which an item was onscreen, the experiment was temporarily suspended and a message appeared onscreen telling the participant 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 we analyzed the data, confidence ratings were converted 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. We were interested in three different kinds of recognition discrimination: studied vs. unrelated lure discrimination (S vs. U), studied vs. switched-plurality lure discrimination (S vs. SP), and switched-plurality vs. unrelated lure pseudodiscrimination (SP vs. U).

To compute sensitivity, we generated 6 ROC curves for each participant (Strong/Weak Interference X the three different types of discrimination: S vs. U, S vs. SP, and SP vs. U); a Gaussian model was fit separately to each of these curves, and we 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, we are using Gaussian models to compute sensitivity for purely pragmatic reasons -- they provide a good fit to the data -- not because we have any commitment to the idea that the underlying distributions are Gaussian.

As predicted, there was a significant LSE for S vs. SP discrimination, indexed using both $A_z$ and $d_a$: for $A_z$, $F(1,79) = 6.076, \text{MSE} = .008$; for $d_a$, $F(1,79) = 7.125, \text{MSE} = .139$. The LSE for S vs. U discrimination was numerically smaller, and not significant: for $A_z$, $F(1,79) = 1.755, \text{MSE} = .005, p = .19$; for $d_a$, $F(1,79) = .704, \text{MSE} = .201, p = .40$. Finally, there was a significant negative LSE for SP vs. U pseudodiscrimination: for $A_z$, $F(1,79) = 21.024, \text{MSE} = .011$; for $d_a$, $F(1,79) = 18.690, \text{MSE} = .221$.

The Gaussian model used to compute $A_z$ and $d_a$ was a very good fit to individual participants' data. Across the 480 model fits (80 participants X Strong/Weak Interference X the three types of discrimination), the average chi-squared (3 df) value for the Gaussian
model was 2.69 (SD = 2.48). To address the concern that $A_z$ and $d_z$ may not accurately reflect sensitivity when the Gaussian model is a poor fit, we re-ran the analysis, excluding participants with high chi-squared values. As in Experiment 1, we 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 re-analysis were qualitatively identical to the original results: There was a significant LSE for S vs. SP discrimination, a significant negative LSE for SP vs. U pseudodiscrimination, and the LSE was not significant for S vs. U discrimination.

**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 vs. SP lure discrimination, insofar as this kind of discrimination depends on recollection. Second, there should be a negative LSE for SP vs. 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 vs. U 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 vs. U pseudodiscrimination, the rightmost point of the Strong Interference ROC is clearly "bumped up", relative to the Weak Interference ROC.

Some, but not all, of the results of this experiment are consistent with a "bias shift only" explanation: For S vs. SP discrimination, the slope and intercept of the normalized ROC were lower in the Strong Interference condition, where 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 vs. U pseudodiscrimination purely in terms of participants shifting their criteria. The Van Zandt theory predicts that adopting a more conservative criterion will result in a monotonic decrease in z-ROC slope, but for SP vs. U pseudodiscrimination, z-ROC slope was significantly higher in the "more conservative" (Strong Interference) condition: slope = .93 (SEM = .07) in the Strong Interference condition and slope = .66 (SEM = .04) in the Weak Interference condition, $F(1,79) = 12.975$, MSE = .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 this data).

The LSE for studied vs. unrelated lure discrimination (indexed using $A_z$ and $d_z$) was smaller in Experiment 2 than in Experiment 1: In Experiment 2, the mean LSE for $A_z$ was .015 (SEM = .011); in Experiment 1, the mean LSE for $A_z$ was .036 (SEM = .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 vs. unrelated lure discrimination
will be a function of exactly how much recollection (relative to familiarity) is contributing. Thus, we may be able to explain the reduced size of the LSE in terms of the idea that recollection was contributing less to studied vs. 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 you do not remember plurality, you can not be sure that you studied this exact word. It therefore stands to reason that participants would weight item recollection less heavily in Experiment 2 than in Experiment 1. Also, we 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 where familiarity also discriminates) than they would otherwise.

General Discussion

In Experiment 1, we obtained a significant list strength effect 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 list strength effect for recognition sensitivity (e.g., Ratcliff et al., 1990). To account for this novel result, we argued that increasing list strength impairs recollection of specific studied details, but 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. We obtained some suggestive evidence in favor of this dual-process hypothesis in Experiment 1, using self-report measures of recollection (“remember” responses): Discrimination computed based on “remember” responses showed a significant list strength effect, but the LSE was not significant when discrimination was computed based on “old” responses (which can be driven by either familiarity or recollection). In Experiment 2, we 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 vs. 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, we used a plurality discrimination paradigm in which participants were given related, switched-plurality (SP) lures as well as unrelated lures at test. The results were exactly as we had predicted: There was a significant LSE for studied vs. SP lure discrimination (indexed using $A_z$ and $d_a$), and there was a significant negative LSE for SP vs. unrelated lure pseudodiscrimination, but the LSE for studied vs. unrelated lure discrimination was not significant.

Relation to other list strength studies

In the experiments reported here, we replicated the null list strength effect that other studies have found, when single-point estimates of recognition sensitivity (e.g., $d'$) are computed based on whether participants think the item is “old” or “new”, and 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 based on the probability of calling an item “old” (i.e., assigning a confidence
The fact that we replicated the null LSE for old-new recognition sensitivity, despite our 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 based on “old” responses) may fail to detect sensitivity differences that are revealed by other, multipoint measures of sensitivity – while the LSE for \( d'(\text{Old}) \) 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 switched-plurality lures we used in Experiment 2). Thus, the fact that we found list strength effects for discrimination using highly similar lures does not directly contradict extant results. Other studies have used related lures (e.g., Shiffrin, Huber, & Marinelli, 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 we did in Experiment 1, Ratcliff et al. (in their Experiments 1 through 5) used nominally unrelated lures, manipulated list strength, collected confidence ratings at test, and plotted ROC curves based on this confidence data. However, in contrast to our Experiment 1 results, the results presented by Ratcliff et al. do not show any clear evidence of a LSE for recognition sensitivity.

At this point, we can only speculate as to why our Experiment 1 found a LSE for recognition sensitivity, but Ratcliff et al.’s experiments did not. In our experiments, we took several steps to boost the size of the LSE: Our experiments used a powerful strength manipulation: Interference items were presented six times in the Strong Interference condition, vs. once in the Weak Interference condition. Also, we 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. 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 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 mentioned earlier, it is possible that the "remember-familiar" instructions that we 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, we should note that our dual process hypothesis clearly predicts a LSE for

\[ d'(\text{Old}) \text{ for S vs. U discrimination} = 1.82 \ (SEM = .08) \text{ in the Weak Interference condition and 1.90 \ (SEM = .07) in the Strong Interference condition, } F(1,79) = .974, \text{ MSE} = .250, p = .33. \]
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). We 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 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 at the same time ensuring some degree of trace overlap) should bolster the LSE for cued recall, relative to a condition where participants are just told 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 discussion of the issue). 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 from 1990 to the present to modify global matching models so 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 list strength effects will always be obtained, the results reported here pose problems for recognition models that predict list strength effects will never be obtained for item recognition sensitivity (e.g., Murdock & Kahana, 1993; Dennis & Humphreys, 2001). Murdock and Kahana (1993) argue that items from the current study list make a negligibly small contribution to the variance of the global match signal, relative to the contribution all of the other items that have been studied (over the person's lifetime); thus, strengthening items from the current list will not boost variance enough to hurt recognition. Dennis and Humphreys (2001) argue 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, non-strengthened list items. These two models, in their present form, can not accommodate the significant list strength effects for
recognition sensitivity reported here, except by arguing that list strength is confounded with some other factor. For example, Dennis and Humphreys (2001) argue that -- in retroactive interference designs -- list length and strength effects could arise spuriously if participants mentally focus on the latter part of the Strong Interference 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 where 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 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., 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, e.g., our finding in Experiment 2 of a LSE for S vs. SP discrimination, coupled with a negative LSE for SP vs. U pseudodiscrimination.

At present, the only model of recognition that has been shown to fit our data is the dual-process Complementary Learning Systems (CLS) neural network model (Norman & O’Reilly, under revision). Unlike single-process models like 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 paper -- see Norman & O’Reilly (under revision) for a detailed discussion of the model’s list strength predictions.

Conclusions

The experiments reported here show that list strength effects 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 when list strength effects will be obtained. Future research will explore the boundary conditions of the list strength effects reported here (i.e., why were they obtained here but not in other studies such as Ratcliff et al., 1994). Also, although the results reported here are certainly consistent with our 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


Author Note

This research was supported by NIH NRSA grant MH12582-01. I thank Tim Curran, Lew Harvey, Doug Hintzman, Elliot Hirshman, David Huber, and two anonymous reviewers for commenting on this manuscript, 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 running participants in Experiments 1 and 2, respectively.

Correspondence concerning this article should be sent to Dr. Kenneth A. Norman, who is now located at Princeton University, Department of Psychology, Princeton, NJ, 08544. Electronic mail may be sent to knorman@princeton.edu.
Table 1

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

<table>
<thead>
<tr>
<th>Condition</th>
<th>&gt;1</th>
<th>&gt;2</th>
<th>&gt;3</th>
<th>&gt;4</th>
<th>&gt;5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak Interference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Studied Old</td>
<td>.98 (.00)</td>
<td>.95 (.01)</td>
<td>.91 (.01)</td>
<td>.83 (.02)</td>
<td>.70 (.03)</td>
</tr>
<tr>
<td>Studied Remember</td>
<td>N/A</td>
<td>N/A</td>
<td>.73 (.03)</td>
<td>.72 (.03)</td>
<td>.67 (.03)</td>
</tr>
<tr>
<td>Studied Familiar</td>
<td>N/A</td>
<td>N/A</td>
<td>.18 (.03)</td>
<td>.12 (.02)</td>
<td>.04 (.02)</td>
</tr>
<tr>
<td>Nonstudied Old</td>
<td>.71 (.03)</td>
<td>.40 (.03)</td>
<td>.22 (.02)</td>
<td>.12 (.02)</td>
<td>.05 (.01)</td>
</tr>
<tr>
<td>Nonstudied Remember</td>
<td>N/A</td>
<td>N/A</td>
<td>.05 (.01)</td>
<td>.05 (.01)</td>
<td>.04 (.01)</td>
</tr>
<tr>
<td>Nonstudied Familiar</td>
<td>N/A</td>
<td>N/A</td>
<td>.16 (.02)</td>
<td>.07 (.01)</td>
<td>.01 (.00)</td>
</tr>
<tr>
<td>Strong Interference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Studied Old</td>
<td>.92 (.02)</td>
<td>.84 (.02)</td>
<td>.77 (.02)</td>
<td>.66 (.03)</td>
<td>.50 (.03)</td>
</tr>
<tr>
<td>Studied Remember</td>
<td>N/A</td>
<td>N/A</td>
<td>.55 (.03)</td>
<td>.54 (.03)</td>
<td>.47 (.03)</td>
</tr>
<tr>
<td>Studied Familiar</td>
<td>N/A</td>
<td>N/A</td>
<td>.23 (.02)</td>
<td>.13 (.02)</td>
<td>.03 (.01)</td>
</tr>
<tr>
<td>Nonstudied Old</td>
<td>.55 (.04)</td>
<td>.26 (.03)</td>
<td>.11 (.02)</td>
<td>.05 (.01)</td>
<td>.02 (.01)</td>
</tr>
<tr>
<td>Nonstudied Remember</td>
<td>N/A</td>
<td>N/A</td>
<td>.02 (.00)</td>
<td>.02 (.00)</td>
<td>.01 (.00)</td>
</tr>
<tr>
<td>Nonstudied Familiar</td>
<td>N/A</td>
<td>N/A</td>
<td>.09 (.01)</td>
<td>.03 (.01)</td>
<td>.00 (.00)</td>
</tr>
</tbody>
</table>

Note. Standard errors in parentheses.
Table 2

Derived Sensitivity Measures for Experiment 1

<table>
<thead>
<tr>
<th>Measure</th>
<th>Weak Interference</th>
<th>Strong Interference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_z$</td>
<td>.93 (.01)</td>
<td>.89 (.01) **</td>
</tr>
<tr>
<td>$d_a$</td>
<td>2.19 (.08)</td>
<td>1.86 (.09) **</td>
</tr>
<tr>
<td>$d'$ (Remember)</td>
<td>2.44 (.09)</td>
<td>2.24 (.08) **</td>
</tr>
<tr>
<td>$d'$ (Old)</td>
<td>2.35 (.12)</td>
<td>2.22 (.10)</td>
</tr>
</tbody>
</table>

Note. Standard errors in parentheses. Two asterisks indicate a significant list strength effect at $p < .05$, two-tailed.
Table 3

Counterbalancing Conditions for Experiment 2

<table>
<thead>
<tr>
<th>Block</th>
<th>Study As</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>Target</td>
<td>Same plurality</td>
</tr>
<tr>
<td>First</td>
<td>Target</td>
<td>Switched plurality</td>
</tr>
<tr>
<td>First</td>
<td>Interference</td>
<td>n/a</td>
</tr>
<tr>
<td>First</td>
<td>Interference</td>
<td>n/a</td>
</tr>
<tr>
<td>First</td>
<td>n/a</td>
<td>Unrelated lure</td>
</tr>
<tr>
<td>Second</td>
<td>Target</td>
<td>Same plurality</td>
</tr>
<tr>
<td>Second</td>
<td>Target</td>
<td>Switched plurality</td>
</tr>
<tr>
<td>Second</td>
<td>Interference</td>
<td>n/a</td>
</tr>
<tr>
<td>Second</td>
<td>Interference</td>
<td>n/a</td>
</tr>
<tr>
<td>Second</td>
<td>n/a</td>
<td>Unrelated lure</td>
</tr>
</tbody>
</table>

Note. Word groups were rotated across these 10 conditions such that each group served equally often in each of the conditions.
Table 4

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

<table>
<thead>
<tr>
<th>Condition</th>
<th>&gt;1</th>
<th>&gt;2</th>
<th>&gt;3</th>
<th>&gt;4</th>
<th>&gt;5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Weak Interference</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Studied</td>
<td>.92 (.01)</td>
<td>.82 (.01)</td>
<td>.77 (.01)</td>
<td>.73 (.02)</td>
<td>.55 (.02)</td>
</tr>
<tr>
<td>Switched-Plurality</td>
<td>.72 (.02)</td>
<td>.52 (.02)</td>
<td>.45 (.02)</td>
<td>.40 (.02)</td>
<td>.24 (.02)</td>
</tr>
<tr>
<td>Unrelated</td>
<td>.65 (.03)</td>
<td>.33 (.02)</td>
<td>.18 (.01)</td>
<td>.13 (.01)</td>
<td>.05 (.01)</td>
</tr>
<tr>
<td><strong>Strong Interference</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Studied</td>
<td>.82 (.02)</td>
<td>.70 (.02)</td>
<td>.64 (.02)</td>
<td>.60 (.02)</td>
<td>.38 (.02)</td>
</tr>
<tr>
<td>Switched-Plurality</td>
<td>.64 (.02)</td>
<td>.43 (.02)</td>
<td>.36 (.02)</td>
<td>.30 (.02)</td>
<td>.14 (.01)</td>
</tr>
<tr>
<td>Unrelated</td>
<td>.41 (.03)</td>
<td>.19 (.02)</td>
<td>.09 (.01)</td>
<td>.06 (.01)</td>
<td>.02 (.00)</td>
</tr>
</tbody>
</table>

**Note.** Standard errors in parentheses.
Table 5

Derived Sensitivity Measures for Experiment 2

<table>
<thead>
<tr>
<th>Measure</th>
<th>Weak Interference</th>
<th>Strong Interference</th>
</tr>
</thead>
<tbody>
<tr>
<td>S vs. SP $A_2$</td>
<td>.72 (.01),</td>
<td>.68 (.01) **</td>
</tr>
<tr>
<td>S vs. U $A_2$</td>
<td>.85 (.01),</td>
<td>.83 (.01)</td>
</tr>
<tr>
<td>SP vs. U $A_2$</td>
<td>.61 (.02)</td>
<td>.69 (.02) **</td>
</tr>
<tr>
<td>S vs. SP $d_a$</td>
<td>.87 (.06)</td>
<td>.71 (.06) **</td>
</tr>
<tr>
<td>S vs. U $d_a$</td>
<td>1.55 (.06)</td>
<td>1.49 (.07)</td>
</tr>
<tr>
<td>SP vs. U $d_a$</td>
<td>.44 (.07)</td>
<td>.76 (.07) **</td>
</tr>
</tbody>
</table>

Note. Standard errors in parentheses. Two asterisks indicate a significant list strength effect at $p < .05$, two-tailed.
List Strength Effects

Figure Captions

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.

Figure 2. ROC curves (plotting hits vs. false alarms for different confidence criteria) for the Weak Interference and Strong Interference conditions of Experiment 1. These ROC curves are plotted based on the pooled data contained in Table 1. The area under the ROC is larger in the Weak Interference condition than the Strong Interference condition, indicating that sensitivity is higher in the Weak Interference condition.

Figure 3. ROC curves for Experiment 2, looking at studied vs. unrelated lure (S vs. U) discrimination, studied vs. switched-plurality lure (S vs. SP) discrimination, and switched-plurality vs. unrelated lure (SP vs. U) pseudodiscrimination, as a function of Interference Strength. These ROC curves are plotted based on the pooled data contained in Table 4.
Stimuli:
Targets (T)
Lures (L)
Interference items (INT)

<table>
<thead>
<tr>
<th>Phase 1</th>
<th>Strong Interf.</th>
<th>Weak Interf.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T mixed with INT</td>
<td>T mixed with INT</td>
</tr>
<tr>
<td></td>
<td>↓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>INT</td>
<td>Play video game</td>
</tr>
<tr>
<td></td>
<td>INT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>INT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>INT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>INT</td>
<td></td>
</tr>
<tr>
<td>Phase 2</td>
<td>Play video game for 2 minutes</td>
<td>Play video game for 2 minutes</td>
</tr>
<tr>
<td>Phase 2a</td>
<td>Recognition test: T mixed with L</td>
<td>Recognition test: T mixed with L</td>
</tr>
<tr>
<td>Phase 3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiment 2 ROC Curves

S vs. U
False Alarms

S vs. SP
False Alarms

SP vs. U
False Alarms

Hits

- Weak Interference
- Strong Interference