Using Recall to Reduce False Recognition: Diagnostic and Disqualifying Monitoring

David A. Gallo
Harvard University

Whether recall of studied words (e.g., parsley, rosemary, thyme) could reduce false recognition of related lures (e.g., basil) was investigated. Subjects studied words from several categories for a final recognition memory test. Half of the subjects were given standard test instructions, and half were instructed to use recall to reduce false recognition. Manipulation checks indicated that the latter instructions did elicit a recall-to-reject strategy. However, false recognition was selectively reduced only when all the words from a category could be recalled (Experiment 1). When longer categories were used, thereby minimizing exhaustive recall, a recall-to-reject strategy was ineffective at reducing false recognition (Experiment 2).

It is suggested that exhaustively recalling a category allowed subjects to disqualify the lure as having occurred, analogous to recall-to-reject demonstrations in other tasks. In contrast, partially recalling a category did not help to diagnose the lure as nonstudied. These findings constrain theories of recall-based monitoring processes.

In the last chapter of *Elements of Episodic Memory*, Endel Tulving (1983) reported a curious phenomenon that was inconsistent with the encoding specificity hypothesis. The phenomenon was observed in a "direct comparison" experiment (p. 303), whereby the probability of recalling a studied word, given a nonstudied associate as a cue, was compared with the probability of falsely recognizing that same nonstudied associate on a recognition test. The encoding specificity hypothesis predicts that these two probabilities should be positively correlated. Nonstudied words that have more overlap (e.g., many common features) with their studied associates should be more likely to cue those associates on a recall test, and also more likely to elicit familiarity-based false alarms on a recognition test. Contrary to expectation—and the encoding specificity hypothesis—a nearly perfect negative correlation was obtained.

To account for this result, Tulving (1983) proposed what has subsequently been called a *recall-to-reject process*. If the recognition lure reminds the rememberer of its associated studied item, then "the rememberer can decide not to call the retrieval cue ‘old’ even if it looks ‘familiar’" (p. 317). Similar appeals to recall-to-reject processes in associatively based false recognition tasks can be found elsewhere in the literature (e.g., Brainerd, Reyna, & Kneer, 1995; Hall & Kozloff, 1970), but the best way to interpret these findings is not without question (Wallace, Malone, Swiergosz, & Amberg, 2000). Brainerd, Reyna, Wright, and Mojardin (in press) have recently reviewed the literature on recall-to-reject processes, couching these phenomena within the theoretical framework of fuzzy-trace theory. For the present investigation I took a more general approach, proposing two ways that recall might reduce false recognition: disqualifying and diagnostic recall-to-reject. Both processes rely on the strategic use of recalled information to monitor or edit memory accuracy at retrieval, but they depend on qualitatively different decision processes.

Disqualifying recall-to-reject occurs when the subject recalls certain information that, because of their knowledge of the situation, eliminates the questionable event as having occurred. For this reason, a disqualifying recall-to-reject strategy is rule-based and is highly dependent on whether the structure of the encoding context (i.e., the study task) permits the application of such a rule. In Tulving’s (1983) direct comparison experiment, the subject might recall that *table* was studied when given *chair* as a lure. If the subject believed that only one associate could have been presented, he or she could have disqualified *chair* as a potential studied item. This type of disqualifying decision process is involved in many (if not all) of the experimental demonstrations of recall-to-reject. For example, in the list-based exclusion task of Jacoby and colleagues (e.g., Jacoby, 1991; McElree, Dolan, & Jacoby, 1999), subjects can reject a familiar word as having been presented in a target list by recalling that it was actually presented in a nontarget list. Other examples are the associative recognition task (e.g., Doshier, 1984; Rotello, Macmillan, & Van Tassel, 2000) and the changed pluralization task (e.g., Hintzman & Curran, 1994; Rotello et al., 2000). In both cases, the subject can reject a familiar lure (e.g., apple—crown or computers) by recalling the corresponding target that made it familiar (e.g., apple—dog or computer).

Diagnostic recall-to-reject occurs when the subject recalls information that suggests, or is diagnostic, that a lure did not occur. This recall-to-reject process is *heuristic-based*, as the term is used in the source-monitoring framework (e.g., Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981). With this process, the subject recalls studied items and compares the memorial evidence for the lure to that for the studied items. If the memorial evidence for the lure is not as compelling as that for the studied items, then...
the subject is less likely to respond “old” to the lure (for related ideas, see Ghetti, 2003; Strack & Bless, 1994; and Israel & Schacter, 1997). Returning to Tulving (1983), the subject might have thought that both table and chair could have been in the study list (i.e., their occurrence was not mutually exclusive), so that a rule-based rejection strategy would not apply. Nevertheless, if they were able to clearly recall table, but chair was only vaguely familiar, then they might reason that chair probably was not studied (otherwise, it too should be clearly recalled). With diagnostic recall-to-reject, recalled information helps to develop expectations about what should be recalled for studied events, which facilitates the rejection of nonstudied events. Of course, diagnostic monitoring could be based on factors other than the recall of studied events, such as plausibility, but the present question is how recall might be involved.

The goal of this study was to investigate whether these recall-to-reject processes could reduce false recognition in a converging associates task, such as those described by Roediger and McDermott (1995) and Shiffrin, Huber, and Marinelli (1995). Subjects in these tasks study multiple associates (e.g., category exemplars) that converge on a nonstudied lure (a related exemplar). These related lures are often mistakenly recognized on a final test, and they frequently evoke illusory recollection (or false recall) of the lure. Associated lures are often falsely recognized on a final test, and they converge on a nonstudied lure (a related exemplar). These findings are in contrast to the feeling of familiarity that is thought to drive false alarms in the aforementioned exclusion tasks (e.g., associative recognition). This is an important difference, because it is unknown whether a recall-to-reject process would be useful when lures elicit false recall. A successful recall-to-reject process necessarily presupposes that recall should be accurate, so that if the lure is also “recalled” (or falsely recollected), then a recall-to-reject process might be difficult or impossible.

Although several studies using converging-associates tasks have highlighted diagnostic monitoring processes in general (e.g., Ghetti, 2003; Schacter et al., 1999), none have directly investigated how recall per se might serve as a diagnostic tool. However, recent work suggests that a recall-to-reject process may be involved in such tasks. Roediger, Watson, McDermott, and Gallo (2001) used multiple regression analysis to investigate the variability across different associative lists in eliciting false recall and recognition. They found that most of the variance was accounted for by associative strength, or the average degree of association between the list items and the related lure. More interesting for present concerns, they also found a strong negative relationship between true recall and false remembering. Lists that elicited greater recall of studied items yielded lower levels of false recall and false recognition.

As a potential explanation, Roediger et al. (2001) proposed that monitoring processes occurring during the test caused this effect: “Subjects may use item-specific information to aid recall of list items and suppress recall of critical items, which would not carry as much item-specific information” (p. 394). That is, more detailed recollections for studied items than for the lure would be diagnostic that the lure was not studied. Similar appeals to recall-based processes, or some variant of dual-process theories, have become increasingly popular to explain otherwise puzzling reductions in false memories from converging-associates tasks (e.g., Benjamin, 2001; Brainerd et al., in press; McDermott & Watson, 2001). For instance, Benjamin (2001) found that repeating studied lists reduced false recognition on self-paced tests, even though repetition was shown to have increased the familiarity of the lures (measured by a speeded test). Repetition enhanced memory for the list items, and this apparently facilitated monitoring processes that reduced false recognition on the self-paced test. However, little is known about how these monitoring processes work and how recall or recollection might be involved, and in none of these cases was the actual use of recall during a recognition test directly manipulated and measured.

In a preliminary experiment to those reported here, subjects studied five words from 40 categories (e.g., parsley, rosemary, thyme), with words mixed by category during study. They then received a recognition memory test. The critical manipulation was whether category labels (e.g., herbs) were presented as cues during study and/or during the recognition test, along with the test item. There were four conditions (YY, NN, YN, NY) with the first and second letters indicating whether labels were presented at study or test, respectively. This cuing manipulation was designed to influence the overall levels of true recall during the recognition test. As expected, category labels did affect cued recall on a separate recall test (e.g., Tulving & Pearlstone, 1966; Epstein, Dupree, & Gronkowski, 1979). Presenting labels at study and test yielded the greatest levels of recall (YY = .36), never presenting labels yielded the lowest levels of recall (NN = .25), and the other two conditions resulted in intermediate levels of recall (YN = .33, NY = .28). In contrast, presenting category labels during the recognition test did not influence false recognition of related lures (e.g., basil), as the overall probability of falsely recognizing critical lures was roughly equivalent across the four conditions (YY = .24, YN = .27, NY = .26, NN = .25). Thus, although there were significant recall differences across conditions, false recognition did not vary, providing no support for a recall-to-reject process. The two experiments reported here were designed as even stronger tests of the recall-to-reject hypothesis.

Experiment 1

In the preliminary experiment, it could be argued that subjects did not realize how recall could be useful during the recognition test. If familiarity-based responding is faster and less effortful than recall-based responding, one might predict that the default mode is to not bother with an elaborate recall-to-reject strategy (cf. Jacoby & Hollingshead, 1990). To address these concerns, Experiment 1 was designed to increase the use of recall on the recognition test. A recall-to-reject strategy was explained to some subjects just before the test, and they were instructed to use this strategy to help reject nonstudied items (the strategic condition). To make the importance of this strategy even more salient, they were asked after each recognition decision if they were able to recall all of the items from that category to help make the decision. A comparison group was not given special instructions and was not asked to make recall-based judgments (the standard condition). Research in other tasks has revealed that making helpful information salient at test can enhance source-monitoring processes (e.g., Dodson & Johnson, 1993; Marsh & Hicks, 1998; Multhaup, 1995). Furthermore, Rotello et al. (2000) found that encouraging subjects to use a recall-to-reject strategy increased estimates of this process (measured via receiver operating characteristic curves) in the changed pluralization task. A similar manipulation might facilitate a recall-to-reject process in a converging associates task.
A secondary goal of Experiment 1 was to add a variable that should selectively affect a disqualifying recall-to-reject process. In this task, recalling a few studied words from a category might contribute to a diagnostic monitoring process, by fostering expectations about the memorability of words that did occur (e.g., “How easily do words from this category come to mind?”). However, in order to disqualify a lure as having occurred, the subject would need to exhaustively recall all of the words that were presented in that category. Only then could they be certain (by design) that the related lure could not have been studied in that category. To test this hypothesis, the structure of the study categories was manipulated (i.e., whether their length was held constant or was varied). It has been shown that subjects are not perfect at estimating the number of items that had been presented in a category after studying categories of varying length (Greene, 1989). Thus, a disqualifying recall-to-reject process should be more difficult when category length varies, because subjects could not be certain whether they had exhaustively recalled all of the items from a category. This manipulation should not affect a diagnostic recall-to-reject process, because that process is based on expected memorability. Such expectations could be developed even when the subject does not know whether he or she has recalled all of the items from a category.

In addition to these manipulations a few other changes were made to increase the likelihood of successful recall during the recognition test (relative to the preliminary experiment). First, study labels were presented at both study and test, which was the condition of a preliminary experiment that elicited the highest levels of recall and the fewest cross-category intrusions. Second, words were blocked by categories at study instead of mixed, which should have enhanced organizational factors that are beneficial to recall (e.g., Bower, 1970). Finally, only three exemplars were presented in each category, which should have made exhaustive recall easier.

Method

Subjects. Subjects were 96 undergraduates at Washington University who received course credit or $10. Twenty-four subjects were arbitrarily assigned to each of the four conditions. Data from 1 subject were replaced because that person reported mistakenly switching the old and new buttons at some point in the task.

Materials. The 30 categories that elicited the highest probability of false recognition in the preliminary experiment were chosen to be critical categories for the present experiment. These categories were drawn from various norms (e.g., Battig & Montague, 1969), and with few exceptions, each category consisted of those exemplars that had the strongest associative connections to a critical nonstudied exemplar (using the Nelson, McEvoy, & Schreiber [1999] word-association norms). Subjects studied 20 of the 30 critical categories, and the remaining 10 categories were used for nonstudied control conditions (i.e., base rate false alarms). Critical targets were chosen so that subjects would be unlikely to guess which items were critical targets and critical lures, and in most cases these items were listed most frequently in category norms other than the critical lures. Critical targets and critical lures did not differ on word frequency (mean count per million = 50.3 vs. 67.6, respectively, t(27) < 1; two pairs were excluded from this analysis because a member was not listed in Kučera & Francis [1967]). Subjects also studied 20 noncritical categories, with lengths that were either the same as critical categories or varied, as described later.

Categories were counterbalanced so that each critical category served as studied or nonstudied equally often across subjects, and so that each critical category contributed a critical target or critical lure to the test equally often across subjects. In the constant length conditions, three items were studied from every category (critical and noncritical). In the varied length conditions, three items were studied from the 20 critical categories, but two items were studied for half of the noncritical categories, and four items were studied from the other half. In this way, the length of all critical categories (three items) was held constant across all conditions, as was the total number of studied categories (40) and the total number of studied items (120). Only the length of the noncritical categories differed between the constant and varied conditions (data from the noncritical categories are not reported).

Procedure. Subjects studied the list for an unspecified memory test and were told that category labels would occur before each set of categorized items to help their memory. Category labels were presented on the computer screen in an uppercase red font for 1,500 ms, followed by a blank screen (250 ms). Studied words from that category were then presented in a lowercase black font for 750 ms each, again separated by a blank screen (250 ms). Categories were presented in a different random order for each counterbalancing condition. Targets occurred in position three for critical categories and in positions one and two for noncritical categories. This manipulation was done to minimize any systematic relationship between targets and their presentation order from the subject’s perspective.

After the study period, subjects completed multiplication problems for 5 min as a filler task. They were then given a self-paced recognition test. The test consisted of 50 randomly presented items, presented in lowercase red letters. Each item was paired with its corresponding category label, presented above the item in uppercase red letters. Ten items were studied exemplars from critical categories (critical targets), and 10 were nonstudied exemplars from the remaining critical studied categories (critical lures). Similarly, 10 items were noncritical targets, and 10 were noncritical lures. The remaining 10 lures were from nonstudied categories (5 critical target controls and 5 critical lure controls). Subjects were told that only one item would be tested per category. In this way they could be sure that any items that they recalled came from the study phase and not from an earlier part of the test phase.

In all conditions, subjects were told that the test consisted of studied (old) and nonstudied (new) words, and that the new words came from both studied and nonstudied categories. They were to press old only for items that were studied, regardless of whether they remembered studying a category, and that the category labels were presented to help their memories. All subjects were told to take their time and to be as accurate as possible. Test instructions and procedures in the standard conditions (with constant or varied category length) did not provide any additional information as to how false recognition could be minimized. Additional directions were given in the strategic conditions to elicit a recall-to-reject strategy.

Subjects in the strategic-constant condition were told that each category contained exactly three words, whereas subjects in the strategic-varied condition were told that the number of words from any individual category was randomly varied from two to four. Both groups were asked to use a recall-to-reject strategy. They were told that if they thought a category was studied, they should try to recall all of the words from that category. If they could recall all of the words other than the test word, then they could be sure that the test word was new. In the constant-length instructions, it was stressed that recalling three words ensured that the test word was new. In the varied-length instructions, subjects were told that they may sometimes be unsure as to how many words were studied in a category but that they should still try to recall all of the words to enhance accuracy. To further encourage both groups to use such a strategy, after each old–new decision subjects were prompted to indicate (by pressing yes or no) whether they had recalled all of the items from that category to help make their old–new decision (a recall-all judgment).

Following the recognition test, all subjects completed a questionnaire asking if they had used recall to help make their recognition decisions (taking approximately 2 min). They then took the cue-recall test. The previous test items and their labels were re-presented to subjects on the
computer. For every test item, they were to recall all the words from that category that they were reasonably sure had been presented in the study phase, and to write these on their response sheet. If they could not recall any words, or did not think that any words had been presented in that category, they were to write none on the corresponding line.

Results and Discussion

There was converging evidence that the instructions were effective at eliciting a recall-to-reject strategy. First, on the questionnaire, subjects in the strategic conditions were more likely to report having used such a strategy \((n = 44)\) than were subjects in the standard conditions \((n = 27)\). Second, recognition response latencies in strategic conditions \((\text{overall} \, M = 4.628)\) were slower than those in standard conditions \((2.249)\), and this difference was significant for every item type \((\text{all } p < .01)\). This finding suggests that subjects in the strategic conditions took more time to recall studied items during the recognition test. The third piece of evidence is based on the recall-all judgments that were made during the recognition test. These judgments can be considered an index of whether subjects were able to recall words from a category on the recognition test. If subjects were using recall during the recognition test, then positive recall-all judgments should have been associated with greater recall of studied items (as measured on the subsequent cued-recall test). This result was obtained. For the strategic-constant condition, recall was greater for categories that elicited positive recall-all judgments \((M = .72, \text{collapsed across cue types})\) than for those that elicited negative recall-all judgments \((.24), \text{and similarly for the strategic-varied condition }(.61 \text{ vs. } .15), \text{both } p < .05.\)

Note: recall was not perfect (100%) for categories that subjects gave positive recall-all judgments, indicating that subjects overestimated their ability to exhaustively recall the categories and/or that giving the recall test after the recognition test lowered recall.

With respect to recall-all judgments, an additional finding of interest was that subjects in the strategic-constant condition were more accurate than were subjects in the strategic-varied condition. Consider responses to critical lures, which are the most relevant for a recall-to-reject hypothesis. Subjects in these two conditions were equally likely to believe that they had recalled all of the items from a category (mean proportions of positive recall-all judgments were .30 and .29, respectively). However, comparison of the actual recall from these categories indicated that recall was greater in the constant condition than in the varied condition \((.74 \text{ vs. } .58, \text{respectively}; t(36) = 2.29, SEM = .072, p < .05)\). This pattern indicates that subjects in the varied condition were more likely to overestimate their ability to recall all of the items from a category. Thus, varying the length of some categories was effective at making it more difficult for subjects to know whether they had recalled all of the items from a category. This manipulation lowered recall and reduced the accuracy of recall-all judgments.

Recognition. Recognition data are presented in the top half of Table 1. Overall, false alarms to critical lures \((M = .34)\) were more frequent than false alarms to critical-lure controls \((M = .11)\), \(t(95) = 10.51, SEM = .022, p < .05\), demonstrating the associative false-recognition effect. These effects were similar in size to those obtained by others using similar categorized materials \((e.g., \text{Dewhurst, 2001; Dewhurst & Anderson, 1999; Hintzman, 1988})\). It has been argued that looking at the difference score between false alarms to critical lures and their control items provides a more purified measure of associatively based false recognition \((e.g., \text{Galloy, Roediger, & McDermott, 2001; see also Schacter, Verfaellie, & Pradere, 1996; Seamon, Luo, & Gallo, 1998})\). This difference defines that portion of the false-recognition effect that is due to having studied related items, as opposed to other factors that may have had a similar influence on related and unrelated false alarms. Adjusted recognition scores are found in the Difference rows of Table 1.

Turning to the constant-length conditions, a 2 (target vs. lure) × 2 (standard vs. strategic) analysis of variance (ANOVA) on adjusted data revealed a main effect of item type, \(F(1, 46) = 8.168, \text{MSE} = .03, p < .05\). There was no effect of instruction, and there was a marginal interaction between item type and instruction, \(F(1, 46) = 3.52, \text{MSE} = .03, p = .07\). Six additional subjects were tested in each condition, and with this additional power the interaction was significant, \(F(1, 58) = 4.51, \text{MSE} = .031, p < .05\). Follow-up \(t\) tests indicated that true recognition did not differ across conditions \((\text{Ms} = .56 \text{ for the standard condition and } .59 \text{ for the strategic condition, } t(58) < 1), \text{whereas instructing subjects to use a recall-to-reject strategy decreased false recognition } (\text{Ms} = .27 \text{ and } .16, \text{respectively}, t(58) = 2.02, SEM = .110, p < .05).\)

A similar ANOVA on data from the varied-length conditions revealed that the careful categorization of the critical targets and lures controlled for any practice effects that may have been affected by the manipulation of the constant-length condition. The varied-length condition was not used because the recall-all response keys were mislabeled.

Table 1

<table>
<thead>
<tr>
<th>Item-type</th>
<th>Standard length</th>
<th>Strategic length</th>
<th>Varied length</th>
<th>Standard length</th>
<th>Strategic length</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td>.68</td>
<td>.74</td>
<td>.72</td>
<td>.63</td>
<td></td>
</tr>
<tr>
<td>CT controls</td>
<td>.11</td>
<td>.15</td>
<td>.14</td>
<td>.12</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>.57</td>
<td>.59</td>
<td>.58</td>
<td>.51</td>
<td></td>
</tr>
<tr>
<td>CL</td>
<td>.40</td>
<td>.53</td>
<td>.29</td>
<td>.35</td>
<td></td>
</tr>
<tr>
<td>CL controls</td>
<td>.10</td>
<td>.15</td>
<td>.07</td>
<td>.13</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>.30</td>
<td>.18</td>
<td>.22</td>
<td>.22</td>
<td></td>
</tr>
</tbody>
</table>

Note. Critical categories contained three items in Experiment 1 and five items in Experiment 2. CT = critical targets; CL = critical lures.

1 The data from some subjects were excluded from this analysis because they never made a positive or negative recall-all judgment for some item types \((n = 3 \text{ in the constant-length condition, and } n = 6 \text{ in the varied length condition})\). Also, data from 1 subject in the constant-length condition were not included because her recall-all response keys were mislabeled.

2 The same pattern was obtained on unadjusted recognition scores. True recognition did not differ between the standard \((M = .69)\) and strategic \((M = .74)\) conditions, \(p > .10\), but false recognition was lower in the latter \((\text{Ms} = .42 \text{ and } .32), t(58) = 2.05, SEM = .097, p < .05\). The addition of these 12 subjects did not change the overall pattern of results or conclusions on any other analyses reported here and for consistency the remaining analyses are based on the original data set.
vealed a main effect of item type, $F(1, 46) = 64.83$, $MSE = .04$, $p < .05$. There was no main effect of instruction and no interaction (both $F$s < 1), indicating that instructing subjects to use a recall-to-reject strategy had no effect when category length was varied.

**Cued recall.** Cued recall of studied items did not differ across the two constant length conditions ($M = .35$). A 2 (target cue vs. lure cue) $\times$ 2 (standard vs. strategic) ANOVA on cued-recall performance from these conditions confirmed that there were no main effects or interactions (all $F$s < 1). A similar ANOVA for the varied-length conditions revealed a main effect of instruction, $F(1, 46) = 11.18$, $MSE = .055$, and no effect of item type or interaction (both $p$s > .10). The effect of instruction indicates that recall of studied items was greater in the standard condition ($M = .52$, collapsed across cue type) than the strategic condition (.36). Telling the strategic subjects that it would be difficult for them to know when they had recalled all of the items in a category (because length varied) reduced their recall performance. As previously discussed with respect to the recall-all judgments, these subjects may have thought that they had recalled all of the items from a category when in fact they had not. Finally, when critical targets were given as cues, false recall of the critical lures ($M = .11$, averaged across conditions) was significantly greater than false recall of these same lures when their corresponding categories had not been studied (i.e., when critical target controls were cues, $M = .05$, $t(95) = 4.30$, $SEM = .014$). This finding demonstrates the associative false-recall effect.

**Recall and recognition.** If true recall had influenced false recognition via a recall-to-reject process, then correct rejections of the critical lures should be associated with greater recall than false alarms. That is, there should be a negative dependency between true recall and false recognition, as suggested by Tulving’s (1983) analysis. To this end, recall data were calculated separately for critical targets and lures that were recognized and those that were not, and these data are reported in the top half of Table 2.

As a further approach, recognition data were calculated separately for those categories from which subjects had recalled all of the items (exhaustive recall) and for those from which they only recalled a few of the items (partial recall). This calculation was done for each subject, and the means for each condition can be found in Figure 1. As discussed earlier, the total number of categories that elicited partial or exhaustive recall differed across the conditions, and some subjects had to be excluded from the statistical analyses because they never exhaustively recalled the items from any category. Nevertheless, the pattern of data was clear. For critical targets, true recognition was greater for categories that were exhaustively recalled ($M = .90$, collapsed across conditions) than for categories that were only partially recalled ($M = .78$, $t(83) = 3.57$, $SEM = .031$, $p < .05$, reflecting the

<table>
<thead>
<tr>
<th>Item type</th>
<th>Constant length</th>
<th>Varied length</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit</td>
<td>.59</td>
<td>.66</td>
</tr>
<tr>
<td>Miss</td>
<td>.26</td>
<td>.37</td>
</tr>
<tr>
<td>CL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False alarm</td>
<td>.39</td>
<td>.43</td>
</tr>
<tr>
<td>Correct rejection</td>
<td>.54</td>
<td>.55</td>
</tr>
</tbody>
</table>

Note. Recall for critical lures (CL) was averaged across the three (Experiment 1) or five (Experiment 2) studied items of each category, whereas recall for critical targets (CT) was averaged across the two or four remaining items, respectively.

As another approach, recognition data were calculated separately for those categories from which subjects had recalled all of the studied items (exhaustive recall) and for those from which they only recalled a few of the items (partial recall). This calculation was done for each subject, and the means for each condition can be found in Figure 1. As discussed earlier, the total number of categories that elicited partial or exhaustive recall differed across the conditions, and some subjects had to be excluded from the statistical analyses because they never exhaustively recalled the items from any category. Nevertheless, the pattern of data was clear. For critical targets, true recognition was greater for categories that were exhaustively recalled ($M = .90$, collapsed across conditions) than for categories that were only partially recalled ($M = .78$, $t(83) = 3.57$, $SEM = .031$, $p < .05$, reflecting the

---

3 The data from 12 subjects (4 in the strategic-constant condition, 4 in the standard-varied condition, and 4 in the strategic-varied condition) were excluded from these analyses because these subjects failed to respond yes or no to at least one type of item, but the inclusion of the partial data from these subjects did not change the pattern of results.
positive relationship between true recall and recognition. This difference was significant in each condition. More important, false recognition from categories that were only partially recalled ($M = .41$) was greater than that from categories that were exhaustively recalled ($M = .12$), $t(76) = 6.10$, $SEM = .047$, $p < .05$. This difference was significant in all but one condition (varied-standard), and again suggests that subjects were able to use a disqualifying recall-to-reject strategy to reduce false recognition.

**Experiment 2**

The recognition results of Experiment 1 suggest that a recall-to-reject strategy was utilized most effectively in the strategic-constant condition, in which subjects were instructed to use a recall-to-reject strategy and could be sure of the category length. Additional analyses indicated that the obtained reductions were driven, at least in part, by exhaustive recall, and that subjects in every condition had lower false recognition for exhaustively recalled categories. As in the preliminary experiment, there was no evidence that partially recalling items from a category was effective at reducing false recognition via a diagnostic recall-to-reject process. Experiment 2 provided one further test of a diagnostic recall-to-reject process. This experiment was identical in all respects to Experiment 1, except five items were studied in each critical category, and four or six items were studied in each noncritical category (and strategic instructions were changed accordingly). On the basis of the results of Experiment 1, increasing category size should have made it extremely difficult for subjects to exhaustively recall all of the items from a category. Thus, if the obtained reduction in false recognition in Experiment 1 was due solely to disqualifying recall-to-reject, and diagnostic recall-to-reject did not contribute, then this effect should be eliminated in the present experiment.

**Results and Discussion**

There was again converging evidence that subjects in the strategic conditions had used recall during the recognition test. First, more subjects reported using a recall-to-reject strategy on the questionnaire in the strategic conditions ($n = 45$) than in the standard conditions ($n = 15$). Second, response latencies on the recognition test were slower in the strategic condition (overall $M = 4.938$ ms) than in the standard condition ($M = 2.373$), all $ps < .01$. Finally, positive recall-all judgments were associated with greater cued recall of studied items. For the strategic-constant condition, recall was greater for categories that elicited positive recall-all judgments ($M = .62$, collapsed across cue types) than for those that elicited negative recall-all judgments ($M = .24$), and similarly for the strategic-varied condition ($Ms = .55$ and .25), both $ps < .05$.

Analysis of the cued-recall data did not reveal any differences across the four conditions (mean recall of studied items was .39 across conditions). A 2 (target cue vs. lure cue) × 2 (constant vs. varied) × 2 (standard vs. strategic) ANOVA confirmed that there were no significant effects on recall of studied items (all $ps > .10$). When critical targets were given as cues, false recall of the critical lures ($M = .18$, averaged across conditions) was significantly greater than false recall of these same lures when their corresponding categories had not been studied (i.e., when critical target controls were cues, $M = .05$, $t(95) = 9.15$, $SEM = .015$, $p < .05$). This finding demonstrates the associative false-recognition effect. Finally, and most important, increasing category length was successful at minimizing exhaustive recall of the categories. When critical lures were given as cues, subjects recalled all of the items from only 3% of all the categories.

**Recognition.** Recognition data are presented in the bottom half of Table 1. As in the previous experiment, false recognition of critical lures ($M = .38$) was greater than false recognition to critical lure controls ($M = .11$), $t(95) = 10.85$, $SEM = .025$, $p < .05$, demonstrating the associative false-recognition effect. Turning to the effects of the instructions, it appears that subjects in the strategic-constant condition had tried to use a recall-to-reject strategy, but the result was an equivalent reduction in both true and false recognition. In the constant-length conditions, a 2 (target vs.

---

4 The data from 8 subjects were excluded from this analysis because they never made an affirmative recall-all judgment ($n = 3$ in the constant condition, and $n = 5$ in the varied condition). Whether the partial data from these subjects were included did not change the pattern of results.
lure) × 2 (standard vs. strategic) ANOVA on adjusted recognition revealed a main effect of item type, $F(1, 46) = 45.91, \text{MSE} = .05$; a main effect of instruction, $F(1, 46) = 5.37, \text{MSE} = .08$; and no interaction, $F < 1$. The main effect on instruction indicates that adjusted recognition was lower in the strategic condition for both targets and lures ($M_s = .52$ and .23, respectively) relative to the standard condition ($M_s = .69$ and .34). A similar ANOVA on data from the varied-length conditions revealed a main effect of item type, $F(1, 46) = 63.18, \text{MSE} = .05$. There was no main effect of instruction and no interaction (both $p > .10$). Unlike the previous experiment, a recall-to-reject strategy was not effective at selectively reducing false recognition.

Recall and recognition. Recall data again were calculated separately for items that were recognized and those that were not, and can be found in the bottom half of Table 2. Turning first to the critical targets, recall of the items from their category was greater ($M = .44$, collapsed across conditions) when these targets were recognized than when they were not ($M = .24$), $t(84) = 8.36, p < .05, \text{SEM} = .024$. This pattern was observed in all of the conditions, replicating a similar effect in Experiment 1. In contrast to Experiment 1, there was no negative relationship between recall and false recognition of critical lures, as most of the means were in the opposite direction. Overall, the mean level of recall was .39 for categories that led to false alarms and .37 for categories that led to correct rejections, and this difference was not significant ($p > .10$). Thus, this experiment provided no evidence that recalling only some of the words from a category could reduce false recognition. This pattern bolsters the conclusion that the effects obtained in Experiment 1 were due to exhaustive recall of the categories.

General Discussion

In these experiments, several manipulations were used to encourage recall on a recognition test. In the preliminary experiment, the presence of recall cues during the recognition test was manipulated. In Experiments 1 and 2, study categories were blocked, category labels were presented at study and test, subjects were instructed to use a recall-to-reject strategy, and subjects were prompted after each recognition decision as to whether they had used recall on that trial. Despite these rather laborious efforts, there was no evidence in any of the experiments that the recall of some studied lists had been studied as pictorially (which was considered more distinctive). False recognition was reduced when all of the studied lists had been studied as sufficient to reduce false recognition relative to a condition in which subjects were not warned. However, if subjects naturally remember during study and then rely solely on this general knowledge of the studied materials to make their recognition decisions. If they do not recall the expected features for a lure’s presentation, then the lure is rejected. Under these circumstances, the recall of other items from a category would not add any useful information for a diagnostic source-monitoring decision. This conclusion resonates with a set of findings from Neuschatz, Payne, Lampinen, and Toglia (2001). In this study subjects were instructed at test to focus on specific differences between true and false memories (e.g., perceptual or emotional) to reduce false recognition in the DRM paradigm. Contrary to expectation, this warning was not sufficient to reduce false recognition relative to a condition in which subjects were not warned. However, if subjects naturally take these characteristics into account, as part of a global monitoring process, then adding the warning—much like adding recall-to-reject instructions—would not have added any useful information for a diagnostic monitoring process.

Some work by Schacter and colleagues also suggests that diagnostic monitoring processes operate at a global level in the DRM task (e.g., Schacter, Israel, & Racine, 1999; see also Dodson & Schacter, 2001). In that work, the relative distinctiveness of a list was manipulated by presenting some lists verbally and others pictorially (which was considered more distinctive). False recognition was reduced when all of the studied lists had been studied as

---

5 As a reviewer suggested, the reduction of both true and false recognition in the constant-strategic condition could have been due to an erroneous recall-to-reject strategy. Because these subjects were instructed that they could be sure to reject an item if they recalled all five items from a category, they may have erroneously rejected both critical targets and lures when they thought they had recalled all of the items from that category.

6 A few subjects never missed a critical target ($n_s = 4$ and 2 in the standard-constant and strategic-constant conditions, respectively, and $n_s = 2$ and 3 in the standard-varied and strategic-varied conditions). All data from these subjects were excluded from the means and analyses reported here, although inclusion of the partial data from these subjects did not change the pattern of results.
pictures. The notion is that subjects extracted a general idea of the distinctiveness of all of the studied materials and then used this knowledge to reduce false recognition (a process dubbed the distinctiveness heuristic). In contrast, when the level of distinctiveness was varied across the materials (i.e., a within-subjects design), so that there was no “global” level of distinctiveness, subjects were less likely to use the distinctiveness heuristic. Of course, this is not to say that recall could never be involved in a diagnostic monitoring process. If subjects in these studies had been instructed to recall the presentation format of the relevant list, then they might have been able to use this local distinctiveness information to reduce false recognition (see Gallo, McDermott, Percer, & Roediger, 2001; see also Hirshman, 1995, for a related discussion on the list-strength paradigm). Whether subjects actually use recall in this way remains an open research question.

**Disqualifying Recall-to-Reject**

The finding that exhaustive recall was associated with reduced false recognition in Experiment 1 implicates the successful use of a disqualifying recall-to-reject strategy. As discussed in the introduction, this process is conceptually the same as the rule-based recall-to-reject strategies that have been observed in exclusion tasks. In those paradigms, the general conclusion was that recall or recollection could reduce false recognition of highly familiar lures. The present findings build upon these other findings by demonstrating that a disqualifying recall-to-reject process can be effective even when there is significant false recall (or illusory recollection) of the lure’s presentation. These findings also extend disqualifying recall-to-reject processes to a qualitatively different type of false memory—that of an event that was never presented but was suggested by associative processes. Both of these extensions are important because such situations may occur quite frequently in nonlaboratory settings (for a discussion see Roediger & McDermott, 2000).

The successful use of a disqualifying recall-to-reject strategy, in the face of significant false recall, suggests that true recall is somehow discriminated from false recall. There are at least two different ways to conceptualize how true and false recall might be distinguished. The first is based on the source-monitoring framework (see Johnson & Raye, 1981; Mitchell & Johnson, 2000). According to this framework, true and false recall can be distinguished by comparatively evaluating the memorial evidence for each event. For instance, true recall should be accompanied by greater recollection of perceptual characteristics than false recall, on average, and this could be used to distinguish between the two. Under this interpretation, diagnostic monitoring processes (e.g., comparative evaluation, criterion setting, and the like) would be involved even in disqualifying recall-to-reject processes.

In contrast to this framework, a dual process framework does not need to assume diagnostic processes. Instead, it could be assumed that recall is a threshold or all-or-none process and that false alarms are based on a separate familiarity-based process (Yonelinas, 2002). Under this framework, false recall (or illusory recollection) could be caused by a qualitatively different process than true recall, such as a familiarity-based attribution process (see Brainerd, Payne, Wright, & Reyna [2003] and Gallo & Roediger [in press] for discussions). In this case, the issue of discriminating between true and false recall might be moot. For example, upon encountering a highly familiar lure, subjects might recall three words from a category via a “direct-access” recall process. If they know (or assume) that only three words had been presented in the category, then this knowledge could prevent an attribution process that otherwise would have led to illusory recollection of the familiar lure. Unlike the source monitoring case, true and false recall would not need to be compared via diagnostic monitoring. Instead, the subject would have a good intuitive sense of when a category is exhaustively recalled, and this knowledge would inoculate him or her from experiencing illusory recall. At the present time, either of these theoretical interpretations of disqualifying recall-to-reject is tenable.

**References**


Received March 20, 2003
Revision received June 5, 2003
Accepted June 19, 2003

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at http://watson.apa.org/notify/ and you will be notified by e-mail when issues of interest to you become available!