Journal of Experimental Psychology:
Learning, Memory, and Cognition
1994, Vol. 20, No. 5, 1063-1087


0278-7393/94/$3.00

# Remembering Can Cause Forgetting: Retrieval Dynamics in Long-Term Memory

Michael C. Anderson, Robert A. Bjork, and Elizabeth L. Bjork

Three studies show that the retrieval process itself causes long-lasting forgetting. Ss studied 8 categories (e.g., Fruit). Half the members of half the categories were then repeatedly practiced through retrieval tests (e.g., Fruit Or_____). Category-cued recall of unpracticed members of practiced categories was impaired on a delayed test. Experiments 2 and 3 identified 2 significant features of this retrieval-induced forgetting: The impairment remains when output interference is controlled, suggesting a retrieval-based suppression that endures for 20 min or more, and the impairment appears restricted to high-frequency members. Low-frequency members show little impairment, even in the presence of strong, practiced competitors that might be expected to block access to those items. These findings suggest a critical role for suppression in models of retrieval inhibition and implicate the retrieval process itself in everyday forgetting.

A striking implication of current memory theory is that the very act of remembering may cause forgetting. It is not that the remembered item itself becomes more susceptible to forgetting; in fact, recalling an item increases the likelihood that it will be recallable again at a later time. Rather, it is other items—items that are associated to the same cue or cues guiding retrieval—that may be put in greater jeopardy of being forgotten. Impaired recall of such related items may arise if access to them is blocked by the newly acquired strength of their successfully retrieved competitors (Blaxton & Neely, 1983; Brown, 1981; Brown, Whiteman, Cattoi, & Bradley, 1985; Roediger, 1974, 1978; Roediger & Schmidt, 1980; Rundus, 1973).

This implication follows from three assumptions underlying what we herein refer to as strength-dependent competition models of interference: (a) the *competition assumption*—that memories associated to a common cue compete for access to conscious recall when that cue is presented; (b) the *strength-dependence assumption*—that the cued recall of an item will decrease as a function of increases in the strengths of its competitors' associations to the cue; and (c) the *retrieval-based learning assumption*—that the act of retrieval is a learning event in the sense that it enhances subsequent recall of the retrieved item. Taken together, these assumptions imply that repeated retrieval of a given item will strengthen that item, causing loss of retrieval access to other related items. We refer to this possibility as *retrieval-induced forgetting*. In this article, we explore two questions regarding retrieval-induced forgetting, one empirical and the other theoretical: (a) Is retrieval-induced forgetting a significant factor producing fluctuations in the long-term accessibility of knowledge? and (b) To what extent do such effects support the strength-dependence assumption? We believe that exploring these questions may help solve the puzzle of why so little of the knowledge available in long-term memory remains consistently accessible.

Many studies illustrate that prior retrievals can make subsequent retrieval of related information more difficult, at least within the context of a single testing session. For example, in the domain of episodic memory, the study of output interference has shown that an item's recall probability declines linearly as a function of its serial position in a testing sequence. This decline has been demonstrated with recall of paired associates (Arbuckle, 1966; Roediger & Schmidt, 1980; Tulving & Arbuckle, 1963, 1966) and categorized word lists (Dong, 1972; Roediger, 1973; Roediger & Schmidt, 1980; Smith, 1971, 1973; Smith, D'Agostino, & Reid, 1970); it occurs regardless of a category's serial position in the learning list (Smith, 1973), and it does not result from the loss of items from primary memory over time (Smith, 1971). In semantic memory, speeded generation of several category exemplars on the basis of letter cues (e.g., Fruit A_____) slows generation of later exemplars and increases the number of generation failures (Blaxton & Neely, 1983; Brown, 1981; Brown et al., 1985). These effects of output interference in both episodic and semantic memory violate expectations derived on the basis of semantic priming and spreading activation, according to which retrieval should facilitate recall of related knowledge, not impair it (Loftus, 1973; Loftus & Loftus, 1974; Neely, 1976; Warren, 1977). These effects show that retrieval-induced forgetting does occur, at least within a single testing session, which some

Michael C. Anderson, Robert A. Bjork, and Elizabeth L. Bjork, Department of Psychology, University of California, Los Angeles.

The research reported herein was supported in part by Grant 4-564040-RB-19900 to Robert A. Bjork and Grant 4-564040-EB-19900 to Elizabeth L. Bjork from the Committee on Research, University of California, Los Angeles, and by Grant MDA 903-89-K-0179 to Keith Holyoak from the Army Research Institute. The article appears on University Microfilms as part of a dissertation submitted to the University of California, Los Angeles, in fulfillment of the degree of PhD for Michael C. Anderson.

We gratefully acknowledge the assistance of Myra Jimenez, Steven Machado, and Shirley Yu in the collection of data and of Catherine Fritz, Dina Ghodsian, Keith Holyoak, Keith Horton, John Shaw, Bobbie Spellman, and Tom Wickens for comments on drafts of this article. We also thank Todd Gross, Steven Machado, Anthony Wagner, and especially Bobbie Spellman for many thoughtful conversations on the topic of retrieval inhibition.

Correspondence concerning this article should be addressed to Michael C. Anderson, Department of Psychology, University of California, 405 Hilgard Avenue, Los Angeles, California 90024-1563.

authors have taken as evidence that retrieval is a basic process underlying forgetting from long-term memory (Roediger, 1974).

Although these initial forays into retrieval-induced forgetting are suggestive, little work has been done to justify the assertion that retrieval plays a significant role in producing long-term fluctuations in accessibility. All studies of retrieval-induced forgetting have emphasized the decline in recall arising from retrievals occurring within a single test session. The extrapolation from these findings to long-lasting impairment hinges crucially on a theoretical interpretation of output interference in terms of strength-dependent competition, which is an interpretation that may not be warranted. For example, no evidence suggests that these effects reflect anything other than temporary suppression occurring within the brief span of an episodic or semantic recall task. However, if the strength-dependence interpretation is correct, such effects should not be restricted to a single output session: A single, effortful recall buried within the context of other thoughts and processes should cause forgetting of related memories on even remote occasions provided that retrieval-based learning endures. When we consider the ubiquity of retrieval in our daily cognitive experiences, retrieval-induced forgetting might be a pervasive source of long-lasting retrieval failures in long-term memory, an implication that starkly contrasts with the cursory weight given to retrieval processes in recent theoretical treatments of interference (e.g., Mensink & Raaijmakers, 1988). Thus, a major goal of the present work is to seek evidence for retrieval-induced forgetting that endures beyond the retrieval event during which it is induced.

The strength-dependence interpretation of retrieval-induced forgetting depends, of course, on the assumptions underlying strength-dependent competition. Although strength-dependent competition has a long history in interference theory (Anderson, 1976; McGeoch, 1936; Melton & Irwin, 1940; Mensink & Raaijmakers, 1988) and remains popular as a means of explaining a variety of phenomena (e.g., the increase in part-set cuing inhibition with the number of cues: Roediger, 1974; Rundus, 1973; the increase in retroactive interference with the degree of interpolated learning: Mensink & Raaijmakers, 1988; list-strength effects in free recall: Ratcliff, Clark, & Shiffrin, 1990; the exacerbation of the tip-of-the-tongue experience with recent presentation of similar words: Baddeley, 1982; Jones, 1989; Reason & Lucas, 1984; Woodworth, 1938), the empirical case for the strength-dependence assumption is not as clearly established as those for the retrieval-based learning assumption (e.g., Allen, Mahler, & Estes, 1969; Bjork, 1975; Gardiner, Craik, & Bleasdale, 1973; Hogan & Kintsch, 1971) and the competition assumption (see Watkins, 1978, for a review). When studies show that strengthening some information in memory impairs recall of other information, there is substantial disagreement on the theoretical interpretation of the impairment (regarding part-set cuing, see Basden, Basden, & Galloway, 1977; Sloman, Bower, & Roher, 1991; regarding retroactive interference, see Greeno, James, DaPolito, & Polson, 1978; Martin, 1971; Postman, Stark, & Fraser, 1968; Riefer & Batchelder, 1988; regarding the tip-of-the-tongue state, see Brown, 1991; Burke, MacKay, Worthley, & Wade, 1991).

More troubling, however, than any such theoretical disagreements are the various findings that strengthening can fail to produce impairment. These failures are illustrated vividly in studies by DaPolito (1966) and Blaxton and Neely (1983). DaPolito explored the amount of proactive interference suffered by a later studied associate to a cue (an A-C item) as a function of the number of presentations of an earlier studied associate to that cue (an A-B item). Although increasing the presentations of the A-B items from one to three increased recall for those items from 49% to 82%, recall of once-presented A-C items went from 30% to 32% (see Riefer & Batchelder, 1988, for detailed analysis of this study). In a different but related theoretical context, Blaxton and Neely (1983) demonstrated that prior presentation of several category exemplars for speeded naming actually facilitated generation of target exemplars from semantic memory. In both studies, strengthening of prior responses should have significantly impaired subsequent retrieval of related items but did not. If strengthening is not sufficient to cause impairment, retrieval-based learning may not cause long-lasting retrieval-induced forgetting.

Given the uncertain empirical status of the strength-dependence assumption, we thought it useful to treat the present work not only as an exploration of retrieval-induced forgetting but also as a test of the strength-dependence assumption itself. In the next section, we introduce a new paradigm for examining the impact of retrieval on the long-term accessibility of related information, and we contrast this method with previous procedures used to investigate strength-dependent competition. The new procedure improves on previous paradigms by unconfounding the strengthening operation from other logical phases of the experiment, a problem that has arguably generated many of the interpretational difficulties surrounding strength-dependent competition. Next, we develop predictions concerning the relative impairment expected for different stimulus materials on the basis of a general class of strength-dependent competition models: ratio-rule models. If impaired recall is observed with the new procedure, then retrieval-induced forgetting will be implicated as a significant factor in producing long-term retrieval failures. Furthermore, if the impairment follows the pattern expected on the basis of the ratio rule, then we will have obtained evidence for strength-dependent competition.

## A Paradigm for Examining Retrieval-Induced Forgetting

In constructing a paradigm to explore retrieval-induced forgetting, we thought it important to consider both the logic of strength-dependent competition and the conditions under which retrieval-induced forgetting might be expected to occur naturally. Because strength-dependent competition among items is thought to occur with respect to a shared retrieval cue, we placed special emphasis on cue–target relationships in all phases of the paradigm. We also sought to minimize opportunities for the formation of item-to-item (as opposed to cue-to-item) associations, the presence of which could provide subjects with retrieval routes for circumventing strength-dependent competition. Because retrieval-induced forgetting may arise from retrieval-based learning that occurs long after initial

learning, we separated initial study and retrieval-based learning into distinct phases; we also included a substantial retention interval between retrieval-based learning and the final test to examine the long-term effects of retrieval.

These considerations led to our designing a retrieval-practice paradigm that consists of three phases: a study phase, a retrieval-practice phase, and a final test phase. In the study phase, subjects study a series of category–exemplar pairs, such as Fruit Orange, with a typical series consisting of six members of each of eight different categories. Because the exemplars of a given category share the category label as a retrieval cue, they should compete for access to conscious recall on later presentation of the category cue. After the study phase, subjects engage in directed retrieval practice on half of the items from half of the categories (e.g., three items from each of four categories). The retrieval practice of a given item is induced by presenting a category name together with an exemplar stem (e.g., Fruit Or_____). Each exemplar test appears several times throughout the practice phase, interleaved with practice trials on other items to maximize the facilitatory effects of retrieval practice. After a substantial retention interval (e.g., 20 min), a final, surprise category cued-recall test is administered: Subjects are cued with each category name and asked to free recall any exemplars of that category that they remember having seen at any point in the experiment. If strengthening due to retrieval practice endures throughout the retention interval, the practiced exemplars in a given category should still create substantial competition for the unpracticed exemplars in that category on the delayed category cued-recall test. The impact of this competition can be assessed by contrasting the final recall of the unpracticed items from the practiced categories with the final recall of items from the unpracticed categories (i.e., those categories for which none of their exemplars had been given retrieval practice). If impairment is observed, we have evidence that retrieval-induced forgetting may contribute to long-lasting retrieval failures and that these failures may result from strength-dependent competition.

The separation of the retrieval-practice paradigm into three phases appears to have several advantages over other well-known procedures thought to provide evidence for strength-dependent competition. These features are highlighted in Figure 1, which contrasts the retrieval-practice paradigm with the retroactive-interference and part-set cuing procedures. These paradigms are represented according to their temporal organization into learning (L), strengthening (S), and final test (T) phases. (Distinct phases are depicted by boxes; contiguous boxes indicate logically distinct, but co-occurring, phases.) In the retroactive-interference paradigm, subjects learn a second list of associates to the same stimuli (L2), and these associates are strengthened by repeated study-test trials (S); this strengthening of second-list associates is thought to impair recall of earlier responses from the first list (L1) on a subsequent test (T) relative to a baseline condition in which subjects never learned the second list (L2). In the part-set cuing paradigm, several exemplars from an earlier studied categorized word list (containing exemplars $L_1 \ldots L_N$) are presented as cues at test (T), presumably strengthening (S) those cues; this strengthening of the cue exemplars is thought to impair recall of the

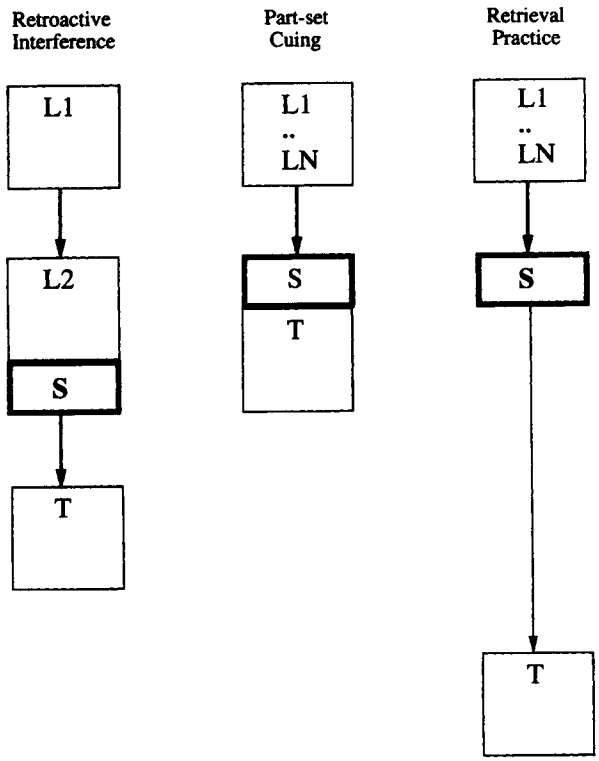


*Figure 1.* The temporal organization of retroactive interference, part-set cuing, and retrieval-practice paradigms into discrete phases. Boxes denote distinct experimental phases; contiguous boxes denote logically distinct but simultaneous phases; arrows indicate the flow of time. The letters *L*, *S*, and *T* designate learning, strengthening, and testing of items, respectively. Note that the strengthening operation is confounded with different phases for all paradigms except the retrieval-practice paradigm. Note also that the retroactive interference paradigm divides the learning of the two competitors (L1, L2) per stimulus into distinct contexts, whereas all items are learned in the same context for other paradigms.

remaining noncue exemplars relative to a baseline condition in which subjects receive no cues. The retrieval-practice paradigm, as described above, is depicted in the right column of Figure 1.

That strengthening does not occur in a distinct phase in the retroactive-interference and part-set cuing paradigms complicates interpreting the effects of that strengthening. The retroactive-interference procedure confounds strengthening of L2 competitors with the acquisition of the new temporal context (List 2) in which those competitors are learned, confusing the relative contributions of strength-dependent competition and response-set suppression to the impaired recall of L1 associates (Postman et al., 1968); in the retrieval-practice procedure, on the other hand, any response-set suppression on the learning list caused by the retrieval-practice phase should be equated across practiced categories and the within-subjects baseline (i.e., those categories that remain unpracticed; see Delprato, 1972, for a similar approach). The part-set cuing paradigm confounds strengthening of competitors with presentation of those items as retrieval cues on the final test,

obscuring the relative effects of strength-dependent competition and those deriving from the role of strengthened items as retrieval cues (Basden et al., 1977; see also Raaijmakers & Shiffrin, 1981; Sloman et al., 1991); in the retrieval-practice procedure, a long interval separates retrieval-based strengthening from the final test, and no items are presented as cues, eliminating the psychological context of cuing. To the extent that confounding the various factors described above with strengthening compromises the measure of strength-dependent competition in the retroactive-interference and part-set cuing paradigms, the retrieval-practice paradigm may provide a better means of testing strength-dependent competition.

## Testing Strength-Dependent Competition Models of Retrieval

Because our paradigm seemed to have certain advantages as a means of testing strength-dependent competition, we took our exploration of retrieval-induced forgetting as an opportunity to evaluate strength-dependent competition more systematically. Because ratio-rule formulations of retrieval are the most widely applied and best articulated strength-dependent models (e.g., Anderson, 1976; Gillund & Shiffrin, 1984; Mensink & Raaijmakers, 1988; Raaijmakers & Shiffrin, 1981; Rundus, 1973), we used a simple ratio-rule model to develop predictions of the relative amount of impairment to be expected across materials differing in their strength of association to a cue.

In the present studies, we manipulated the taxonomic frequency of exemplars in a category. In Experiments 1 and 2, to test an implication of the basic ratio-rule equation, we contrasted categories consisting entirely of strong exemplars with categories consisting entirely of weak exemplars. For a broad range of learning-rate assumptions, ratio-rule models predict that retrieval-based strengthening should impair weak exemplar categories to a proportionally greater extent than strong exemplar categories (see Appendix A for a numerical example). Qualitatively, the reason for this prediction is straightforward. The ratio-rule model asserts that the probability of retrieving an item is a function of the strength of association of that item to the retrieval cue, relative to the strength of association of all other memory items to that cue. This relation can be expressed as a simple recall probability ratio, as in the following example: P(recall Orange given the cue Fruit) = Strength of the Fruit–Orange association/sum of strengths for all Fruit associates. When other items, such as Banana, are strengthened through retrieval practice, the denominator in the equation for Orange increases, decreasing its recall probability ratio. Because retrieval practice will increase the associative strength of a weaker item to a proportionally greater extent (see Appendix A), proportional impairment of its competitors will also be greater. If retrieval-induced forgetting manifests this pattern of impairment across strong- and weak-exemplar categories, specific evidence in favor of ratio-rule formulations of strength-dependent competition will have been obtained; if it does not, the ratio rule, and perhaps strength-dependent competition in general, may be inadequate as an account of retrieval-induced forgetting.

## Experiment 1

In Experiment 1, we used the retrieval-practice paradigm to determine whether retrieval-based learning causes long-lasting memory failures. In the initial study phase, subjects studied 8 six-item categories. Four of these categories were composed of strong exemplars (e.g., Fruit Orange), and four were composed of weak exemplars (e.g., Tree Hickory). After the study phase, three exemplars from two strong and two weak categories received retrieval practice (e.g., Fruit Or_____) three times each. The three retrievals for each item, interleaved with tests of other items, were ordered to produce an expanding sequence of intertest intervals for each item to maximize the consequences of retrieval practice (see Landauer & Bjork, 1978). After a 20-min retention interval, a final unexpected category cued-recall test was administered: Subjects were cued with each category name and asked to free recall any members of that category they could remember having been presented at any point in the experiment.

To describe our predictions (for each of the experiments we report) more concisely and to simplify discussions throughout this article, we have labeled the different types of categories and items that occur in the retrieval-practice paradigm as follows: Categories for which some of their members receive retrieval practice are labeled *Rp* categories (i.e., retrieval practice categories); categories for which no members receive any retrieval practice are labeled *Nrp* categories (i.e., no retrieval practice categories). The items within an Rp category that actually receive retrieval practice are labeled *Rp+* items (i.e., Rp category, practiced items); items within an Rp category that do not receive retrieval practice are labeled *Rp−* items (i.e., Rp category, unpracticed items); and, finally, items within an Nrp category, none of which, of course, receive any retrieval practice, are simply labeled *Nrp* items. If retrieval-induced forgetting produces long-lasting retrieval failures, retrieval practice of Rp+ items should impair later recall of Rp− items (relative to recall observed for the Nrp baseline), even though retrieval-based learning occurred in a context separated from the final test by 20 min. If impaired recall of Rp− items is caused by strength-dependent competition from the Rp+ items, the impairment of weak Rp− items should be proportionally greater than the impairment of strong Rp− items.

### *Method*

#### *Subjects*

The subjects were 36 introductory psychology students from the University of California, Los Angeles, whose participation partially fulfilled a course requirement.

#### *Design*

Two factors, retrieval-practice status and category composition, were manipulated within subjects. Retrieval-practice status had three levels: (a) Rp+ items, which were practiced three times each by means of an expanding schedule of category-plus-stem cued-recall tests (e.g., Fruit Or_____ ) during the retrieval practice phase; (b) Rp− items, which were not practiced, but were members of the same category as the Rp+ items, and (c) Nrp items, which received no additional

retrieval practice and were not members of a practiced category. Nrp items, which were divided into two subgroups of three (called Nrpa and Nrpb) for counterbalancing purposes, served as a baseline against which to measure the positive effects of practice in the case of Rp+ items, and the hypothesized negative effects of practice on Rp− items.

Category composition had two levels: Strong categories, which contained exemplars whose taxonomic frequency had an average rank order of 8 (Battig & Montague, 1969); and weak categories, which contained exemplars with an average rank order of 33. The dependent measure was the proportion of each type of item recalled on a final category cued-recall test.

## Procedure

The experiment was conducted in four phases: a learning, a practice, a distractor, and a surprise category cued-recall phase. In the learning phase, subjects were randomly assigned to one of two random orders of the learning materials. Each subject was given a learning booklet, face down, as well as an instruction page, which they followed as the experimenter read the instructions aloud. Subjects were told that (a) they were participating in an experiment on memory and reasoning, (b) they would be given 5 s to study category–exemplar pairs and should spend all of this time relating the exemplar to its category, (c) after each 5 s passed, a voice on a tape recording would signal them to turn the page, and (d) the sequence was to be repeated until all pairs in the learning booklet had been presented. On completion of the instructions, subjects were told to turn their booklets over and begin studying.

Booklets and instructions were collected as soon as the learning phase was completed. Subjects were then randomly assigned to one of four practice counterbalancing conditions and to one of three retrieval-practice orders for that counterbalancing condition. Subjects received a booklet face down and a new instruction page, which they followed as the experimenter read it aloud. Subjects were told that (a) each page would contain one of the category labels that they had received in the previous phase along with a hint about what exemplar they were to retrieve; (b) the hint consisted of the first two letters of the appropriate exemplar; and (c) they were to retrieve an item that they had seen, rather than responding with any exemplar that fit the letter cues. Subjects then turned their booklets over and began the test: They were given 10 s to recall each cued exemplar, and a tape-recorded voice instructed them when to turn pages. After the practice phase, subjects participated in an unrelated causal reasoning experiment for 20 min.

In the testing phase, subjects were randomly assigned to one of three random testing orders of the categories. Booklets were distributed face down and the experimenter read instructions aloud. Subjects were told that, at the top of each page, there would be a name of one of the categories studied previously and that they should recall all exemplars of that category that they had been shown at any time in the experiment. Subjects were given 30 s for each category, and were then instructed to turn the page.

## Materials

*Category selection.* Ten categories, two of which were used as fillers, were drawn from several published norms (Battig & Montague, 1969; Marshall & Cofer, 1970; Shapiro & Palermo, 1970). The 8 experimental categories were selected in the following manner. Relatively unrelated categories (i.e., dissimilar and nonassociated categories) were chosen to ensure that measures of category-recall performance were as independent as possible. Intercategory similarity and association were first determined by the experimenters carefully assessing the relatedness of the knowledge domains (e.g., If Fruit were to be used, Vegetable would not be selected); these judgments were reinforced, using the Marshall and Cofer (1970) norms, by minimizing (a) the pairwise associations between category labels and (b) the interexemplar associations (after particular exemplars had been chosen). The phonemic similarities among the category labels was also minimized.

To reduce variations in stimulus complexity and associability, category labels were constrained to be semantically unambiguous and only one word in length (e.g., no categories such as Earth Formations were included). Finally, the word frequencies (Kucera & Francis, 1967) of category labels were kept in the low to moderate range, with all labels falling between 25 and 100 occurrences per million.

*Exemplar selection.* Once eight categories were found that met these constraints, particular exemplars were chosen for each one (see Appendix B). Four of the categories were randomly chosen to contain all strong exemplars and four to contain all weak exemplars. Exemplars in three of the strong categories had an average rank order of 8 (median = 7, i.e., average position in a list rank ordered by frequency of report), according to Battig and Montague (1969) category norms. Exemplars in the remaining strong category (Leather) were drawn from the Shapiro and Palermo (1970) norms and had an average rank order of 3.8. Exemplars in the four weak categories had an average rank order, according to Battig and Montague, of 33 (median = 23). Thus, there was a clear difference in the taxonomic frequency of exemplars in the strong versus the weak categories.

Exemplars were also constrained to be low-frequency, unambiguous, noncompound words. The average word frequency (Kucera & Francis, 1967) for all eight categories was 13 occurrences per million, $SD$ = 3.8. No two exemplars began with the same first two letters, ensuring that each two-letter cue in the retrieval-practice task would be unique. In addition, to avoid interference of extraexperimental items, no chosen category exemplar had the same first two letters as an unchosen category exemplar that was listed in the Battig and Montague (1969) norms. For example, the word *trumpet* could not be chosen as a musical instrument because the word *trombone* might produce extraexperimental interference. Items with strong a priori item-to-item associations (e.g., cat and mouse as members of the set animals) were avoided.

Finally, two constraints were used to match the effectiveness of the first two letters of an exemplar as a retrieval cue for the retrieval practice task: versatility matching and syllable matching. The versatility (Solso & Juel, 1980) of a set of letters corresponds to the number of words containing those letters in the specified positions. For example, an estimate of the versatility of the letter combination BA in the first two positions of a word is 413 because there are approximately 413 words that begin with that combination of letters in the Kucera and Francis (1967) norms. Versatilities of the two-letter stems of exemplars were constrained to be at a moderate level of difficulty ($M$ = 281, $SD$ = 12) as measured by Solso and Juel. Finally, stems were constrained to provide less than one syllable of information. In ambiguous cases, we used *Webster's New Collegiate Dictionary* (1980) to determine where syllabic breaks occurred.

*Learning booklets.* Learning booklets were constructed from the 48 experimental and 12 filler items. The placement of these items in the learning booklet was designed to minimize interexemplar associations because such associations could provide secondary retrieval routes to unpracticed items in the practiced categories, offsetting the impairment caused by the competition for the primary retrieval cue. Two measures were taken to minimize interitem association among category members and to maximize attention to category-exemplar relationships. First, category-exemplar pairs were presented to subjects centered on individual pages in paired-associate format (e.g., Fruit Orange). Second, rather than presenting all exemplars from a given category at once, the order of exemplars within a booklet was determined by blocked randomization in which each block contained one exemplar from each category, resulting in six blocks of 10 items

(each block containing 8 items from experimental categories and 2 items from filler categories). The ordering of exemplars within each block was determined randomly except that (a) in the first block, filler items appeared in the beginning to control for primacy effects; (b) in the last block, filler items appeared at the end to control for recency effects; and (c) throughout the booklet, no two categories appeared in sequence more than once. Two different learning booklets were constructed, in which both the ordering of categories within blocks and the list position of particular category items varied.

*Retrieval-practice booklets.* Each page of a retrieval-practice booklet contained one test of a single category exemplar. The category label appeared centered on the page with the first two letters of the exemplar printed two spaces to the right of it, followed by a solid line to indicate that the item was incomplete (e.g., Fruit Or_____). The stem of the exemplar was provided to direct subjects to retrieve a particular item. The solid line was the same length for all items so that no cues for word length would be given.

To construct retrieval-practice booklets, we first defined an abstract ordering of exemplar tests using the following constraints. The first and last few items in all practice booklets were tests of filler items to acquaint subjects with the practice task and to control for primacy and recency effects on final recall. All experimental items were tested three times on an expanding schedule, with an average spacing of 3.5 trials between the first and second test and 6.5 trials between the second and third test. In general, no two category members were tested on adjacent pages, and the average test position of each category in the test booklet was kept constant. To the extent possible, we prevented particular sequences of category–exemplar tests from appearing consecutively more than once (as is prone to occur with systematic spacing manipulations) by inserting tests of filler items.

To control for specific-category effects, we counterbalanced which categories were practiced and which were not. The eight experimental categories were divided into two random sets of four (referred to as Set A and Set B), with the constraint that two strong and two weak categories appeared in each set. Half of the subjects performed retrieval practice on Set A and the other half of the subjects on Set B. To control for specific-exemplar effects, we further divided Set A and Set B into two random subsets (referred to as Subsets A1, A2, B1, and B2). For Subset A1, three exemplars were randomly selected from each of the four categories in A, with the remaining three exemplars constituting A2. Half of the subjects who practiced the Set A categories practiced A1 exemplars, and the remaining subjects practiced A2 exemplars. Subsets B1 and B2 were constructed and distributed in the same manner (see Appendix B for the materials and their divisions into these sets). These procedures ensured that every item participated in every condition equally often, and resulted in four sets of 12 items (A1, A2, B1, and B2) from which we constructed retrieval-practice booklets.

Each of the four 12-item counterbalancing sets was assigned to the abstract ordering of exemplar tests three times, resulting in 12 booklets of 51 pages (three practice orders for each of the four counterbalancing sets). Distractor materials were booklets containing causal-reasoning tasks.

*Test booklets.* Each page of the nine-page test booklets contained one category cue centered at the top. The first page for all testing booklets was one of the filler categories (mountains), which was inserted to minimize variance due to output interference. The order of the remaining experimental categories was random, except that across the three testing orders, the average test position for each category and each condition was approximately the same. Each of the three testing orders was combined with each of the 12 practice booklets, yielding 36 distinct combinations.

Finally, we used a portable tape recorder to play the tape instructing subjects when to turn booklet pages and a stopwatch to time subjects in the final test phase.

## *Results and Discussion*

### *Retrieval Practice*

The retrieval practice success rates for Rp+ items varied as a function of category composition, with 74% and 90% success rates being obtained across weak and strong Rp+ items, respectively. (Note that potential difficulties of interpretation created by the differing rates of retrieval-practice success are addressed in Experiment 3).

### *Final Test Performance*

All analyses were first conducted treating the counterbalancing subgroups of Nrp items as distinct levels of the retrieval practice factor. Because no significant difference was obtained between the recall means of these subgroups ($M$ = 48.8% and 48.1% for Nrpa and Nrpb items, respectively) nor was there a simple interaction between the Nrpa-Nrpb and the strong–weak manipulation, the data from these subgroups were combined in the results reported below.

Table 1 shows the percentages of each type of item that were correctly recalled for the strong and weak categories, respectively. As expected, repeatedly retrieving several members of a studied category improved the recall of those items (Rp+ = 73.6%) relative to the baseline (Nrp = 48.4%) on the final delayed recall test, $F(1, 32) = 136.9$, $p < .0001$, $MS_e$ = .022. More important, however, is the finding of impaired recall for the remaining unpracticed category exemplars (Rp− = 37.5%) relative to the same baseline, $F(1, 32) = 30.3$, $p < .0001$, $MS_e = .019$. This pattern of improved recall for Rp+ items and impaired recall for Rp− items is consistent with the item-specific interference predicted by strength-dependent competition models of forgetting: That is, retrieval practice appears to have produced enduring retrieval-based learning of the Rp+ items, as evidenced by their improved recall performance, thereby reducing the competitiveness of the Rp− items during the final recall test, as evidenced by their impaired recall performance. Furthermore, this pattern of results indicates that retrieval-induced forgetting is not restricted to a single output session and may, in fact, contribute to long-lasting retrieval failures.

As expected, the main effect of our category composition manipulation was significant, with strong exemplars being recalled at a higher level than weak exemplars ($M$ = 58.3% and 45.7%, respectively), $F(1, 32) = 53.2$, $p < .0001$, $MS_e$ =

Table 1
*Mean Percentage of Items Recalled on a Category Cued-Recall Test as a Function of Category Composition in Experiment 1*

| | Retrieval practice status of item | | |
|---|---|---|---|
| Category composition | Rp+ | Rp− | Nrp |
| Strong exemplars | 81.0 | 40.3 | 56.0 |
| Weak exemplars | 66.2 | 34.7 | 41.0 |

*Note.* Rp+ = practiced exemplars from practiced categories; Rp− = unpracticed exemplars from practiced categories; Nrp = unpracticed exemplars from unpracticed categories.

.022. An analysis of the magnitudes of retrieval-practice facilitation for strong and weak exemplars, however, revealed that the absolute improvement for weak items was not reliably different from that for strong items, (Rp+ − Nrp = 66.2 − 41.2 = 25.0% for weak items vs. 81.0 − 56.0 = 25.0% for strong items), $F(1, 32) < 1$. Furthermore, although the proportional facilitation of weak items—measured as a percent of their Nrp baseline—appeared to be greater than the facilitation of strong items (61.5% vs. 44.6%, respectively), this difference was not statistically reliable, $F(1, 32) < 1$. This failure for weak exemplars to show greater facilitation is probably because final recall performance underestimates the facilitation of those items; final recall reflects both the facilitation of successfully practiced items and the lack of facilitation for the larger number of weak items missed entirely during practice.

Examining next the pattern of impairment for strong and weak exemplars, we first determined that reliable impairment had been obtained for both strong and weak categories, $F(1, 32) = 27.4, p < .0001, MS_e = .022$; $F(1, 32) = 4.5, p < .05$, $MS_e = .021$, respectively. Additional analyses, however, revealed that the recall of strong Rp− items exhibited both more absolute impairment and more proportional impairment than did the recall of weak Rp− items: absolute impairments being 15.7% (56.0 − 40.3) for strong Rp− items versus 6.3% (41.0 − 34.7) for weak Rp− items, $F(1, 32) = 4.6, p < .05$, $MS_e = .023$; and proportional impairments being 28.0% for strong items versus 15.4% for weak items, $F(1, 32) + 7.5, p < .01, MS_e = .194$.

Thus, whereas the overall tradeoff between facilitation and impairment observed in the present recall results is consistent with an interpretation in terms of strength-dependent competition, the results obtained from our manipulation of category composition are not what would be expected from ratio-rule models. If, for example, one assumes that weak items would be strengthened at a proportionally greater rate than strong items by retrieval practice (as we had originally expected to find), then the ratio-rule model predicts proportionally greater impairment for weak categories than for strong. If, rather, one assumes that strong and weak items would be facilitated to a proportionally equivalent degree by retrieval practice, the assumption consistent with the present results, the ratio-rule model predicts—as shown in Appendix A—greater absolute impairment for strong-exemplar categories than for weak-exemplar categories but equivalent proportional impairments, an outcome not observed in the present results. (One exception to the previous predictions, arising under certain unrealistic assumptions, is addressed in Experiment 3)

The observed pattern of impairment as a function of exemplar strength is, thus, both surprising and potentially important, appearing as it does to be inconsistent with the predictions of ratio-rule models. One approach to explaining this discrepancy would be to propose an additional mechanism that either selectively impairs recall of strong Rp− items, or that selectively facilitates recall of weak Rp− exemplars. For instance, the retrieval-practice phase may set in motion some process other than strengthening that affects the pattern of impairment, the effects of which persist throughout the retention interval. Unfortunately, the present experiment provides no way to disentangle dynamics arising at test from those arising during the retrieval-practice phase. It is possible, for example, that impaired recall of Rp− items was produced entirely at final test, arising as a consequence of the prior retrieval of strengthened Rp+ items. Indeed, an inspection of the output order of items on the final recall test of the present study supports such an interpretation: Rp+ items were reported far earlier, on average, than Rp− items, similar to the early recall of cue items in studies of part-set cuing (Roediger, Stellon, & Tulving, 1977).

In summary, then, the temporal locus (or loci) of the mechanism (or mechanisms) contributing to the impaired recall of Rp− items cannot be determined with precision on the basis of the results of Experiment 1 alone. We thus designed Experiment 2 to test whether impaired recall of Rp− exemplars would still be observed when the output order of the exemplars in a given category was controlled at the time of the final test.

## Experiment 2

In Experiment 2, we used the same procedure and materials as in Experiment 1 except that we replaced the category-cued free-recall test with a category-plus-stem cued-recall test, which allowed us to control for the order in which Rp+ and Rp− items were output at the time of the final test. More specifically, each item on the final test, as in the retrieval-practice phase, was tested on a single page by presenting a category name and the first two letters of that exemplar. Using the first two letters of an exemplar to direct the subjects' search enabled us to manipulate whether Rp− items were tested first or second in their categories—hereinafter referred to as *Rp−1st* and *Rp−2nd* items, respectively—and whether Nrp items were tested first or second—hereinafter referred to as *Nrp1st* and *Nrp2nd* items, respectively.

By comparing the recall of Rp−1st items to that of Nrp1st items, we would be able to obtain a measure of Rp− recall that was free of any potential output interference effects from the recall of Rp+ items. Thus, any recall impairment observed for these Rp−1st items would have to reflect the long-term consequence of events that had occurred during the retrieval-practice phase, rather than the consequence of output interference dynamics occurring during the final test phase. Similarly, by comparing the recall of Rp−2nd items to that of Nrp2nd items, we would obtain a measure of Rp− impairment from which potential interference effects owing to the earlier recall of Rp+ items had been eliminated: The recall tests for both sets of these items would follow the tests for items recalled first in their respective categories (i.e., Rp+1st and Nrp1st items), thus, their recall should be equally affected by output interference. If output interference actually does contribute to recall in this task, a comparison of the recall levels for Nrp1st and Nrp2nd items should reveal that the former are recalled better than the latter. Given this result, we would expect the difference in recall performance for Rp−1st versus Nrp1st items or for Rp−2nd versus Nrp2nd items, either of which would be a measure of Rp− recall impairment uncontaminated by output interference, to be less than the difference between the recall for Rp−2nd and Nrp1st items because this

latter difference should reflect the recall of Rp− items impaired by both output interference and any potential long-term effects from the retrieval-practice phase. That is, a comparison between the recall of Rp−2nd items and Nrp1st items would produce a measure of Rp− recall that would be subject to the same effects as had influenced the Rp− recall observed in Experiment 1.

## Method

### Subjects

The subjects were 48 introductory psychology students from the University of California, Los Angeles, whose participation partially fulfilled a course requirement.

### Design

The design of Experiment 2 differed from that of Experiment 1 in how final recall was measured: Accessibility of category exemplars was assessed with a category-plus-stem completion task rather than a category-cued free-recall task, so that the order for testing category exemplars could be manipulated. Thus, the design involved three factors, all manipulated within-subjects: retrieval practice, category composition, and testing position, with retrieval practice and category composition being manipulated exactly as they had been in Experiment 1.

The final test booklet was blocked by categories. The testing order of exemplars within category blocks was manipulated on two levels: The first half of the block constituted the tested-first exemplars (e.g., Rp−1st and Nrp1st items) and the last half constituted the tested-second exemplars (e.g., Rp−2nd and Nrp2nd items). The dependent measure was the percentage of words recalled in a category-plus-stem cued-recall test.

### Procedure

To the point of the final test, the procedure we used in Experiment 2 exactly matched the procedure used in Experiment 1. In the final test phase, subjects were instructed that they would be tested in a way similar to that in which they had been tested in the practice phase. More specifically, subjects were told that on each page of the test booklet they would see the name of a category with the first two letters of an exemplar next to it and that their task was to retrieve the exemplar, from any portion of the experiment, that corresponded to those cues. Subjects were given 10 s to recall each item, after which time a tape-recorded voice instructed subjects to turn the page. This sequence was repeated until all trials in the test booklet were completed.

### Materials

The apparatus, as well as the learning, practice, and distractor materials, were identical to those used in Experiment 1.

Each page of the final test booklets had one category-plus-stem cued-recall test. Tests of exemplars were blocked by category to match the recall conditions of Experiment 1 as closely as possible. Finally, items of a particular type (e.g., Rp+, Rp−, Nrpa, and Nrpb) were always tested in sequence, being either the first three or the last three items tested within their respective categories.

The average test booklet position of category types (i.e., Strong and Weak) was controlled by creating the following order of category types: S, W, W, S, S, W, W, S. This general order of category types was used to construct two specific counterbalanced orderings of categories: The first ordering was constructed by selecting categories from the strong and weak sets and randomly assigning them to appropriate positions; the second ordering was constructed by switching categories from the first half of the first test sequence with those of the second. The average testing position of practiced and unpracticed categories was controlled by implementing one pattern (Rp, Nrp, Nrp, Rp, Nrp, Rp, Rp, Nrp), which was then inverted when we counterbalanced the categories that were practiced.

The testing order of particular exemplars within a category was counterbalanced by switching the first three exemplars with the second three. The exemplar-position counterbalancing crossed with the category-position counterbalancing (resulting in four test booklet types) ensured that all items contributed to all testing-order and practice-condition combinations (e.g., Rp+1st, Rp+2nd, Rp−1st, Rp−2nd, etc.) and that all categories and exemplars had the same average testing position.

Each of the four retrieval-practice counterbalancing conditions (A1, A2, B1, and B2, as in Experiment 1), each having three random orders, was paired with each of the four final test booklet types, resulting in 48 practice-book–test-book combinations (one for each subject).

## Results and Discussion

### Retrieval-Practice Performance

As in Experiment 1, the retrieval practice success rates for Rp+ items varied as a function of category composition, with a 76.1% and 85.0% success rate being obtained across weak and strong Rp+ items, respectively.

### Final Test Performance

As for Experiment 1, all statistical analyses were initially conducted treating the counterbalancing subgroups of Nrpa and Nrpb as distinct levels of the retrieval-practice factor. However, because the mean correct recall percentages for these subgroups (71.2% and 74.1%, respectively) did not differ significantly, $F(1, 44) = 1.6, p = .21$, their data were combined into a single Nrp measure for ease of exposition. Similarly, data were collapsed across our other two counterbalancing factors because they did not interact with the variables of interest.

Table 2 shows the percentages of each type of item that were correctly recalled on the final category-plus-stem cued-recall test for strong and weak exemplars, respectively, as a function of their within-category testing position. As might have been expected, the addition of a two-letter cue during the final test substantially increased the overall level of recall in Experiment 2 as compared with that of Experiment 1 ($M = 75.7\%$ vs. 52.0%, respectively). The overall correct recall percentages increased from 59% to 82.8% for strong exemplars and from 47% to 68.5% for weak exemplars. As can be seen from observing the means reported in Table 2, retrieval practice appeared to facilitate weak exemplars more than strong exemplars (Rp+ − Nrp = 79.9 − 62.7 = 17.2% for weak exemplars and 91.0 − 82.7 = 8.5% for strong exemplars), $F(1, 40) = 3.9, p = .054$, a result that is likely to be an artifact of the very high recall performance of the strong exemplars and, as such, not likely to be meaningful.

### *Final Test Performance Averaged Across Output Position*

In general, the findings of Experiment 2 replicated those of Experiment 1, despite our use of a substantially different testing method. We obtained a significant main effect for category composition, with strong exemplars being recalled more frequently than weak exemplars ($M$ = 82.7% and 67.0%, respectively), $F(1, 40) = 73.6, p < .0001, MS_e = .064$. Planned comparisons revealed that retrieval practice improved the recall of Rp+ items over that of Nrp items ($M$ = 85% and 73%, respectively), $F(3, 120) = 37.2, p < .0001, MS_e = .056$, but, on the whole, did not reliably damage the recall of Rp− items relative to that of Nrp items ($M$ = 68.8% and 73%, respectively), $F(1, 40) = 2.3$, $p = .13$. This main-effect comparison for Rp− impairment, however, is obscured by a marginal interaction with category composition, $F(1, 40) = 2.8$, $p = .10$, $MS_e = .076$. Because Experiment 1 had led us to expect an interaction between our category-composition and our retrieval-practice factors and because strong items, but not weak items, may have been subject to ceiling effects, we reasoned that any inhibiting effects on the recall of strong categories may have been artificially reduced, lessening the chance for obtaining a significant interaction. We, therefore, regarded this marginal interaction as sufficient grounds to examine the potential inhibitory effects of retrieval practice on strong items and weak items in isolation. Comparisons revealed that Nrp items were recalled at a significantly higher rate than Rp− items (82.7% vs. 74.7%) for strong categories, $F(1, 40) = 7.2, p < .01, MS_e = .060$, whereas there was no evidence for a difference in the recall of Nrp and Rp− items (62.7% vs. 62.9%) for weak categories. As in Experiment 1, there was a proportionally greater degree of impairment for strong Rp− items than for weak Rp− items (9.7% vs. 0%), $F(1, 44) = 5.8, p < .05$. Interestingly, this finding, like those of Blaxton and Neely (1983) and DaPolito (1966) discussed in the introduction of this article, appears to be an instance in which strengthening fails to cause impairment.

Finding impairment with the category-plus-stem cued-recall testing procedure used in Experiment 2 is surprising for at least two reasons. First, it is surprising to the degree that stem completion, which was essentially what this testing procedure required, resembles recognition testing. It is well known that retroactive interference effects are greatly attenuated (and often eliminated) when a recognition testing procedure is used instead of modified-modified free recall (see, e.g., Postman & Stark, 1969), suggesting that such interference effects reflect difficulties in retrieval. Second, other effects of retrieval inhibition (e.g., part-set cuing inhibition and the list-strength effect) are either rather small (Todres & Watkins, 1981) or are nonexistent (Ratcliff et al., 1990; Slamecka, 1975) with recognition testing, unless more sensitive tests (e.g., recognition time, see Neely, Schmidt, & Roediger, 1983) are used. Because we did observe retrieval-induced forgetting for a stem-completion testing procedure, however, it follows that either (a) the retrieval demands of stem completion are more similar to those imposed by recall than to those imposed by recognition, or (b) the current impairment is qualitatively different from part-set cuing and retroactive interference effects.

Table 2
*Mean Percentage of Items Recalled on a Category-Plus-Stem Cued-Recall Test as a Function of Category Composition and Within-Category Testing Position in Experiment 2*

| | Retrieval practice status of item | | |
|---|---|---|---|
| Category composition | Rp+ | Rp− | Nrp |
| Strong exemplars | | | |
| Tested first | 91.0 | 77.8 | 85.4 |
| Tested second | 91.0 | 71.5 | 79.9 |
| *M* | 91.0 | 74.7 | 82.7 |
| Weak exemplars | | | |
| Tested first | 79.9 | 63.2 | 59.7 |
| Tested second | 79.9 | 62.5 | 65.7 |
| *M* | 79.9 | 62.9 | 62.7 |

*Note.* Rp+ = practiced exemplars from practiced categories; Rp− = unpracticed exemplars from practiced categories; Nrp = unpracticed exemplars from unpracticed categories. Tested first or second = items tested in the first three or second three positions of a category block. Comparisons of Rp− and Nrp items within a given row reflect practice-induced inhibitory effects alone. Comparison of Rp− tested second and Nrp tested first reflects the combined effects of practice- and test-induced inhibition.

### *Impact of Testing Order on Final Test Performance*

As the output order of items in Experiment 1 had led us to suspect, the prior recall of other category members at the time of the final test did impair the recall of later items in Experiment 2. Although the main effect of testing position did not reveal an advantage for earlier items ($M$ = 75.3%) over later items ($M$ = 74.5%), this factor showed a marginal interaction with category composition, $F(1, 40) = 3.9, p = .056, MS_e = .063$. Consistent with the tendency observed in Experiment 1 for strong exemplars to be more impaired than weak exemplars, the effect of output interference at the time of the final test was greater for strong exemplars than it was for weak exemplars in Experiment 2. That is, whereas the overall correct recall percentage for strong exemplars tested first (84.7%) was significantly better than that for strong exemplars tested last (80.6%), $F(1, 40) = 4.0, p < .05, MS_e = .045$, the overall correct recall percentages for weak exemplars tested first showed no advantage over that for weak exemplars tested last (65.6% vs. 68.4%, respectively), $F(1, 40) = 1.1, p > .05$. Interestingly, for strong items, the two sources of impairment—the impairment due to testing position and the impairment due to practice of other category members—appear to be independent effects: Collapsing across testing order, the impairment due to the retrieval-practice factor (Nrp − Rp− = 82.7 − 74.7) was significant, $F(1, 44) = 7.2, p < .01$, and this factor did not interact with testing position, $F(1, 40) < 1$.

Perhaps the most important findings of Experiment 2 concern the variations in Rp− impairment as a function of our testing order manipulations. First is the demonstration of impairment even when Rp− items were tested prior to Rp+ items. As noted, the reliable impairment observed for strong Rp− items did not vary with the position in which Rp− items were tested, Nrp1st − Rp−1st = 7.6% and Nrp2nd − Rp−2nd = 8.4%. Because Rp− items that were tested first were

not contaminated by the potentially interfering effects of Rp+ output, we can attribute the impairment of strong Rp−1st items to effects enduring from the retrieval practice phase. Second is the demonstration that the output of Rp+ items before Rp− items did result in some additional impairment for the strong Rp− exemplars. Looking at Table 2, if one compares Rp−2nd performance, which is subject to both retrieval-practice and output sources of inhibition, with Nrp1st performance, which is free from both sources of inhibition, the difference (13.9%) is larger than that between Rp−1st and Nrp1st performance (7.6%), which is a measure of Rp− impairment free of any potential output interference effects, and that between Rp−2nd and Nrp2nd performance (8.4%), which is a measure of Rp− impairment from which potential output interference effects have been eliminated. It appears, then, that under circumstances in which output order is not constrained, practiced items will tend to be recalled first, adding to the long-term debilitating effects of retrieval practice, at least for strong items.

### *Possible Explanations*

The finding of impairment when Rp− items were tested first rules out the possibility that the retrieval-induced forgetting observed in the present paradigm can be entirely due to output interference dynamics operating at the time of the final recall test. We turn now to a consideration of explanations for Rp− impairment in terms of enduring consequences of processes set in motion by the retrieval practice given to Rp+ items and to a consideration of our failures in both Experiments 1 and 2 to obtain a pattern of Rp− impairment consistent with predictions of ratio-rule models. Four accounts of this apparent violation of the strength-dependence assumption are outlined and then tested in Experiment 3: (a) covert retrieval and strengthening bias, (b) extraexperimental interference, (c) lateral inhibition, and (d) suppression.

*Covert retrieval and strengthening bias.* Although the present findings clearly violate the most straightforward predictions of the ratio-rule model, perhaps aspects of our procedure conspired to make our results appear as though the ratio-rule model had been violated. For instance, covert retrievals during the retrieval-practice phase of our paradigm might have influenced the relative impairment across strong and weak categories. Perhaps the present pattern of impairment could be made consistent with ratio-rule models if additional strengthening deriving from such retrievals selectively reduced the impairment expected for weak Rp− exemplars.

Analysis of the expected pattern of covert retrievals illustrates, however, that such intrusions, were they to occur spontaneously (as opposed to strategically), should, in fact, decrease impairment more for strong Rp− items than for weak Rp− items. Strong Rp− items should be more likely to intrude and be strengthened than should weak Rp− items; covert retrieval, therefore, should favor the recall of strong Rp− items. The question remains, however, whether subjects used some strategy during practice of weak categories that enabled selective rehearsal of weak Rp− items, thereby reducing the final recall impairment to weak categories. Subjects might have adopted such an intentional rehearsal strategy if there was a clear difference in difficulty between strong and weak Rp+ items that highlighted the necessity of giving extra rehearsal to weak items. If the difficulty of weak Rp+ items triggers strategic rehearsal of weak Rp+ and Rp− items, impairment should not arise whenever Rp+ items are weak and should arise whenever Rp+ items are strong, provided that significant strengthening of the practiced items occurs.

A second aspect of the present data that complicates the interpretation of the greater impairment for strong items is that ceiling effects prevented us from accurately assessing the relative facilitation of strong and weak Rp+ items. Although ceiling effects were clearly not a problem in Experiment 1, a potentially greater strengthening of strong Rp+ items in Experiment 2 might have caused the greater impairment of strong Rp− items. Such concerns are fueled by the differences in retrieval-practice success rates observed in both Experiments 1 and 2. If either strengthening bias or strategic covert rehearsal occurred, competition might still be strength dependent in the sense predicted by the ratio rule.

*Extraexperimental interference.* A second explanation of the greater impairment for strong items emerges if extraexperimental exemplars contributed to the patterns of impairment observed in Experiments 1 and 2, as might occur if subjects failed to use a representation of the experimental context as a retrieval cue. When the potential contribution of extraexperimental interference is considered, the ratio-rule model can predict greater proportional impairment for strong categories and minimal impairment for weak categories. These predictions derive from differences in the composition of the set of extraexperimental exemplars across strong and weak categories. To illustrate, because strong studied categories included many of their strongest exemplars as part of the study list, their extraexperimental sets should contain mainly weak exemplars; in contrast, extraexperimental sets for weak categories should contain the strong exemplars. Because the negative impact of retrieval-based learning on Rp− items can be shown to be far greater when the net strength of the extraexperimental set is low than when it is high (assuming that the experimental context is not used as a cue, see Appendix A), the impairment to strong categories can be great, whereas the impairment to weak categories can be minimal, owing to the differential makeup of their extraexperimental sets of exemplars.

*Lateral inhibition.* A third possibility consistent with the results thus far is that competition may be strength dependent but in a way that we did not expect: Practice of strong Rp+ items might produce more absolute and proportional impairment than practice of weak Rp+ items. Although this would not be consistent with the ratio rule, greater impairment deriving from the practice of strong exemplars might result if strong Rp+ items were more effective inhibitors than were weak Rp+ items, as might be the case if impairment were caused by automatic lateral inhibition among category exemplars. Such models have been suggested to account for the negative effects of part-list cues on retrieval of related material (Blaxton & Neely, 1983; Martindale, 1981; Roediger & Neely, 1982).[1]

[1] It is not a necessary property of lateral-inhibition models that they predict greater impairment for strong categories than for weak categories. For example, one might assume that exemplar nodes in a

*Suppression.* A final possibility is that the greater impairment of strong Rp− items results from a process of active suppression (as suggested by Keele & Neill, 1978, in their model of attention; see also Blaxton & Neely, 1983; Carr & Dagenbach, 1990; Dagenbach, Carr, & Barnhardt, 1990; Neill & Westberry, 1987), which is an inhibitory process that acts on those Rp− items during the retrieval-practice phase. Suppose that we assume that spontaneous covert retrievals did occur during retrieval practice but not in a way that led to covert strengthening of competitors. Instead, suppose that the provision of the category cues during retrieval practice primed all category members but that the stem cues directed access sufficiently so that competitors were not consciously intruded. Activation of Rp− items in this manner, however, may have created retrieval discrimination problems, slowing access to Rp+ items. If inhibition were used to overcome such discrimination problems, and if strongly associated exemplars interfered more frequently than weak exemplars—and were, thus, suppressed or inhibited more frequently than weak exemplars—the greater impairment of strong Rp− items could be explained.

Like the lateral-inhibition approach, the suppression account explains the impaired recall of Rp− items by an inhibitory process; unlike lateral inhibition, however, the amount of impairment suffered by Rp− items is thought to be modulated by the amount of interference caused by Rp− items rather than the strength of the Rp+ items. Thus, the suppression hypothesis need not make the strength-dependence assumption inherent to both the ratio rule and lateral inhibitory models because the extent to which Rp− items are impaired depends only on their own strength. Experiments 1 and 2 cannot distinguish between lateral inhibition and suppression because we used homogeneous categories; thus, the greater impairment for strong items could have resulted from either the greater strength of Rp+ or of Rp− items. Experiment 3 was designed to discriminate among these possible accounts of the greater impairment for strong categories.

## Experiment 3

Experiment 3 explores mechanisms that might underlie the greater retrieval-induced forgetting for strong categories observed in Experiments 1 and 2. In particular, we attempt to distinguish among the four accounts proposed in the discussion of Experiment 2: (a) the strengthening bias and covert retrieval hypothesis, which asserts that the greater impairment for strong categories is an artifact of biases in the strengthening of Rp+ items and in the covert rehearsal of Rp− items during retrieval practice; (b) the extraexperimental interference hypothesis, which asserts that greater impairment for strong categories derives from the differential composition of the set of extraexperimental exemplars across strong and weak categories; (c) the lateral inhibition hypothesis, which asserts that strong Rp+ items are better inhibitors than are weak Rp+ items; and (d) the suppression hypothesis, which asserts that the greater impairment for strong categories arises because strong Rp− items are more interfering than weak Rp− items, and thus, are more vulnerable to suppression during retrieval practice.

We implemented several modifications of the design and procedure in Experiment 3. First, to eliminate the ceiling effects on the recall of Rp+ and Nrp items observed in Experiment 2, we made the final test more difficult by using single-letter rather than double-letter word-stem cues. Second, category composition was manipulated between subjects in the present experiment to reduce subject strategies arising from contrasts in the difficulty of strong versus weak Rp+ items during retrieval practice. Finally, we expanded our manipulation of category composition to include mixed categories (i.e., categories composed of three strong and three weak exemplars), resulting in four levels of category composition instead of two: the pure strong condition, with strong items practiced (hereinafter designated the SS condition, where the underlined letter denotes the subset that is practiced), the mixed condition with strong items practiced (SW), the mixed condition with weak items practiced (WS), and the pure weak condition with weak items practiced (WW).

The inclusion of mixed categories in the present experiment should allow us to discriminate among the four accounts of the greater impairment for strong categories obtained in Experiments 1 and 2. The predictions of these four hypotheses are summarized in Table 3 in terms of the hypothesized influence of retrieval practice on Rp− items. Note that the four hypotheses make identical predictions for the pure category conditions (i.e., SS and WW), but vary in what they predict for the mixed categories (i.e., SW and WS). Consider first the covert retrieval and extraexperimental interference hypotheses, depicted in Rows 1 and 2, either of which, if confirmed,

Table 3
*Hypothesized Influence of Retrieval Practice on Rp− Recall as a Function of Rp+ and Rp− Exemplar Strength*

| Hypotheses | Category composition (example items) | | | |
|---|---|---|---|---|
| | SS (Orange, Banana) | SW (Orange, Kiwi) | WS (Guava, Banana) | WW (Guava, Kiwi) |
| Covert retrieval plus strengthening bias | − | − | + | 0 |
| Extraexperimental interference | − | − | − | 0 |
| Automatic lateral inhibition | − | − | 0 | 0 |
| Suppression | − | 0 | − | 0 |

*Note.* SS, SW, WS, and WW designate categories composed of either all strong exemplars (SS), all weak exemplars (WW), or half strong and half weak exemplars (SW and WS). The strength of the practiced and unpracticed items (Rp+ and Rp− items) is indicated by underlined and nonunderlined letters respectively. − = inhibitory effects; + = facilitatory effects; 0 = neutral effects.

lateral-inhibitory network had nonlinear activation functions that reduced or enhanced inhibitory inputs, dependent on the current activational state of the node. For present purposes, the important point is that the amount of impairment inflicted by an inhibiting item does depend on the strength of the association between the cue and the inhibiting item and that this strength-dependent process can, under certain assumptions, cause greater impairment for strong categories.

would support a ratio-rule interpretation of our results. According to the covert-retrieval hypothesis, subjects give extra rehearsal to weak Rp+ and Rp− items because weak Rp+ items seem difficult. If subjects rehearse in this manner, there should be no impairment whenever Rp+ items are weak (WS and WW) with the potential for facilitation when Rp− items are more accessible for rehearsal (WS). Furthermore, there should be significant impairment in the SW condition because subjects should not consider it necessary to perform extra rehearsal on strong Rp+ items. The inclusion of mixed categories also controls for variations in extraexperimental interference because the contents of the extraexperimental exemplar sets for SW and WS conditions are identical; thus, there should be impairment in both mixed conditions, provided that significant strengthening occurs for Rp+ items.

Next, consider the two inhibitory hypotheses—lateral inhibition and suppression depicted in Rows 3 and 4. If the greater impairment for strong categories resulted because strong Rp+ items are better inhibitors, there should be more impairment for conditions containing strong Rp+ items than for conditions containing weak Rp+ items (i.e., average of SS and SW impairment > average of WS and WW impairment). Finally, if the greater impairment for strong categories arises because strong items are more vulnerable to suppression, more impairment should occur for conditions containing strong Rp− items than for conditions containing weak Rp− items, irrespective of the strength of the practiced set (i.e., the average of SS and WS impairment > average of SW and WW impairment).[2]

An additional benefit arising from the inclusion of mixed categories in Experiment 3 is that it affords further tests of the ratio-rule model. Ratio-rule models make two predictions with respect to performance on tests of our Nrp baseline items. First, the probability of recalling a strong exemplar should be greater for strong items in an SW baseline category than for strong items in an SS baseline category. This prediction arises because the presence of additional strong items in the SS category reduces the relative strength of those strong items. Second, for similar reasons, weak items in SW baseline categories should be recalled less well than weak items in WW baseline categories because the presence of strong items should reduce their relative strengths. Thus, our mixed baseline categories enable us to test predictions of the ratio-rule model on the basis of results that are not likely to have been affected by any special dynamics that may have arisen in our retrieval-practice phase.

## Method

### Subjects

The subjects were 64 students (16 in each of the four between-subjects conditions) from the University of California, Los Angeles. Of these, 48 students participated in partial fulfillment of a course requirement and 16 students (8 in condition SW and 8 in condition WS) were paid for their participation.

### Design

The design of Experiment 3 differed from that of Experiment 2 in that category composition was manipulated between subjects and had four levels instead of two: The strong–strong (SS) and the weak–weak (WW) conditions contained only strong and weak categories, respectively; and the remaining two conditions, SW and WS, contained categories composed of three strong and three weak exemplars. In the SW condition, subjects practiced the strong items, whereas in the WS condition, subjects practiced the weak items. As in Experiment 2, both the practice status of an item and testing order were manipulated within subjects.

The dependent measure was the percentage of words recalled in a category-plus-stem cued-recall test, in which single-letter stems were used instead of two-letter stems as had been used in Experiment 2.

### Materials and Procedure

The materials used in Experiments 1 and 2 were revised to meet the constraints imposed by our expanded manipulation of category composition. As illustrated in Appendix C, eight large categories were constructed, each with 12 exemplars (6 strong and 6 weak) so that each category could participate in the SS, SW, WS, and WW conditions. The newly constructed categories and exemplars had characteristics similar to those used in previous experiments. According to Battig and Montague (1969) category norms, strong exemplars had an average rank order of 8, and weak exemplars had an average rank order of 50, which was substantially lower than that of weak items in Experiments 1 and 2 ($M = 33$). Thus, there was a clear difference in the taxonomic frequency of exemplars across the strong and weak item sets.

As before, exemplars were constrained to be low-frequency, noncompound words. The average word frequency (Kucera & Francis, 1967) for all eight categories was 12 occurrences per million, not differing substantially between strong ($M = 15$) and weak exemplars ($M = 8$). Because the new final test used only the first letters of exemplars to cue subjects, no two exemplars within a category were allowed to begin with the same first letter. Exemplars from different categories could begin with the same first letter (for the obvious reason that we have more than 26 words), but efforts were taken to distribute this overlap among letters, categories, and conditions. Because our materials pool was large, we relaxed the constraints that no exemplar could begin with the same first two letters as any extraexperimental exemplar from its own category or as any exemplar from other presented categories, although these constraints were honored to the degree possible. As before, versatilities of the two-letter stems were constrained to be at a moderate level of difficulty ($M = 246$), and did not differ substantially across strong ($M = 244$) and weak ($M = 248$) exemplars. The construction of such large categories in accordance with these constraints required us to replace two of our previous categories, Leather and Hobbies, with new categories, Insects and Fish.

*Learning booklets.* The strong and weak exemplars of each category were randomly divided into two subsets, S1 and S2 in the case of strong exemplars and W1 and W2 in the case of weak exemplars, as illustrated in Appendix C. We used these materials to construct six different types of learning booklets: SS booklets, containing only

[2] A further prediction might be made that strong Rp− items should be more impaired in the WS than in the SS condition because those items might cause more interference during the practice of weak Rp+ items. This prediction requires that either (a) the probability that a strong Rp− item will intrude is a function of its strength relative to Rp+ items in that category rather than a function of its own absolute strength, or (b) the intrusion probability for strong Rp− items is equivalent in the WS and SS conditions but that the longer search time necessary for weak Rp+ items provides more occasions for intrusion, and thus, inhibition. Although the former approach can be questioned on the basis of the failures of strength-dependent competition in Experiment 2, the latter assumption seems plausible.

categories having six strong exemplars each; WW booklets, containing only categories having six weak exemplars each; and four SW booklets, containing only categories having three strong and three weak exemplars each. (Note that no underlining is needed to denote the contents of the learning booklets and that the order of S and W is irrelevant.) The latter four booklets were designed by making all four possible combinations of strong and weak subsets of our categories: S1W1, S1W2, S2W1, and S2W2. Thus, we completely counterbalanced for exemplar-specific effects within each exemplar type (S or W), and, in the case of SW categories, ensured that all combinations of strong and weak exemplars were presented for study.

*Retrieval-practice booklets.* As in Experiments 1 and 2, the eight categories were randomly divided into two subsets of four each: sets A and B. For each of our four category-composition types, SS, SW, WS, and WW, one half of the subjects were given retrieval practice on Set A, the other half on Set B. In the cases of SS and WW, the exemplar-specific counterbalancing was identical to that used in the previous experiments: Half of the subjects practiced condition S1 (or W1) and half practiced S2 (or W2), resulting in four retrieval-practice counterbalancing conditions: AS1, AS2, BS1, and BS2 (or AW1, etc. in the case of weak exemplars). In the SW and WS conditions, only the category-level counterbalancing was used because the distinction between these two conditions reflects the item counterbalancing (i.e., the only difference between WS and SW subjects was which items they practiced). Thus, for both SW and WS conditions, there were only two retrieval-practice counterbalancing conditions. Eight retrieval-practice booklets were constructed to implement these counterbalancing measures: four booklets—S1, S2, W1, and W2—for each of our two category subsets, A and B. Unlike our previous studies, however, only one random order for each booklet type was constructed instead of three.

*Final test booklets.* The format of the testing pages of the final test booklets was identical to that of Experiment 2: one category-plus-stem cued-recall test per page. The test-phase-counterbalancing and average-position-matching measures were also carried over from Experiment 2, with the following exceptions: (a) Because, for any given subject, all categories were of one type only (e.g., SS), matching of the average testing position of category types was unnecessary, and (b) the counterbalancing of the half of the testing sequence in which a category appeared was eliminated. These measures resulted in 2 test counterbalancing conditions (corresponding to the exemplar-order counterbalancing) for each of our six different learning booklet types. Because testing orders for SW and WS conditions were identical, however, only eight booklet types were actually required to implement these 12 conditions.

The two practice counterbalancing booklets for each of the four combinations of SW learning booklets (S1W1, S1W2, S2W1, and S2W2), when crossed with the 2 different test booklet types, resulted in 16 practice–test booklet combinations, one for each subject. The 4 practice counterbalancing booklets for SS and WW learning booklets, when combined with testing order counterbalancing, resulted in 8 different practice–test booklet combinations, one for every 2 subjects. Filler materials were identical to those used previously. The procedure used in Experiment 3 was identical to that of Experiment 2.

## Results and Discussion

### Retrieval-Practice Performance

The retrieval-practice success rates varied across the SS ($M$ = 82%), SW ($M$ = 82%), WS ($M$ = 67%), and WW ($M$ = 68%) conditions, as one might have expected on the basis of the differing taxonomic frequencies of practiced items across these sets. Note that the retrieval-practice success rates were equivalent for conditions in which the taxonomic frequencies of items were the same (e.g., for SS and SW and for WS and WW).

### Final Test Performance

As in Experiments 1 and 2, we collapsed across most of our counterbalancing factors because they did not interact with the variables of interest. The statistical treatment of Nrpa and Nrpb subdivisions, however, differed somewhat from that of the previous two experiments. Whereas it was feasible to collapse across these two measures in the SS and WW groups, in which Nrpa and Nrpb subsets represented the same item pools, it was not feasible in the SW and WS conditions, in which Nrpa and Nrpb subsets reflected different item pools (strong and weak items). To avoid differences in the number of observations entering into Nrp measurements between homogeneous categories (SS and WW) and heterogeneous categories (SW and WS), we restricted our comparisons of Rp− items to the Nrpb subset (which always matched the taxonomic frequency of Rp− exemplars) and our comparisons of Rp+ items to Nrpa subsets (which always matched the taxonomic frequency of Rp+ exemplars).

Table 4 shows the percentages of each type of item that were correctly recalled on the final category-plus-stem cued-recall test as a function of category composition and within-category testing position. As expected, overall performance in Experiment 3 ($M$ = 56.2%) decreased relative to that observed in Experiment 2 ($M$ = 74.8%), most likely owing to the use of single-letter rather than two-letter stems to cue the recall of exemplars during the final test. This decrease in performance eliminated the possibility of a ceiling-effect problem as had occurred in Experiment 2, allowing us to assess reliably the

Table 4
*Mean Percentage of Items Recalled on a Category-Plus-Stem Cued-Recall Test as a Function of Category Composition and Within-Category Testing Position in Experiment 3*

| | Retrieval practice status of item | | | |
|---|---|---|---|---|
| Category composition | Rp+ | Rp− | Nrpa | Nrpb |
| Strong–strong (SS) | 79.6 (S) | 56.8 (S) | 64.1 (S) | 66.2 (S) |
| Tested first | 83.2 | 54.2 | 62.6 | 60.4 |
| Tested second | 75.9 | 59.3 | 65.6 | 71.9 |
| Strong–weak (SW) | (78.1 (S) | 47.9 (W) | 55.2 (S) | 44.2 (W) |
| Tested first | 78.1 | 52.1 | 56.2 | 46.8 |
| Tested second | 78.1 | 43.7 | 54.2 | 41.6 |
| Weak–strong (WS) | 66.2 (W) | 51.0 (S) | 48.9 (W) | 60.5 (S) |
| Tested first | 63.7 | 52.2 | 49.9 | 64.6 |
| Tested second | 68.7 | 49.9 | 47.9 | 56.3 |
| Weak–weak (WW) | 62.0 (W) | 42.2 (W) | 42.2 (W) | 33.4 (W) |
| Tested first | 58.4 | 43.7 | 40.6 | 32.3 |
| Tested second | 65.6 | 40.7 | 43.8 | 34.5 |

*Note.* Rp+ = practiced exemplars from practiced categories; Rp− = unpracticed exemplars from practiced categories; Nrpa and Nrpb = unpracticed exemplars from unpracticed categories. An S or a W in parentheses denotes the strength of the exemplars in that cell. Tested first or second = items tested in the first or second three positions of a category block. Comparisons of Rp− and Nrpb baseline items reflect impairment. Comparisons of Rp+ and Nrpa baseline items reflect facilitation.

absolute and proportional differences in facilitation and inhibition. The absolute facilitation owing to retrieval practice obtained for weak items was not different from that obtained for strong items, (Rp+) − (Nrp) = 64.1 − 45.6 = 18.5% and 78.9 − 59.7 = 19.2%, respectively, $F(1, 60) < 1$, reinforcing the conclusion that the difference in facilitation observed in Experiment 2 arose from the influence of ceiling effects on the recall of strong items. Contrary to expectation, weak exemplars also failed to show proportionally greater facilitation than strong exemplars (28.9% and 24.3%, respectively), $F(1, 60) < 1$, as in Experiment 1. Again, the failure for weak exemplars to exhibit greater facilitation than strong exemplars may reflect the fact that final recall performance underestimates facilitation due to retrieval practice (see Experiment 1). However, the strengthening-bias explanation proposed to account for the greater impairment for strong categories obtained in Experiment 2 is clearly not supported by the present results.

### *Final Recall Performance Averaged Across Output Position*

Except for the lower level of overall performance, the results of Experiment 3 were similar to those of Experiment 2. A significant main effect for category composition was obtained, $F(3, 60) = 8.2, p < .0001$, with the average recall of subjects in the SS condition (66.6%) being superior to the average recall of subjects in the SW (56.3%) and the WS (56.7%) conditions, $F(1, 60) = 7.1, p < .01$, and the recall of subjects in the latter two sets being superior to that of subjects in the WW conditions (44.9%), $F(1, 60) = 9.2$, $p < .01$. Thus, our manipulations of taxonomic frequency clearly had the desired impact on recall performance. Furthermore, as expected, planned comparisons revealed that retrieval practice improved overall recall of Rp+ items ($M = 71.5\%$) over Nrpa items ($M = 52.6\%$), $F(1, 60) = 53.0, p < .0001, MS_e = .043$, but, on the whole, did not reliably damage recall of the Rp− items ($M = 49.5\%$) relative to Nrpb items ($M = 51.1\%$), $F(1, 60) < 1$. Facilitation of practiced items did not interact with category composition whether the taxonomic strengths of the practiced items were contrasted (SS and SW vs. WS and WW = 19.2% vs. 18.5%) or whether the taxonomic strengths of the Rp− competitor items were contrasted (SS and WS vs. SW and WW = 16.4% vs. 21.4%), with $F(1, 60) < 1$ in all cases.

The crucial comparisons, however, regard interactions of inhibition with the levels of our category composition factor. In particular, the suppression hypothesis predicts greater impairment for conditions in which Rp− items were strong (SS and WS) than for those in which Rp− items were weak (SW and WW). This interaction was found to be significant, appearing when absolute impairment was considered, $F(1, 60) = 10.5$, $p < .01$, as well as when proportional impairment was considered, although the latter interaction was only marginally significant, $F(1, 60) = 3.2, p = .08$. Interestingly, the interaction resulted both from significant absolute inhibition in strong Rp− conditions, (Rp−) − (Nrpb) = 53.9 − 63.4 = −9.5%, $F(1, 60) = 7.6$, $p < .01$, and from marginally significant facilitation in weak Rp− conditions, 45.1 − 38.8 = +6.5%, $F(1, 60) = 3.3, p = .07$.

As can be seen in Table 5, which summarizes facilitation and impairment effects for Rp− and Rp+ items as a function of Rp+ and Rp− strength, there is little evidence that variations in the strength of Rp+ items modulated impairment of Rp− recall: The impairment to Rp− items when the Rp+ items were strong (−2.9%) was not significantly different from the impairment to Rp− items when the Rp+ items were weak (−0.3%), $F(1, 60) < 1$, failing to support the lateral inhibition hypothesis. Furthermore, the impairment to Rp− items was nonsignificant in both cases, presumably because the facilitatory and inhibitory effects on the recall of Rp− items as a function of Rp− strength cancelled each other out. The pattern of results presented in Table 5 implies that the variable modulating the degree of retrieval-induced forgetting is not the strength of the Rp+ item but the strength of the Rp− item, as predicted by the suppression hypothesis. Specifically, if nontarget competitors are strong, they are more likely to be inhibited than if they are weak, regardless of whether practiced items are strong or weak.

It is important to emphasize that the present findings replicate the complete absence of impairment that was observed for weak Rp− items in Experiment 2, despite variations in materials and testing procedure. Indeed, there is even some indication that weak Rp− items may profit from the practice of their competitors. There are several reasons why these surprising results cannot be explained by either the strengthening-bias and covert-retrieval hypothesis or the extraexperimental interference hypothesis. First, if strong Rp+ items received more strengthening, they should have displayed greater absolute and proportional facilitation with respect to their Nrp baseline than did the weak Rp+ items. As noted earlier, however, both the absolute and the proportional facilitation for strong and weak exemplars were statistically equivalent, and, if anything, evidenced proportionally greater facilitation for the weak Rp+ items. Furthermore, the impairment observed for Rp− items in the WS condition, in which the hypothetically less facilitated weak items were practiced, makes an explanation of the greater impairment for strong Rp− items in terms of less facilitation for weak Rp+ items unlikely. Second, if weak categories were less impaired because the difficulty of weak Rp+ items led subjects selectively to rehearse Rp− items, we should have observed (a) no impairment, and perhaps facilitation in the WS condition, and (b) substantial impairment in the SW condition. Because

Table 5
*Impairment of Rp− Items and Facilitation of Rp+ Items on a Category-Plus-Stem Cued-Recall Test as a Function of the Taxonomic Strength of Rp+ and Rp− Items in Experiment 3*

| Strength of Rp+ items | Strength of Rp− Items | |
|---|---|---|
| | Strong | Weak |
| Strong | −9.4 (+15.5) | +3.7 (+22.9) |
| Weak | −9.5 (+17.3) | +8.8 (+19.8) |
| *M* | −9.5 | +6.3 |

*Note.* Impairment = (Rp−) − (Nrp); facilitation = (Rp+) − (Nrp). Rp+ = practiced exemplars from practiced categories; Rp− = unpracticed exemplars from practiced categories; Nrp = unpracticed exemplars from unpracticed categories.

neither the impairment of strong nor the facilitation of weak Rp− items showed a significant effect of the strength of the practiced exemplar, $F(1, 60) < 1$ in both cases, biases in covert rehearsal cannot explain the present data. Finally, because the SW and WS conditions had the same extraexperimental exemplar set and because the Rp+ items in those conditions were strengthened to a proportionally equivalent degree, the lack of impairment in the SW condition (and probably in the WW condition as well) cannot be explained by the extraexperimental interference hypothesis. Thus, it appears that the failure of retrieval-based strengthening in the SW and WW conditions to impair Rp− items constitutes a genuine violation of the strength-dependence assumption. The implications of these findings for ratio-rule models are elaborated further in the General Discussion section.

We also examined the performance of strong and weak exemplars in our Nrp baseline conditions to determine whether they conformed to the patterns predicted by relative strength models. Ratio-rule models predict that strong exemplars in the SW and WS conditions should be recalled better than those in the SS condition because a strong item's relative strength is reduced in the latter case. Not only did we fail to observe this pattern, we observed what may be a trend in the opposite direction: As can be seen in Table 4, recall of strong exemplars in SS categories (65.2%) appeared to be better than the average recall of strong exemplars in the SW and WS categories (57.9%), although this was not significant, $F(1, 30) = 2.3$, $p = .14$. Similarly, weak exemplars in the WW condition should be recalled better than weak items in the WS or SW conditions. This trend also failed to occur, and the opposite pattern was suggested: The recall of weak exemplars in WW categories (37.8%) appeared to be worse than the average recall of weak exemplars in the SW and WS categories (46.6%), although this difference was only marginally significant, $F(1, 30) = 3.3$, $p = .08$. This pattern of results constitutes yet another violation of the strength-dependence assumption, contradicting the predictions of a ratio-rule model.

### *Impact of Testing Order on Final Recall Performance*

The most important testing-order finding of Experiment 3 was the replication of significant Rp− inhibition at different positions in the testing sequence. As illustrated in the rows labeled *Tested first* in Table 4, the recall of strong Rp− items was impaired when they were tested before Rp+ items. As in Experiment 2, the reliable impairment observed for strong Rp− items (SS and WS) did not vary with the position in which Rp− items were tested: (Nrpb1st) − (Rp−1st) = 9.3%; (Nrpb2nd) − (Rp−2nd) = 9.5%, with the interaction, $F(1, 60) < 1$. Nor did the greater impairment for strong Rp− items than for weak Rp− items interact with testing order, $F(1, 60) < 1$. Again, because Rp− items that are tested first are not contaminated by the potentially interfering effects of Rp+ output, we can attribute the impairment of strong Rp−1st items to effects enduring from the retrieval-practice phase. Thus, the finding of enduring inhibition was replicated.

As in Experiment 2, items recalled later in a category ($M = 56.1\%$) were not, in general, recalled worse than items recalled earlier in a category ($M = 56.2\%$). Unlike Experiment 2, however, testing order did not interact with our category composition factor, $F(3, 60) = 1.4$, $p > .2$, even when attention was restricted to only those conditions used in Experiment 2 (SS and WW), $F(1, 60) < 1$. Because the number of subjects in each condition ($n = 16$) was smaller than in the previous experiment ($n = 48$), and because there is considerable variability in the effects of testing order for both strong items (overall, four cells show impairment, three show facilitation, and one is a tie) and weak items (overall, four cells show impairment and four show facilitation), comparisons of individual cells are not likely to be meaningful. However, when all cells with strong and weak exemplars are considered (i.e., Rp+, Rp−, Nrpa, and Nrpb for all conditions), strong items tested first ($M = 63.9\%$) are no different than strong items tested last ($M = 63.9\%$), nor are weak items tested first ($M = 48.4\%$) different than weak items tested last ($M = 48.3\%$). The reasons for this failure to replicate the output interference of Experiment 2 are unclear.

In summary, the results of Experiment 3 replicated those of Experiment 2 in most major respects, including (a) the greater impairment for strong than for weak Rp− items; (b) the complete absence of impairment for weak Rp− items; and (c) the presence of Rp− impairment when Rp− items were tested before their Rp+ competitors. In addition, Experiment 3 demonstrated that the greater impairment for strong categories observed in Experiments 1 and 2 is attributable to a greater susceptibility of strong Rp− items to impairment, rather than to either a greater potency of strong Rp+ items as inhibitors or to the covert strengthening of weak Rp− items.

## General Discussion

Three general findings emerge from the current work. First, retrieving information repeatedly can impair recall performance on related information. In Experiment 1, retrieval practice on three members of a studied category, such as Fruit, improved recall performance for those items on a subsequent test but often at the cost of decreasing recall performance for the remaining three members. Experiments 2 and 3 replicated this impairment and generalized it to a category-plus-stem cued-recall test. Thus, the act of remembering can cause forgetting of semantically related material on a later recall test.

Second, the present experiments demonstrate that the negative effects of retrieval can endure well beyond the immediate context in which a competitor is retrieved. In all three experiments, the impairment of nonpracticed exemplars was still in evidence after the 20 min retention interval between retrieval practice and the final test. This finding contrasts with those from previous studies that focused exclusively on retrieval-based impairment within a single testing session (e.g., Blaxton & Neely, 1983; Brown, 1981; Dong, 1972; Roediger, 1973; Roediger & Schmidt, 1980; Smith, 1971, 1973). These previous studies did not address the durability of output interference, leaving it unclear whether output interference contributed to long-term forgetting or reflected a transient interference. The present finding demonstrates that the negative effects of retrieval are not restricted to a single output session and suggests that the reasons for this enduring quality are more complex than we anticipated.

Initially, we expected that impairment would occur after 20 min because the practice-based facilitation would persist, allowing practiced items to block unpracticed competitors. In Experiments 2 and 3, we studied these assumptions more closely by manipulating the output order of Rp+ and Rp− items at test. Interestingly, Rp− impairment still occurred when category-plus-stem cues (e.g., Fruit Or_____) were used to force subjects' output of Rp− items before Rp+ items. This result suggests that output interference at test cannot be the sole explanation of the Rp− impairment and that an additional inhibitory component persists throughout the 20-min retention interval. This impairment may be the first demonstration of inhibition at a long retention interval that cannot be explained by prior output of dominant items. Whatever the contributions of practice- and test-based sources of impairment may be, the present experiments show that retrieval is a significant factor contributing to long-lasting memory failure.

Finally, and unexpectedly, retrieval appears to have its greatest negative effects on items strongly associated to the current retrieval cue. In Experiment 1, recall of unpracticed members from strong-exemplar categories (e.g., Fruit Orange) suffered significantly more retrieval-induced forgetting than did recall of unpracticed members from weak-exemplar categories (e.g., Tree Hickory). This general pattern was replicated with the category-plus-stem cued-recall task of Experiments 2 and 3, except that unpracticed members of weak-exemplar categories were not simply less impaired than members of strong-exemplar categories, they were either unimpaired altogether or they were even facilitated by the retrieval of their competitors. Experiment 3 demonstrated that the strength of the unpracticed item, not the strength of the practiced item, had determined the impairment observed in Experiments 1 and 2: Strong competitors were impaired independently of the type of item that was practiced (strong or weak), whereas weak competitors were unimpaired by practice of those same items. These findings suggest the surprising conclusion that highly accessible items will be the most vulnerable to retrieval-induced forgetting.

When trying to explain why retrieval of some items has negative effects on other items, one is inevitably drawn to the significant facilitatory effects of retrieval practice as a potential cause. The intuition that strong items block the retrieval of weaker ones is compelling, even though the empirical justification for this intuition is not as strong as one might like. If the impairment observed at present related sensibly to the degree of strengthening, it would clearly support the strength-dependence assumption. In the next two sections, we argue that strength-dependent competition has difficulty accounting for the pattern of impairment across our experiments and that a retrieval-based suppression mechanism provides a better account. We then discuss relations of the present findings to research on retroactive interference, part-set cuing and the list-strength effect.

### *Strength-Dependent Competition*

The impairment of unpracticed category members might seem to result from the retrieval-based strengthening of their practiced companions. Indeed, the retrieval-practice procedure was designed to maximize this strengthening because the prediction of retrieval-induced forgetting was based on the strength-dependence assumption. Several of the present findings, however, lead one to question whether an item's recall probability is affected by the strength of its competitors.

The most compelling findings are summarized in Table 6, which displays the facilitatory and inhibitory effects of retrieval practice as a function of the strength of unpracticed and practiced competitors for all three experiments. The mean facilitation of Rp+ items, illustrated in the right column of Table 6, makes it clear that retrieval practice strengthened practiced items (average facilitation across all three experiments, $M = 17.7\%$). If this facilitation caused impairment by blocking access to Rp− items, we should have observed Rp− impairment whenever facilitation of Rp+ items was in evidence. Yet, the inhibitory effect of retrieval practice (left column) depended greatly on whether unpracticed items were weak exemplars (bottom left) or strong exemplars (top left). When Rp− items were weak, no impairment occurred (bottom left, averaged across experiments, $M = +2.7\%$; the impairment in Experiment 1 will be addressed in the *Suppression* section), even though their practiced companions were strongly facilitated (bottom right, $M = 20.7\%$). Furthermore, as shown in Row 8 of Table 6, recall of weak Rp− items remained unaffected ($M = 3.7\%$), even when their practiced competitors were already more accessible because they were strong exemplars of the category. In contrast, when Rp− items were strong, significant impairment occurred (top left, $M = -9.9\%$), even though their practiced competitors were no more, and possibly less, facilitated than the aforementioned practiced items (see top right, $M = 14.7\%$). This pattern of Rp− impairment across strong and weak exemplars was consistent across three experiments that varied in materials and testing procedures, and it was not influenced by the taxonomic strength of the practiced competitors (as can be seen by

Table 6
*Impairment (Rp−) − (Nrp) and Facilitation (Rp+) − (Nrp) Due to Retrieval Practice Across Experiments 1, 2, and 3 as a Function of the Taxonomic Strength of the Rp− Set and the Strength of the Rp+ Set*

| Strength of Rp− and strength of Rp+ | Exp. | *N* | Effect of retrieval practice: Impairment (Rp−) − (Nrp) | Effect of retrieval practice: Facilitation (Rp+) − (Nrp) |
|---|---|---|---|---|
| Strong items | | | −9.9 | +14.7 |
| Strong | 1 | 36 | −15.7*** | +25.0*** |
| Strong | 2 | 48 | −8.0** | +8.4** |
| Strong | 3 | 16 | −9.4** | +15.5*** |
| Weak | 3 | 16 | −9.4** | +17.3*** |
| Weak items | | | +2.7 | +20.7 |
| Weak | 1 | 36 | −6.3* | +21.9*** |
| Weak | 2 | 48 | +0.2 | +17.2*** |
| Weak | 3 | 16 | +8.8* | +19.8*** |
| Strong | 3 | 16 | +3.7 | +22.9*** |

*Note.* Rp+ = practiced exemplars from practiced categories; Rp− = unpracticed exemplars from practiced categories; Nrp = unpracticed exemplars from unpracticed categories.
*$p < .05$. **$p < .01$. ***$p < .001$.

comparing Row 3 vs. Row 4 and Row 7 vs. Row 8 of Table 6). It appears from these results that the strengthening of a competitor (whether defined in terms of taxonomic frequency or in terms of retrieval-based facilitation), though correlated with the events that lead to impairment, is not the cause of the effect; the critical variable is the strength of the unpracticed item.

The failure of strong competitors to impair recall is not restricted to the retrieval-practice manipulations summarized in Table 6. In Experiment 3, recall of baseline items (i.e., Nrp items) varying in taxonomic frequency showed a similar pattern. Neither the recall of strong nor the recall of weak Nrp exemplars decreased when strong competitors were substituted for weak ones: As can be seen in Table 4, recall of strong Nrp items in the SW and WS conditions (55.2 and 60.5, respectively; $M = 57.9$) was not different than recall of those same Nrp items in the SS condition (64.1 and 66.2, $M = 65.2$); similarly, recall of weak Nrp items in the WW condition (42.2 and 33.4, $M = 37.8$) was not different than recall of those same Nrp items in the SW or WS conditions (44.2 and 48.9, respectively, $M = 46.6$). Indeed, if there was any effect of adding strong competitors, it was positive, not negative. This pattern of results clearly violates the strength-dependence assumption. Even when differences in the relative strength of competitors were operationalized according to variations in taxonomic frequency (which did, in fact, result in highly significant differences in recall rates) rather than according to retrieval-based learning, the predicted strength-dependent competition effects failed to occur.

One might object that these failures of the strength-dependent competition predictions arise from the category-plus-stem testing procedure we used in Experiments 2 and 3. In this procedure, subjects may have treated the category and the exemplar stem as a joint retrieval cue, focusing memory search to category exemplars beginning with that stem. Because all exemplar stems were constructed to be unique in the category (and, in most cases, in the experiment), such a search would exclude Rp+ items from the search set. If the stem-completion testing procedure eliminated Rp+ items from the search set, it should not be surprising (from the standpoint of relative strength models) to find that Rp− items were unimpaired by the greater strengths of Rp+ items. The difficulty with this reasoning is that although it may account for the lack of impairment for weak Rp− items in Experiments 2 and 3, it leaves the impairment of strong Rp− items in those same experiments unexplained. Thus, the results of Experiments 2 and 3 imply either that (a) the stem-completion testing procedure eliminates the blocking predicted by strength-dependent competition and that a mechanism other than blocking is contributing to the retrieval-induced forgetting observed for strong items or that (b) impairment is not a necessary consequence of the strengthening of competitors.

But even if we focused exclusively on the category-cued free-recall testing procedure of Experiment 1, the relationship between the degree of impairment and the degree of facilitation does not fit the strength-dependent competition model. In Experiment 1, as in Experiments 2 and 3, both absolute and proportional impairment were greater for strong-exemplar categories than for weak-exemplar categories. Yet, the opposite pattern should be true according to strength-dependent competition models (augmented with fairly common learning assumptions). Greater proportional impairment for weak categories is predicted because retrieval practice should increase the associative strength of weaker items to a proportionally greater extent. Although this assumption appears justified, the difference in facilitation for strong and weak items was not statistically reliable; nonetheless, even with proportionally equivalent facilitation, impairment should not be greater for strong-exemplar categories (as shown in Appendix A), as it was found to be in all three experiments. As argued in the discussions of Experiments 1 and 3, these findings cannot be explained by such factors as covert rehearsal or biases in the strengthening of practiced items in strong categories. Even when we focus on the category-cued free-recall procedure of Experiment 1, the pattern of impairment does not relate sensibly to the strengthening of competitors.

Thus, although it is compelling to attribute the impairment of unpracticed exemplars to the strengthening of their practiced competitors, this approach appears to be inadequate, if not mistaken. The facilitation of practiced items does not relate in any orderly way to the degree of impairment; rather, the strength of unpracticed exemplars is the best predictor of their own impairment. When trying to explain these failures of strength-dependent competition, one must keep in mind that retrieval is functionally distinct from other strengthening procedures such as multiple presentations of an item (see, e.g., Blaxton & Neely, 1983, for an informative contrast of these procedures). In particular, retrieval involves the search for an item in memory and the discrimination of that target item from among a set of partial matches. Thus, when strengthening occurs through retrieval, as opposed to other strengthening methods in which the full item is presented to subjects, the activation of these partial matches may have significant implications for success on later retrieval tasks. These special qualities of retrieval led us to consider the contribution of suppression in the production of retrieval-induced forgetting.

### *Suppression*

The failure of strength-dependent competition to account for the pattern of results obtained in the present research argues for some other mechanism associated with retrieval that causes forgetting. One possibility is that the observed impairment reflects the inhibition of the affected items, as suggested in some modified spreading-activation theories of memory retrieval. In these theories, presenting a cue should activate all associated responses in parallel; this initial spread of activation may then need to be focused to isolate the target response from interfering competitors. Although focusing can be achieved in various ways, inhibition is often thought to subserve this function (Blaxton & Neely, 1983; Carr & Dagenbach, 1990; Gernsbacher, Barner, & Faust, 1990; Keele & Neill, 1978; Martindale, 1981; Neely & Durgunoglu, 1985; Neill & Westberry, 1987; Walley & Weiden, 1973). If nontarget items are inhibited during retrieval of target exemplars, subsequent recall of those inhibited items should be impaired. This inhibition may be sufficient to produce retrieval-induced forgetting.

An inhibitory theory of retrieval-induced forgetting can

account for several important features of the present findings. First, it offers an explanation for the greater impairment of strong items observed in all three experiments (Table 6, top left). Strong Rp− items should be more impaired because their greater associative strength should lead them to interfere more with the retrieval practice of their competitors, and this greater interference should, in turn, render those strong items more vulnerable to inhibition. In contrast, weak Rp− items may remain totally unimpaired (Table 6, lower left) or may even be facilitated by their initial activation (Table 6, Row 7), provided that their level of activation does not interfere with the retrieval practice of their competitors. Second, the impairment of Rp− items that were tested before Rp+ items (i.e., Rp−1st items) in Experiments 2 and 3 would be explained: Impaired recall of Rp−1st items would reflect inhibition that endured from the prior retrieval-practice phase, as suggested previously. Finally, the many failures of the strength of a competitor to affect recall probability can be explained if we assume that a competitor's strength decreases retrieval speed without affecting retrieval probability. The mere presence of Rp+ items (or strong Nrp exemplars) in memory would then slow retrieval of Rp− items (or Nrp competitors) on the final test, but should not prevent their recall. The recall of those Rp− items, however, should be impaired on the final test if their strength had impeded the retrieval practice of their practiced companions.

Although inhibitory processes can account for the present findings better than can strength-dependent competition, some aspects of the results are inconsistent with both hypotheses. First, the same strong items exhibited output interference (Strong 1st − Strong 2nd = 4.1%) in Experiment 2, but did not in Experiment 3 (0.0%).[3] Second, Rp+ items never showed output interference in Experiments 2 or 3 (Rp+1st − Rp+2nd = 0.6%, averaged across strong and weak items for both experiments). According to the inhibition hypothesis, prior retrieval of category members at final test should inhibit the remaining strong items (whether those items are Rp+ items or strong exemplars); according to strength-dependent competition, these prior retrievals should strengthen the retrieved exemplars, blocking access to subsequent items. It is possible that a single retrieval of each item on the final test may not be sufficient to produce the expectation of reliable differences in recall for either theory. Whatever the proper explanation may be, these inconsistencies afflict both theories. Given this observation, the results are most consistent with a model in which inhibition is used to overcome interference from competing items.

The present results support some inhibitory theories of retrieval-induced forgetting more than others. Many theories assume that the degree to which a target inhibits competitors depends on the strength of that target item. For instance, in their recent center-surround theory of semantic memory retrieval, Carr and Dagenbach (1990) proposed that inhibition enhances the discriminability of weakly activated targets that may be overcome by the activation of competing codes. In this theory, the weaker the target item, the more inhibited competitors should be (with the strength of competitors held constant), even when the target is not successfully retrieved. Other formulations of lateral inhibition might assert that strong targets produce more, not less, inhibition than weak targets. If highly associated targets become more active when presentation of the cue occurs and if increases in target activation lead to increases in the inhibition that is spread laterally to competitors, strong exemplars should cause more inhibition than weak exemplars. Both approaches assume that the severity of inhibition relates to the strength of the target item, yet the findings of Experiment 3 suggest that this assumption may not be correct: The degree of impairment suffered by Rp− items did not depend on whether strong or weak category exemplars were practiced (see Rows 3, 4, 7, and 8 in Table 6). The failure for impairment to be related to target (Rp+) strength suggests that inhibition may not be an automatic process mediated by the representations of competing target items. The results are consistent, however, with a process of active suppression, applied directly to competing items to the extent that those items interfere with task demands (see, e.g., Blaxton & Neely, 1983; Keele & Neill, 1978; Neely & Durgunoglu, 1985; Neill & Westberry, 1987).

Although suppression provides the best single account of our data, it must be emphasized that this hypothesis is not incompatible with strength-dependent competition. Indeed, there is some indirect evidence for a two-process interpretation of retrieval-induced forgetting. Weak Rp− items exhibited small, but reliable recall impairment in Experiment 1 but did not in Experiments 2 and 3, whereas strong Rp− items exhibited reliable impairment in all three experiments. An interesting two-process interpretation of this pattern of impairment is as follows: If the stem-completion testing procedure used in Experiments 2 and 3 eliminated strength-dependent competition (as suggested previously), the lack of impairment for weak Rp− items can be explained, but the impairment for strong Rp− items in those same experiments cannot. If this testing procedure remained sensitive to suppression, however, then the results of Experiments 2 and 3 show that strong items suffer suppression but weak items do not. This interpretation suggests that the impairment of weak Rp− items in the category-cued free-recall test of Experiment 1 may have arisen entirely from strength-dependent competition. Whatever the contributions of strength-dependent competition, however, the present results argue that an active suppression mechanism causes much of the long-lasting retrieval-induced forgetting in the retrieval-practice paradigm.

### *Relation to Other Empirical Findings*

Retrieval-induced forgetting resembles several other phenomena in which enhancing recall of some items impairs memory for related information. For example, our findings resemble both retroactive interference effects and part-set cuing inhibition to the extent that retrieval practice is similar to repeated learning trials and cuing, respectively. Despite these similarities, the pattern of impairment in the present experi-

[3] Although the present experiment did not obtain output interference, subsequent experiments with the same materials and procedure have obtained sizable output interference effects (8 to 10%). The reason for the failure to find such effects in the present Experiment 3 are unclear.

ments argues that retrieval-based learning is not the primary cause of retrieval-induced forgetting; rather, impairment appears to result from an active suppression of unpracticed exemplars. This interpretation raises the possibility that the commonly assumed link between strengthening and impairment in the aforementioned phenomena has been overstated or perhaps even misinterpreted. In this section, we show that these and other findings that support a causal link between strengthening and impairment stem from paradigms that confound strengthening and retrieval-induced forgetting. Thus, what appears to be strength-dependent competition may often be retrieval-based suppression. Although this general argument applies to many phenomena, we focus on three for the purpose of illustration: retroactive interference, part-set cuing inhibition, and the list-strength effect.

### *Retroactive Interference*

Perhaps nowhere has the apparent connection between strengthening and impairment been more vividly demonstrated than in a classic study of retroactive interference by Barnes and Underwood (1959). In their study, Barnes and Underwood showed that recall for items from a first list of paired associates systematically decreased with increases in the number of learning trials administered on a second list of associates. Decreases in the recall of first-list responses correlated well with increases in the recall for second-list responses, suggesting that strengthening second-list items caused the decrease in recall of their first-list competitors. This negative correlation between second- and first-list recall has been successfully modeled with strength-dependent competition mechanisms (Mensink & Raaijmakers, 1988), without proposing the additional unlearning process included in both the classical two-factor theory of interference (Melton & Irwin, 1940) and in modern connectionist learning approaches (see, e.g., Lewandowsky, 1991; Sloman & Rumelhart, 1992). Despite the success of the strength approach in modeling these data, the present findings question whether the conditions of strength-dependent competition are sufficient or even necessary to produce retroactive interference.

Although it is compelling to focus on the orderly relationship between the degree of strengthening on second-list responses and the amount of retroactive interference, an alternative view arises when we consider that second-list responses in Barnes and Underwood's (1959) study were strengthened by the method of anticipation. In this method, each cycle through a learning list entails two events for each paired associate: (a) presentation of that associate's stimulus as a cue, to which subjects must recall or "anticipate" the associated response and then (b) presentation of the response as feedback. By cuing recall in this manner, Barnes and Underwood effectively gave subjects retrieval practice on the second list. If the present analysis of retrieval practice is correct, repeated suppression of first-list responses during these trials may have caused the observed increases in retroactive interference rather than (or perhaps, in addition to) strengthening of second-list competitors. This account of retroactive interference effects parallels the classical notion of unlearning (Melton & Irwin, 1940) in its emphasis on intrusions of first-list responses during tests of the second list; the suppression account, however, attributes impairment to inhibition of the first-list target items rather than to weakening of their cue–target associations (see the response-set suppression hypothesis of Postman et al., 1968, for a similar emphasis on response inhibition). The important point, for present purposes, is that theoretical treatments of interference data that focus exclusively on the strengthening of second-list responses greatly understate the role of retrieval-induced forgetting. Indeed, if suppression contributes to retroactive interference as suggested by the present data, it becomes difficult to assess whether strengthening by itself is sufficient to produce impaired recall.

### *Part-Set Cuing Inhibition*

A second illustration of the connection between strengthening and impairment was provided in a study of part-set cuing inhibition by Rundus (1973). In this experiment, subjects studied categorized word lists and then recalled items from each category with varying numbers of exemplars provided as cues. Rundus found that as the number of cues increased from zero to four, recall of the remaining noncue items decreased. Based on the assumption that cue exemplars were strengthened by their presentation at test, Rundus concluded that the decline in recall of noncue items was caused by the strengthening of their cued competitors. Several replications of this basic finding (see, e.g., Roediger, 1973, and Watkins, 1975) have supported Rundus's interpretation, although manipulations of cue type that should induce variations in strengthening (e.g., taxonomic frequency of exemplars; intralist vs. extralist exemplars) have failed to cause the predicted variations in impairment (Basden et al., 1977; Karchmer & Winograd, 1971; Watkins, 1975). Nonetheless, Rundus's strength approach retains its popularity because it accounts for a range of part-set cuing findings (see Nickerson, 1984, and Roediger & Neely, 1982, for reviews).

Although the robust relationship between the number of cues and impairment supports strength-dependent competition, an alternative interpretation arises when we consider that strengthening cues often causes subjects to retrieve those items before noncues. Cue items may be retrieved before noncues either overtly, if both cues and noncues are to be recalled (see, e.g., Karchmer & Winograd, 1971; Roediger et al., 1977 for data on this point), or covertly during attempts to recall noncues, as is often presumed to occur in "blocking" models of part-set cuing inhibition (see, e.g., Rundus, 1973). When cue items are retrieved early, noncues should suffer more retrieval-induced forgetting than the corresponding items for control subjects for whom recall order has not been biased. As more cues are provided, more items should be retrieved prior to noncues, further impairing noncue recall. Although decreases in noncue performance may be caused by strengthening of cue items during their covert retrieval—a possibility noted by both Roediger (1974) and Rundus (1973), the present analysis suggests that noncue impairment reflects retrieval-based suppression. This interpretation receives support from a study by Blaxton and Neely (1983) in which speeded recall of several prime exemplars from a semantic category slowed subsequent

recall of a target exemplar, whereas speeded naming of those same primes facilitated target recall. If strengthening were sufficient to impair competing items, then both the recall and presentation of prime items should have impaired retrieval of target exemplars. Thus, cuing by itself may not impair recall; rather, the strengthening of cues may indirectly impair recall to the extent that early retrieval of cue items suppresses noncues at the time of test.

### *List-Strength Effect*

A final illustration of the apparent relationship between strengthening and impairment comes from a recent series of studies on what has been termed the list-strength effect by Ratcliff et al. (1990). The list-strength effect can be thought of as an analog to the well-known list-length effect, except that performance on a target item (or set of items) is predicted to decrease from the strengthening of other list members rather than from the addition of new list members. To test this prediction, Ratcliff et al. developed the mixed–pure paradigm, the goal of which was to show that strengthening one half of a list of words would both (a) impair performance on the remaining nonstrengthened list-half to a greater extent than would be the case were the words to be on a list in which no items were strengthened (i.e., a pure–weak list) and (b) facilitate performance on the strengthened list-half to a greater extent than would be the case were the words to be on a list in which all items were strengthened (i.e., a pure–strong list). Strengthening may be accomplished either by increasing the exposure time or the number of repetitions of the to-be-strengthened items, and either free recall, cued recall, or recognition memory can be tested. In a series of experiments using this paradigm, Ratcliff et al. found reliable list-strength effects in free recall, small and inconsistent effects in cued recall, and either no effect or reverse effects in recognition memory. Although the authors' interpretation of their entire pattern of results involved more than strength-dependent competition, this factor was thought to be crucial in producing the observed free- and cued-recall effects.

Two points should be made concerning Ratcliff et al.'s (1990) findings as evidence for the relationship between strengthening and impairment. First, although the authors successfully demonstrated an overall list-strength effect in free recall, the component of their data that produced this effect was not impairment of the weak-list half: The weak half of the study list was impaired by 2.7%, even though the remainder of the list was strengthened by 25% (i.e., relative to a pure–weak baseline, see Ratcliff et al., 1990, p. 172). Rather, the significant list-strength effect in free recall was produced by the 8% advantage of strong items in a mixed list over strong items in a pure–strong list (i.e., part "(b)" of the above list-strength prediction). Second, even the small amount of impairment that did occur in free recall cannot be confidently attributed to strength-dependent competition because Ratcliff et al.'s free-recall measure suffers from the same output-order bias present in studies of part-set cuing inhibition. If strengthened items were retrieved before nonstrengthened items, retrieval-based suppression may have occurred. When such output-order biases were eliminated, as was the case in their cued-recall experiments, impairment of weak items disappeared entirely (in Experiment 3, there was 0.0% impairment, despite 21.2% facilitation of strong items; in Experiment 6, there was 0.7% impairment, despite 27.2% facilitation). Thus the existing data on the list-strength effect provide no support for the relation between strengthening and impairment.

### *Concluding Remarks*

Although previous work has demonstrated the negative side effects of retrieval, these effects have received surprisingly little attention in modern theories of interference. The relative neglect of these phenomena may stem from two factors. First, retrieval-induced forgetting resembles other varieties of forgetting in which facilitating recall of some items impairs memory performance on related competitors. Because retrieval clearly facilitates those items that are retrieved, it is tempting to reduce the associated impairment of related items to strength-dependent competition. Second, the characterization of retrieval-induced forgetting as output interference may have hampered generalization of the phenomenon from the empirical context in which it was initially investigated. Indeed, the term *output interference* connotes a fleeting source of interference, muddying measures of recall in list-learning experiments. Together, these factors may have discouraged the separate study of retrieval-induced forgetting.

The present research has stressed the key role that retrieval may play in producing long-lasting forgetting. Our findings show that forgetting due to retrieval can last for at least 20 min, afflicting what we know the best, the most severely. Furthermore, the pattern of impairment in the present experiments suggests that the reduction of retrieval-induced forgetting to strength-dependent competition, though parsimonious, has been misleading. Though strengthening correlates with impairment, it may not, by itself, be the cause of forgetting; rather, impairment may instead reflect the negative side effects of a suppression process that assists in the resolution of retrieval competition. If this hypothesis is correct, it suggests that the recall impairments observed in other paradigms in which the effects of strengthening have not been adequately separated from the effects of retrieval-induced forgetting (e.g., retroactive interference, part-set cuing paradigms) may actually reflect retrieval-based suppression rather than strength-dependent competition. Thus, the contrary reduction may be possible: Strength-dependent competition may reflect the mechanisms of retrieval-induced forgetting. Regardless of how the theoretical interpretation of these effects evolves, the present research illustrates that retrieval can be a cause of long-lasting forgetting. The ubiquity of retrieval processes in our daily cognitive experience may render the mere use of "what we know" the most common source of fluctuation in the accessibility of our knowledge.

## References

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463–470.

Anderson, J. R. (1976). *Language, memory and thought.* Hillsdale, NJ: Erlbaum.

Arbuckle, T. Y. (1967). Differential retention of individual paired

associates within an RTT "learning" trial. *Journal of Experimental Psychology, 74,* 443–451.

Baddeley, A. D. (1982). Domains of recollection. *Psychological Review, 89,* 708–729.

Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology, 58,* 95–105.

Basden, D. R., Basden, B. H., & Galloway, B. C. (1977). Inhibition with part-list cuing: Some tests of the item strength hypothesis. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 100–108.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms [Monograph]. *Journal of Experimental Psychology, 80,* 1–46.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Blaxton, T. A., & Neely, J. H. (1983). Inhibition from semantically related primes: Evidence of a category-specific inhibition. *Memory & Cognition, 11,* 500–510.

Brown, A. S. (1981). Inhibition in cued retrieval. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 204–215.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin, 109,* 204–223.

Brown, A. S., Whiteman, S. L., Cattoi, R. J., & Bradley, C. K. (1985). Associative strength level and retrieval inhibition in semantic memory. *American Journal of Psychology, 98,* 433–447.

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language, 30,* 542–579.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning.* New York: Wiley.

Carr, T. H., & Dagenbach, D. (1990). Semantic priming and repetition priming from masked words: Evidence for a center-surround attentional mechanism in perceptual recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 341–350.

Dagenbach, D., Carr, T. H., & Barnhardt, T. M. (1990). Inhibitory semantic priming of lexical decisions due to failure to retrieve weakly activated codes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 328–340.

DaPolito, F. J. (1966). *Proactive effects with independent retrieval of competing responses.* Unpublished doctoral dissertation, Indiana University.

Delprato, D. J. (1972). Pair-specific effects in retroactive inhibition. *Journal of Verbal Learning and Verbal Behavior, 11,* 566–572.

Dong, T. (1972). Cued partial recall of categorized words. *Journal of Experimental Psychology, 93,* 123–129.

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1,* 213–216.

Gernsbacher, M. A., Barner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 430–445.

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

Greeno, J. G., James, C. T., DaPolito, F., & Polson, P. G. (1978). *Associative learning: A cognitive analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562–567.

Jones, G. V. (1989). Back to Woodworth: Role of interlopers in the tip-of-the-tongue phenomenon. *Memory & Cognition, 17,* 69–76.

Karchmer, N. A., & Winograd, E. (1971). The effects of studying a subset of familiar items on recall of the remaining items: The John Brown effect. *Psychonomic Science, 25,* 224–225.

Keele, S. W., & Neill, W. T. (1978). Mechanisms of attention. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of Perception,* (Vol. 9, pp. 3–47). New York: Academic Press.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Skykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Lewandowsky, S. (1991). Gradual unlearning and catastrophic interference: A comparison of distributed architectures. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays in honor of Bennet B. Murdock* (pp. 445–476). Hillsdale, NJ: Erlbaum.

Loftus, E. F. (1973). Activation of semantic memory. *American Journal of Psychology, 86,* 331–337.

Loftus, G. R., & Loftus, E. F. (1974). The influence of one memory retrieval on a subsequent memory retrieval. *Memory & Cognition, 2,* 467–471.

Marshall, G. R., & Cofer, C. N. (1970). Single-word free association norms for 328 responses from the Connecticut cultural norms for verbal items in categories. In L. Postman & G. Keppel (Eds.), *Norms of word association* (pp. 321–360). New York: Academic Press.

Martin, E. (1971). Verbal learning theory and independent retrieval phenomena. *Psychological Review, 78,* 314–332.

Martindale, C. (1981). *Cognition and consciousness.* Homewood, Ill: Dorsey Press.

McGeoch, J. A. (1936). Studies in retroactive inhibition: VII. Retroactive inhibition as a function of the length and frequency of presentation of the interpolated lists. *Journal of Experimental Psychology, 19,* 674–693.

Melton, A. W., & Irwin, J. M. (1940). The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *American Journal of Psychology, 3,* 173–203.

Mensink, G. J. M., & Raaijmakers, J. G. W. (1988). A model of interference and forgetting. *Psychological Review, 95,* 434–455.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitory and inhibitory processes. *Memory & Cognition, 4,* 648–654.

Neely, J. H., & Durgunoglu, A. Y. (1985). Dissociative episodic and semantic priming effects in episodic recognition and lexical decision tasks. *Journal of Memory and Language, 24,* 466–489.

Neely, J. H., Schmidt, S. R., & Roediger, H. L., III. (1983). Inhibition from related primes in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 196–211.

Neill, W. T., & Westberry, R. L. (1987). Selective attention and the suppression of cognitive noise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 327–334.

Nickerson, R. S. (1984). Retrieval inhibition from part-set cuing: A persisting enigma in memory research. *Memory & Cognition, 12,* 531–552.

Postman, L., & Stark, K. (1969). The role of response availability in transfer and interference. *Journal of Experimental Psychology, 79,* 168–177.

Postman, L., Stark, K., & Fraser, J. (1968). Temporal changes in interference. *Journal of Verbal Learning and Verbal Behavior, 7,* 672–694.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178.

Reason, J. T., & Lucas, D. (1984). Using cognitive diaries to investigate naturally occurring memory blocks. In J. E. Harris & P. E.

Morris (Eds.), *Everyday memory actions and absent-mindedness* (pp. 53–70). London: Academic Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 93,* 318–339.

Roediger, H. L., III. (1973). Inhibition in recall from cueing with recall targets. *Journal of Verbal Learning and Verbal Behavior, 12,* 644–657.

Roediger, H. L., III. (1974). Inhibiting effects of recall. *Memory & Cognition, 2,* 261–269.

Roediger, H. L., III. (1978). Recall as a self-limiting process. *Memory & Cognition, 6,* 54–63.

Roediger, H. L., III, & Neely, J. H. (1982). Retrieval blocks in episodic and semantic memory. *Canadian Journal of Psychology, 36*(2), 213–242.

Roediger, H. L., III, & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired associate lists. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 91–105.

Roediger, H. L., III, Stellon, C. C., & Tulving, E. (1977). Inhibition from part-list cues and rate of recall. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 174–188.

Rundus, D. (1973). Negative effects of using list items as retrieval cues. *Journal of Verbal Learning and Verbal Behavior, 12,* 43–50.

Shapiro, S. I., & Palermo, D. S. (1970). Conceptual organization and class membership: Normative data for representatives of 100 categories. *Psychonomic Monograph Supplements, 3* (11, Whole No. 43).

Slamecka, N. J. (1975). Intralist cueing of recognition. *Journal of Verbal Learning and Verbal Behavior, 14,* 630–637.

Sloman, S. A., Bower, G. H., & Roher, D. (1991). Congruency effects in part-list cuing inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 974–982.

Sloman, S. A., & Rumelhart, D. E. (1992). Reducing interference in distributed memories through episodic gating. In A. Healy, S. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 227–248). Hillsdale, NJ: Erlbaum.

Smith, A. D. (1971). Output interference and organized recall from long-term memory. *Journal of Verbal Learning and Verbal Behavior, 10,* 400–408.

Smith, A. D. (1973). Input order and output interference in organized recall. *Journal of Experimental Psychology, 100,* 147–150.

Smith, A. D., D'Agostino, P. R., & Reid, L. S. (1970). Output interference in long-term memory. *Canadian Journal of Psychology, 24,* 85–87.

Solso, R. L., & Juel, C. L. (1980). Positional frequency and versatility of bigrams for two- through nine-letter English words. *Behavior Research Methods and Instrumentation, 12,* 297–343.

Todres, A. K., & Watkins, M. J. (1981). A part-set cuing effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 91–99.

Tulving, E., & Arbuckle, T. Y. (1963). Sources of intratrial interference in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 1,* 321–334.

Tulving, E., & Arbuckle, T. Y. (1966). Input and output interference in short-term associative memory. *Journal of Experimental Psychology, 72,* 89–104.

Walley, R. E., & Weiden, T. D. (1973). Lateral inhibition and cognitive masking: A neuropsychological theory of attention. *Psychological Review, 80,* 284–302.

Warren, R. E. (1977). Time and the spread of activation in memory. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 458–466.

Watkins, M. J. (1975). Inhibition in recall with extralist "cues." *Journal of Verbal Learning and Verbal Behavior, 14,* 294–303.

Watkins, M. (1978). Engrams as cuegrams and forgetting as cue-overload: A cueing approach to the structure of memory. In C. R. Puff (Ed.), *The structure of memory* (pp. 347–372). New York: Academic Press.

*Webster's new collegiate dictionary.* (1980). Springfield, MA: G. & C. Merriam.

Woodworth, R. S. (1938). *Experimental psychology.* New York: Henry Holt.

# Appendix A

## Numerical Examples of a Ratio-Rule Model

We provide several numerical examples of ratio-rule predictions for the retrieval-practice paradigm. First, we show how the simplest formulation of the ratio rule predicts facilitation and impairment. We then extend the basic model to derive predictions for our taxonomic frequency manipulation.

### Basic Model and an Example

Assume that our categories are represented as a set of exemplars, each with a univalent association to the category cue. The simplest ratio-rule equation for this representation would then express the probability of recalling an exemplar, given a category cue, in the following form:

$$P(E1 \mid C1) = S(C1, E1)/\text{Sum}(S(C1, Ex))$$

In this equation, E1 is a particular exemplar; C1 is a particular category; and S(C1, E1) is the associative strength between category C1 and E1. Thus, the probability of recalling a particular exemplar, E1, is governed by the ratio of that exemplar's associative strength to the category cue, to the summed strengths of association of all exemplars (*Ex*) to that cue.

To see why this equation predicts facilitation for practiced exemplars and impairment for unpracticed exemplars, consider a simple four-member category, each exemplar having a cue-item associative strength of .2. The probability of recalling an item from this set would then be proportional to the ratio of its own strength of association to the cue to those of all competitors' strengths [.2/(.2 + .2 + .2 + .2) = .25]. If retrieval practice on two items from this set increased their associative strengths, say, to .3, then for those two practiced items we should observe facilitation [.3/(.2 + .2 + .3 + .3) = .3]; however, that same increase should result in impairment for the two items of that set that were not practiced [.2/(.2 + .2 + .3 + .3) = .2].

### Extended Model With Examples

Because the basic model, as currently specified, incorrectly predicts equal recall for items from strong and weak sets [e.g., strong: .4/(.4 + .4 + .4 + .4) = .25, weak: .2/(.2 + .2 + .2 + .2) = .25], it must be modified so that recall probability is dependent on an item's absolute strength as well as its relative strength. One way in which this goal can be accomplished is to distinguish between *trace-access* probability and *response-recovery* probability, the former governed by the target item's relative strength and the latter by its absolute cue-target strength (see, e.g., Raaijmakers & Shiffrin, 1981). Thus, recall probability for a strong item would be its trace-access probability multiplied by its response-recovery probability, which would result in greater recall for items from strong sets than for items from weak sets (e.g., from the previous example, .4 and .2 might be recovery probabilities, yielding .25 × .4 = .10 vs. .25 × .2 = .05, for strong and weak sets, respectively).

To make predictions about the relative impairment for strong and weak sets, we must specify both how retrieval practice increases cue–target associative strengths across strong and weak sets and how recovery probabilities differ across these sets. To simplify the analysis, first suppose that retrieval practice increases cue–target associative strengths to a proportionally equivalent degree across strong and weak sets. For example, an item in a four-item strong set having an initial strength of .4 might be incremented by 50% to .6, in which case an item from a weak set having an initial strength of .2 would be incremented to .3. Given proportionally equivalent strengthening for items in strong and weak sets, the reduction in *target accessibility* would be the same for unpracticed items in either set (e.g., for the strong set, Nrp − Rp− is: [.4/(.4 + .4 + .4 + .4)] − [.4/(.4 + .4 + .4 + .6)] = .03; for the weak set: [.2/(.2 + .2 + .2 + .2)] − [.2/(.2 + .2 + .2 + .3)] = .03). Superior recovery probabilities for items in strong sets, when multiplied by a strong item's target-access probability, would increase the absolute recall impairment expected for strong sets above that expected for weak sets (deficit in strong-item recall = [.25 × .4] − [.22 × .4] = .012; deficit in weak-item recall = [.25 × .2] − [.22 × .2] = .006). However, regardless of the magnitude of the difference in recovery probabilities across these sets, impairment for each set relative to its baseline should be proportionally equivalent (for strong items, proportional impairment = .012/[.25 × .4] = .12; for weak items, .006/[.25 × .2] = .12).

If we revise the somewhat unrealistic assumption that learning rates are proportionally equivalent across strong and weak items by assuming that items increase by the same constant amount (e.g., retrieval practice results in an increment of .1, regardless of an item's existing strength), or that growth in strength is a negatively accelerated function of current strength (as would be the case with linear operator models of learning, e.g., Bush & Mosteller, 1955; Rescorla & Wagner, 1972), the proportional impairment should be less for strong items than for weak items. This outcome obtains because weak items will increase in strength to a proportionally greater degree than strong items. Because we know that proportionally equivalent strengthening leads to proportionally equivalent impairment, proportionally greater facilitation for weak categories should lead to proportionally greater impairment for weak items.

### Extended Model With Extraexperimental Exemplars

Suppose that each category has four strong and four weak exemplars and that four are presented in the experiment and four remain as extraexperimental exemplars. Suppose, also, that strong and weak exemplars begin with extraexperimental strengths of .2 and .1, respectively, which are then incremented to .4 and .2 respectively upon their presentation in the study list.[A1] With these assumptions, the four category types in Experiment 3 can be represented with sets of eight strengths—four experimental and four extraexperimental strengths: SS = (.4, .4, .4, .4 | .1, .1, .1, .1); SW = (.4, .4, .2, .2 | .2, .2, .1, .1); WS = (.2, .2, .4, .4 | .2, .2, .1, .1); and WW = (.2, .2, .2, .2 | .2, .2, .2, .2). Note that the SS and WW category types vary in the strengths of their respective extraexperimental items, whereas the SW and WS category types do not.

Under these assumptions, the ratio rule predicts that impairment for strong categories should be proportionally greater than impairment for weak categories. To see this, suppose that two items in each SS and WW category are strengthened by 50% of their original

[A1] Note that this example assumes that the learning rates for strong and weak exemplars are proportionally equivalent, as discussed in the previous section of Appendix A. Although this assumption is not reasonable given the wealth of data showing that learning rate is a negatively accelerating function of prior strength, this learning assumption is the one that is most consistent with the present pattern of facilitation for Rp+ items across strong and weak categories. Without this particular learning rate assumption, it is unclear whether the ratio-rule model could account for the greater impairment of strong-exemplar categories in the manner suggested in this section.

strengths; that is, to .6 and .3, respectively. The probability of recalling an Rp− item from a strong category would then become .4/2.4 × .4, or .0627, whereas the probability of recalling a weak Rp− item would then become .2/1.8 × .2, or .0222. Relative to the baseline for strong (.08) and weak (.025) categories, strong and weak Rp− items would be impaired by .0173 and .0028 respectively. Thus, absolute impairment for strong categories would clearly be greater than that for weak categories. However, proportional impairment for strong categories (.0173/.08 = .216) would also be greater than proportional impairment for weak categories (.0028/.025 = .112). Thus, the relative impairment for strong and weak categories would depend on the composition of the extraexperimental set, given that we assume that subjects do not use experimental context as a retrieval cue to restrict memory search.

## Appendix B

### Categories and Exemplars Used in Experiments 1 and 2, Divided Into the Four Practice Counterbalancing Sets (A1, A2, B1, B2) and Sorted by Category Composition (Strong or Weak)

| Category | Exemplar Set 1 | Exemplar Set 2 |
|---|---|---|
| Set A: Strong | | |
| Fruits | Orange, nectarine, pineapple | Banana, cantaloupe, lemon |
| Leather | Saddle, gloves, wallet | Shoes, belt, purse |
| Set A: Weak | | |
| Trees | Palm, hickory, willow | Poplar, sequoia, ash |
| Professions | Tailor, florist, farmer | Critic, grocer, clerk |
| Set B: Strong | | |
| Drinks | Bourbon, scotch, tequila | Brandy, gin, rum |
| Hobbies | Gardening, coins, stamps | Ceramics, biking, drawing |
| Set B: Weak | | |
| Metals | Chrome, platinum, magnesium | Mercury, pewter, tungsten |
| Weapons | Hammer, fist, lance | Rock, arrow, dagger |

## Appendix C

### Categories From Experiment 3, With the 12 Exemplars From Each Category Divided Into Four Subsets (S1, S2, W1, and W2) and With the Categories Divided Into Practice Counterbalancing Sets A and B

| Category | S1 | S2 | W1 | W2 |
|---|---|---|---|---|
| Set A | | | | |
| Drinks | Vodka | Bourbon | Sake | Moonshine |
| | Rum | Ale | Tequila | Cognac |
| | Gin | Whiskey | Drambuie | Kahlua |
| Weapons | Sword | Bomb | Arrow | Nail |
| | Rifle | Pistol | Dagger | Foot |
| | Tank | Club | Hatchet | Lance |
| Fish | Catfish | Bluegill | Walleye | Yellowtail |
| | Trout | Flounder | Snapper | Muskie |
| | Herring | Guppy | Angler | Puffer |
| Fruits | Tomato | Orange | Fig | Coconut |
| | Strawberry | Lemon | Mango | Raisin |
| | Banana | Pineapple | Nectarine | Guava |
| Set B | | | | |
| Professions | Engineer | Nurse | Veterinarian | Critic |
| | Accountant | Plumber | Janitor | Investor |
| | Dentist | Farmer | Gardener | Soldier |
| Metals | Iron | Silver | Francium | Lithium |
| | Aluminum | Brass | Tungsten | Pewter |
| | Nickel | Gold | Chrome | Mercury |
| Trees | Birch | Elm | Mimosa | Palm |
| | Hickory | Spruce | Cedar | Willow |
| | Dogwood | Redwood | Juniper | Ash |
| Insects | Beetle | Fly | Locust | Tick |
| | Roach | Mosquito | Weevil | Cicada |
| | Hornet | Grasshopper | Aphid | Scorpion |

*Note.* Assignments of subsets to A1, A2, B1, and B2 are not shown. S = strong; W = weak.

Received March 12, 1992
Revision received August 25, 1993
Accepted October 12, 1993 ■