Journal of Experimental Psychology: General
1998, Vol. 127, No. 1, 55–68

0096-3445/98/$3.00

The Mismeasure of Memory: When Retrieval Fluency Is Misleading as a Metamnemonic Index

Aaron S. Benjamin and Robert A. Bjork
University of California, Los Angeles

Bennett L. Schwartz
Florida International University

The experiments address the degree to which retrieval fluency—the ease with which information is accessed from long-term memory—guides and occasionally misleads metamnemonic judgments. In each of 3 experiments, participants' predictions of their own future recall performance were examined under conditions in which probability or speed of retrieval at one time or on one task is known to be negatively related to retrieval probability on a later task. Participants' predictions reflected retrieval fluency on the initial task in each case, which led to striking mismatches between their predicted and actual performance on the later tasks. The results suggest that retrieval fluency is a potent but not necessarily reliable source of information for metacognitive judgments. Aspects of the results suggest that a basis on which better and poorer rememberers differ is the degree to which certain memory dynamics are understood, such as the fleeting nature of recency effects and the consequences of an initial retrieval. The results have pedagogical as well as theoretical implications, particularly with respect to the education of subjective assessments of ongoing learning.

There has been a surge of interest in *metamemory*—the study of what people know and understand about their own memory and memorial processes. From a theoretical standpoint, there has been a particular effort to explain why certain metamnemonic measures, such as the *feeling of knowing* (FOK; see Hart, 1965) or *judgments of learning* (JOL; see Arbuckle & Cuddy, 1969), are accurate or inaccurate under various conditions (e.g., Dunlosky & Nelson, 1992, 1994; Koriat, 1995; Reder & Ritter, 1992; Schwartz & Metcalfe, 1992). Theories have also been advanced to account for metamemory phenomena computationally (Metcalfe, 1993; Reder & Schunn, 1996) and at the process level (Koriat, 1993). From a practical standpoint, researchers have investigated the degree to which metamnemonic judgments play a role in determining study behavior (e.g., Nelson & Leonesio, 1988) and selecting efficient—or inefficient—training regimens (for a review, see Jacoby, Bjork, & Kelley, 1994).

One important characteristic of such theories is that metamnemonic judgments are assumed to be highly inferential in origin. This position stands in stark contrast to certain historical views of metamemory, which posited the existence of an internal monitor that surveyed memory contents in a relatively unbiased manner and formulated judgments about future retrievability based on the presence or absence of the to-be-retrieved item (Hart, 1967; Burke, MacKay, Worthley, & Wade, 1991). The nature of the inferential task facing the metacognizer is to evaluate the *objective* status or future retrievability of a particular memory given certain *subjective* cues as to its current status. Such a process must incorporate an understanding of which subjective cues are diagnostic and, furthermore, when that diagnosticity may be compromised.

Our analysis focuses on one particular subjective cue that is accessible to humans: the fluency or ease with which information comes to mind. We operationalize such *retrieval fluency* (RF) as the speed with which information is accessed from memory and reported, the probability that it is accessed and reported, or both. The goal of our research is to evaluate whether RF is used in the formation of metamnemonic judgments, and to assess whether humans know how to modulate such use in the face of predictive tasks on which such reliance is potentially misleading.

We do not wish to argue that the use of RF as a metamnemonic index is a poor heuristic. Quite to the contrary, it is very often true that ease or speed of retrieval now predicts fluent retrieval in the future. For example, Blake (1973) showed that both feeling-of-knowing judgments and eventual recall probability correlate with the amount of partial information accessed during a retrieval attempt (see also Koriat, 1993; Schacter & Worling, 1985). Also, judgments of learning appear to be differentially accurate in predicting future cued-recall performance depending on whether a diagnostic retrieval immediately prior to the judgment is possible (Dunlosky & Nelson, 1992). Dunlosky and Nelson (1994) showed that when a retrieval is made diagnostic by delaying it from the study episode, predictions of cued recall accurately reflect the efficacies of different encoding procedures, such as distributed versus

Aaron S. Benjamin and Robert A. Bjork, Department of Psychology, University of California, Los Angeles; Bennett L. Schwartz, Department of Psychology, Florida International University.

Many thanks to Stephanie Jacobs, Barbara Bui, and Janet Chung for their assistance with running participants and coding data. Thanks also to Tom Nelson, Asher Koriat, and Lynne Reder for their helpful suggestions.

Correspondence concerning this article should be addressed to Aaron S. Benjamin, Department of Psychology, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, California 90095-1563. Electronic mail may be sent via Internet to benjamin@psych.ucla.edu.

massed trials and interactive imagery versus rote rehearsal (see also Begg, Duft, Lalonde, Melnick, & Sanvito, 1989, Experiment 2) Predictions do not mirror such important encoding distinctions if they are made immediately after study, at which time an attempted retrieval is not diagnostic of future performance (e.g., Rabinowitz, Ackerman, Craik, & Hinchley, 1982).

The important issue is how sophisticated humans' conceptualizations of memory are: Can they evaluate when and to what degree RF is and is not diagnostic? The general strategy of our research was to pick experimental paradigms in which either speed or probability of retrieval is at odds with later probability of retrieval. If RF is an important basis for predictions, then we expect participants to use such information even when it misleads predictions to a disastrous degree. Additionally, such heavy reliance on RF as a metacognitive index implies a lack of appreciation for the multidimensionality of memory, a theme we have developed elsewhere (Benjamin & Bjork, 1996).

Other authors have investigated the accuracy with which participants predict their own recall performance, particularly with respect to the role such judgments play in the execution and selection of control processes (e.g., Groninger, 1979; Mazzoni & Cornoldi, 1993; Mazzoni, Cornoldi, & Marchitelli, 1990; Nelson, 1993). These authors have found predictions to be generally accurate. Judgments appear to correctly incorporate characteristics of word frequency and imagability (Groninger, 1979), for example, and use such information in the allocation of study time (Mazzoni et al., 1990; Nelson, 1993), even if that study time is not necessarily used productively (Nelson & Leonesio, 1988).

Dunlosky and Nelson (1992, 1994; see also Nelson & Dunlosky, 1992) have found predictions of cued-recall performance to be maximally accurate when the JOL is solicited with only the cue term present and at a substantial delay after learning. Although there is some controversy as to the nature of such an effect (Nelson & Dunlosky, 1992; cf. Spellman & Bjork, 1992), we emphasize here the attenuation of metamnemonic accuracy under nondiagnostic retrieval conditions.

Our position is that a greater understanding of the heuristics and indexes participants use in making metamnemonic predictions—accurate or inaccurate—sheds light on why judgments are sometimes accurate and sometimes not and thus better informs our theories as to how such inaccuracies can be rectified in a pedagogical or training milieu. A growing number of researchers are investigating not simply the information-processing aspects of metamemory, but also its relation to learning strategy differences in education (Maki & Berry, 1984; Owings, Petersen, Bransford, Morris, & Stein, 1980), training (Bjork, 1994; Jacoby et al., 1994), in children (for a review, see Siegler, Adolph, & Lemaire, 1996), and in different neuropsychological syndromes (for a review, see Shimamura, 1994). In the final portion of this article, we discuss individual differences in the degree to which participants are apt to be misled by retrieval fluency and potential ramifications of our findings for educational considerations.

Experiment 1

The purpose of Experiment 1 was to examine whether participants could distinguish between ease of access from *semantic memory* and ease of access from *episodic memory.* Semantic memory, as initially formulated by Tulving (1972), involves the storage of factual or conceptual information in an abstract (context-free) associative structure, whereas episodic memory involves the storage of personally experienced episodes from an autobiographical and context-dependent perspective.

These two memory classes may overlap in the information they contain, as in the following example. Some of us may know that the Toronto Blue Jays won the World Series in 1993. We may even know that the winning home run was hit by Joe Carter off of a bad pitch from Mitch Williams. Such information holds a place in our semantic memory. However, we may also remember exactly how and where we first heard (or saw) the event, how we felt, or whom we were with at the time. This information is episodic in nature.

Such a distinction implies that ease of access to information in one type of memory may not mean equally probable access to related information in the other type of memory. We may, for example, remember where we were when the Challenger space shuttle exploded but not the names of any of the astronauts on board. Or, we may remember that smoking causes lung cancer but not where we first learned that fact. For our purposes, we needed to find a task in which the ease of retrieval of nominally equivalent information was negatively correlated between episodic and semantic access.

An experimental procedure used by Gardiner, Craik, and Bleasdale (1973) is one such possible task. Gardiner et al. demonstrated that words retrieved with difficulty during a general knowledge question-answering task were later free recalled with a higher probability than words that had been retrieved more easily on the initial question-answering task. Such a relationship, although initially counterintuitive, is easily explained by the semantic–episodic distinction. When participants answer the question initially, they are being guided by the question on a search through semantic memory that is concluded by the arrival at an answer, right or wrong, or by a failure to arrive at any answer to the question. The longer they spend on such a search, the more salient or elaborated the entry they create in episodic memory for the event of having searched for that answer. The more elaborated the episodic trace is, the more easily it can be accessed on a later free-recall task, which is primarily an episodic memory task.

In the Gardiner et al. (1973) paradigm, those items that have more accessible semantic memory entries (i.e., that are known "better" and thus retrieved more quickly) may yield episodic traces that are less accessible. However, if participants assume that memory is homogeneous—that is, if their mental model of their own memory does not incorporate the semantic–episodic distinction—their predictions should reflect RF from semantic memory and thus run opposite to actual later recall performance (from episodic memory). That is, participants should predict more facile free recall for

items initially retrieved quickly and poorer recall for items initially retrieved with greater difficulty.

Method

The experiment follows the general procedure of Gardiner et al. (1973). After answering each of 20 general information questions, participants gave a prediction of free recall (PFR) for the answer they provided. Following a distractor task that took 10 min, participants were then asked to free recall the answers they gave during the first phase of the experiment.

Participants. Eighty undergraduates from the University of California, Los Angeles (UCLA), participated in the study and did so to partially fulfill course requirements.

Materials and apparatus. The 20 general knowledge questions used in the experiment were drawn from a compendium provided by Nelson and Narens (1980b). All participants answered the same 20 questions, which were selected to be of a moderate-to-easy level of difficulty. Presentation of the questions and the collection of response information was controlled by a personal computer. Two pages of drawings of individual states of the United States were used in a distractor task, and one blank page (unlined) was used for the final free recall.

Procedure and design. Participants were told that they would be asked 20 trivia questions and that the time it took to answer each question was of primary interest to the experimenters. Thus, it was explained, they were to press the *Enter* key as soon as they knew the answer to the presented question. No explicit instructions concerning guessing were given.

It was emphasized to participants that they should not press the Enter key before mentally generating an answer but that they should do so as soon as possible after coming up with an answer. Participants were told that the question would then disappear from the screen and that they were to enter their answer, a space, and then their prediction of free recall for that word. This prediction was defined as the subjective percentage probability that they would be able to produce that answer again in 20 min on a blank sheet of paper with no cues. It was explained that because their predictions were to represent probabilities, the values of those predictions should range from 0 (meaning no chance of later recall) to 100 (meaning certain later recall). They were told that no prediction was to be made if they had not come up with an answer. Participants were further instructed that the nature of the final recall task was such that the questions would *not* be re-presented and were shown a blank sheet of paper to emphasize the conditions under which that task would be performed.

Participants then initiated the experiment by pressing the Enter key. The 20 questions were presented one at a time in one of two random orders. After completing this task, they were given the two pages of line drawings of the outlines of individual states of the United States and told that the second part of the experiment involved examining the relationship between general trivia knowledge (as assessed in the previous part of the experiment) and geographical knowledge. They were instructed to label each state with its appropriate name and to emphasize accuracy over speed. This activity was stopped after 10 min, which, together with the initial question-answering period, totalled approximately the 20-min delay indicated to the participants during the instructions.

Finally, participants were provided with the blank sheet of paper and asked to recall all of the answers they had generated in the first part of the experiment. They were instructed to persevere for the full 10 min allotted, despite the fact that they may feel as though they had recalled as many as possible earlier in that period.

Before the participants were released, they were debriefed on the nature of the task and the hypotheses and were given credit for the completion of the experiment.

Scoring. For each of the 20 items, three dependent measures were recorded: (a) initial response latency, (b) predicted free recall, and (c) whether that word was free recalled on the final test. Because the general design of the later analyses was quasi-experimental in nature, there were no independent variables.

One participant was excluded for failing to provide responses to eight of the questions. For the remaining 79 participants, the goal was to examine the probability of final recall and PFR as a function of initial response latency. Simply examining rates of final free recall as a function of initial response times (or intervals thereof), however, introduces a selection bias across participants: The longer response times disproportionately weigh participants with long mean response latencies, and the shorter response times disproportionately weigh participants with short mean response latencies. To circumvent this problem, each participant's individual data set was split into four quartiles based on their own median response latency. For each quartile, for each participant, average rates of final recall and average PFRs were calculated. All participants, therefore, contribute equally to each response-time quartile, which was then averaged across subjects. For a given participant, only those items for which an answer was provided were included in the analysis. Incorrect and correct responses were analyzed separately but are combined in the following analyses because there were no substantive differences between the two.

Our analytic strategy differs somewhat from the norm in metamemory research. Nelson (1984) has argued convincingly that the gamma statistic (*G;* Goodman & Kruskal, 1954) is the most appropriate measure of association for data from the feeling-of-knowing paradigm, and its use has been extended to other domains of metamemory as well. The nature of the question that we are addressing motivates the alternative analysis that we have suggested. We digress here briefly to discuss differences in the goals of our research from the goals of research for which the gamma statistic is the preferred measure.

The gamma coefficient is the most satisfactory of a family of measures of association in the treatment of ordinal data, such as FOK ratings. Whereas FOKs themselves are often made on an interval scale, Nelson (1984) has emphasized the lack of an empirical foundation for the treatment of FOK ratings as interval-level data. Such confusion is avoided in paradigms such as the one advocated by Nelson and Narens (1980a) in which FOK data are solicited in a manner that emphasizes their ordinal nature.

As a measure of association, the gamma coefficient provides an index of the degree of variability present in one variable that can be explained by variability in the other. Such a statistic is important in explaining the degree to which predictions derive from some other measured source or the degree to which performance reflects predictions. Because our primary focus is not the relative role that retrieval fluency plays in the generation of metamemory judgments, but rather whether judgments vary reliably with differences in retrieval fluency, the analyses presented in the body of this article take a similar form to those presented in Experiment 1. Because correlational measures such as *G* are, however, the most appropriate indices of the degree of the relationships we discuss, those measures are presented in Table 1.

Because a great many factors influence any metamnemonic judgment—including a wide range of word-related, participant-related, and word-by-participant-related factors that are not germane to the issues motivating Experiments 1 and 2—the values of *G* in this research are in fact quite low. However, as can be seen in the data, the influence of RF on participants' predictions is quite clear in the quartile analyses we consider most appropriate.

Table 1.
Mean Gamma Correlations and Standard Errors of the Mean (SEM) From Experiments 1, 2A, and 2B

Experiment/Correlation	*G*	*SEM*
Experiment 1		
Response time–recall probability	.133	.036
Response time–prediction	−.348	.035
Prediction–recall probability	−.035	.056
Experiments 2A and 2B		
Output position–recall probability	.143	.047
Output position–prediction	−.146	.078
Prediction–recall probability	−.177	.064

Note. *G* = Gamma; Goodman–Kruskal's index of relationship.

Results

The results presented next are reliable at the $p < .05$ level unless otherwise noted. For each participant, there were two data points for each of four response-latency quartiles—one point for the mean rate of final recall for items in that quartile and the other corresponding to the average PFR provided for items in that quartile. Average response latencies for the four quartiles were 2,795 ms, 3,918 ms, 5,452 ms, and 9,975 ms, respectively. The top panel of Figure 1 shows recall rate as a function of response time quartile. The increasing rates of final recall with response-latency quartile are consistent with the findings of Gardiner et al. (1973). This relationship was reliable by a Friedman test, $\chi^2(3, N = 79) = 40.356$. A Wilcoxon signed-ranks test indicated reliable pairwise differences between all groups except Quartiles 2 and 3.

The bottom panel of Figure 1 shows mean PFR ratings by response latency quartile. Consistent with our predictions, ratings decrease with increasing response latency quartile, Friedman test, $\chi^2(3, N = 79) = 77.787$. All pairwise comparisons were reliable by a Wilcoxon test except between Quartiles 2 and 3.

Discussion

Clearly, predictions of free recall reflect how quickly the relevant information was accessed during a semantic memory search. It would thus appear that participants fail to appreciate the mnemonic nature of the to-be-predicted free-recall task, and they erroneously rely on RF on a wholly different task to predict their own future performance. More generally, it appears as though participants use a predictive strategy that entails the conceptualization of memory as unidimensional in nature and thus tacitly endorse the notion that memory is homogeneous. The failure of participants to accurately predict their own future recall performance is particularly interesting in that their mistaken beliefs concerning the homogeneity of memory lead them to rely on the correct cues for prediction but to use them incorrectly.

The quasi-experimental design of Experiment 1 leaves open the possibility that predictions of recall are not supported by retrieval fluency itself but rather reflect some other variable, such as how well an item is known. This variable then could serve both to foster fluent retrieval and elevated predictions of future recallability. However, this conclusion appears unlikely in light of several related findings. First, Lee, Narens, and Nelson (1993) found that subthreshold target priming increased judgments of learning while leaving eventual recall unaffected, thus supporting the notion that metamnemonic judgments reflect more than covariation with those factors supporting eventual recall (see also Mazzoni & Nelson, 1995). In particular, judgments seem to incorporate the kinds of fleeting elevation of retrieval fluency afforded by simple perceptual priming.

More fundamentally, if predictions reflected some subjective index of knowingness that served to enhance both predictions and retrieval fluency, then that factor would most likely also serve to enhance eventual recall. More well-known material is more readily retrieved than less well-known material (see, e.g., Freedman & Loftus, 1971; Loftus & Freedman, 1972). If such underlying dynamics were at work, the negative association between predicted and actual levels of recall would not obtain.

Experiment 1 demonstrates that participants lack the

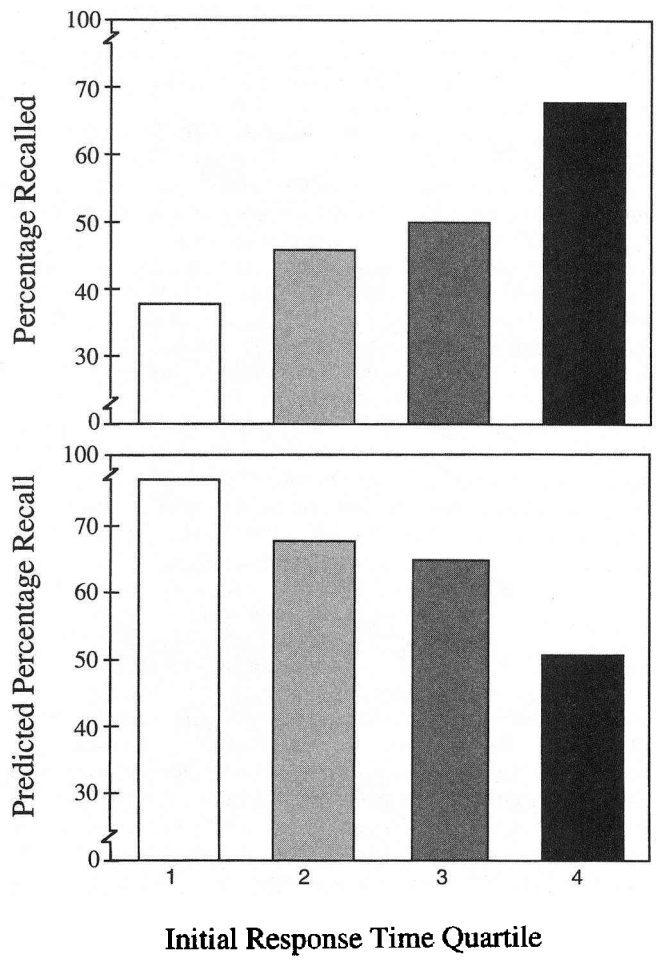

Figure 1. Mean final free recall (top panel) and prediction of free recall (bottom panel) as a function of response time quartile (Experiment 1).

ability to predict performance on an episodic memory task from performance on a semantic memory task. We have argued that this inability arises because of an lack of appreciation for the multidimensionality of memory. There are also experimental paradigms in which episodic access at one time does not predict episodic access at a later time. The goal of Experiments 2 and 3 is to examine whether participants understand or fail to understand the conditions under which fluency of initial access to episodic memories does and does not predict fluent access to those memories at a later time.

Experiments 2A and 2B

As in the previous experiment, Experiments 2A and 2B involved prediction making at the time of an initial retrieval. However, unlike the previous experiment, the to-be-predicted task and the task during prediction are both episodic and highly similar. Thus these experiments address metacognitive appreciation of the effects of time and intervening events on a single task, rather than dissociations between types of recall tasks.

Like the previous task, this experiment uses a paradigm in which RF during an initial task has been shown not to predict retrieval probability on a later task. In these experiments, participants learn a list of unrelated words that they are to immediately free recall. Upon doing so, they also make predictions, for each word that they recall, of their likelihood of re-recalling that word later. Thus, they are predicting performance on a later free-recall task while engaging in free recall during the predictive task. However, the nature of free-recall performance differs dramatically between an immediate and a delayed free-recall test. These differences arise from two major factors.

First, the recency portion of a participant's recall protocol—a portion that, on an immediate test, shows substantially higher recall than the rest of the list (e.g., Murdock, 1962)—is depressed to the recall level of items from the middle of the list when tested at a delay (see, e.g., Postman & Phillips, 1965). This *recency-to-primacy shift,* as it has been termed, has been attributed to the fleeting nature of short-term memory storage, which provides for a short-term dump of recency items on an immediate but not on a delayed test. Moreover, when there is an additional end-of-experiment recall test, there is sometimes an additional *negative recency effect* (Craik, 1970), whereby recency items are recalled with even lower probability than items from the middle of the list.

This additional variance arises from the differential effects of retrieval during the initial test on recency and nonrecency items. Bjork (1975) demonstrated that the later recall probability of nonrecency items is enhanced by immediate recall to a much greater degree than that of recency items. Such a conceptualization is consistent with the idea introduced earlier that labored retrieval is a more effective learning event than relatively passive retrieval. In this case, recency items that are dumped quickly and easily but actually only learned to a weak degree are accorded little benefit of their initial retrieval.

It is crucial in the development of metacognitive skill that the effects of retrieval practice are recognized and understood. Retrieval practice has been shown to have powerful enhancing effects on future retrieval (see, e.g., Landauer & Bjork, 1978), and we have emphasized here the role that the fluency of a particular retrieval episode may have in promoting future retrieval: namely, that more involved, difficult retrievals are more effective learning events than are very fluent retrievals. A goal of Experiments 2A and 2B is to address the degree to which participants appreciate such a contingency.

Reliance on RF, however, once again misleads the predictive process. Those items that are output first during free recall—and are thus of high retrieval fluency—suffer doubly on a later test. First, they accrue little benefit as a result of their initial retrieval owing to their quick and easy access. Second, the items recalled first tend to be those from the end of the list, ones that are still in short-term memory and readily accessible, but items that may never have been processed to a level that would support delayed recall. For these reasons, items output late in the retrieval process tend to be re-recalled with a higher rate than items output earlier in that process. Predictions that are reliant on RF, however, would predict the opposite of this pattern. Items output late in retrieval are, by definition, of lower RF than items output earlier in retrieval and would accordingly be rated lower on future recallability. Experiments 2A and 2B address this hypothesis.

An additional goal of Experiment 2A was to examine whether the act of prediction making affects the recall on either the initial or final test. Thus, in Experiment 2A, half of the immediate recall tests were not accompanied by predictions. In Experiment 2B, participants made PFRs for all items recalled on all lists.

Method

Participants. In Experiment 2A and 2B the participants were 24 and 79 undergraduates at UCLA, respectively, with 13 women and 11 men in Experiment 2A and 42 women and 37 men in Experiment 2B. They participated in partial fulfillment of course requirements.

Materials and apparatus. Six lists of 13 unrelated words were created and put onto slides. These 78 words were randomly assigned to the six lists and randomly assigned to a position within that list. Two different orders of slides were created. They were shown with a slide projector in a dimly lit room, and participants were provided with a response booklet in which to record the words that they could recall and their predictions. This booklet contained eight pages. The first page was used for a practice trial and had 13 lines, one for each potentially recalled word. Next to each line was a scale from 0 to 100, marked in increments of 10. Participants were asked to indicate their PFR by circling the appropriate percentage on this scale. In Experiment 2B, the following six pages—to be used for the test lists—were exactly the same as the practice page. In Experiment 2A, only three of the six were the same; the remaining pages had no prediction scales. The final page of the booklet was completely blank and was used for final recall of all words from all six lists.

Procedure and design. Participants were run in groups ranging from 3 to 8 in size. They were seated around a large table, and all

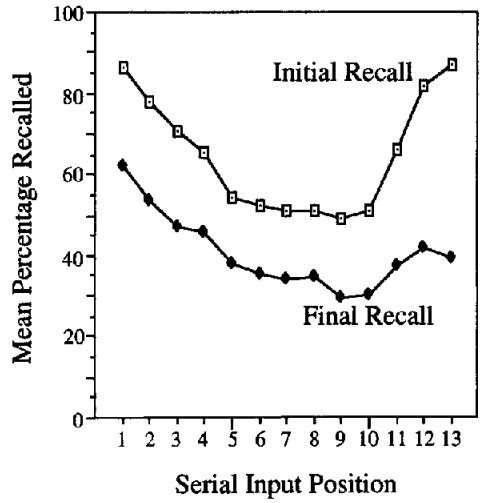

Figure 2. Mean immediate and final (delayed) recall as a function of an item's serial input position (Experiments 2A and 2B).

participants faced the front of the room and the screen onto which the to-be-remembered words were projected one at a time. After being instructed as to the nature of recall and the predictive task, participants viewed a practice list of 13 words. They immediately attempted to recall this list and made predictions. The purpose of this practice list was to familiarize the participants with the prediction making and to be sure that participants understood that a prediction was to be made after each word, rather than after all words in a given list had been recalled. Participants were told that they would be viewing six lists, each with 13 words.

During the experimental phase, participants viewed each of 13 words in six lists at a rate of 2 s per word. After an entire list, a blank slide and a "Recall" instruction from the experimenter cued the immediate recall phase. Participants were given 150 s to recall and make their predictions. After the final (sixth) list, they were told to turn the page and recall as many words from all of the previously viewed lists as possible. They were given 10 min for this final recall episode. After completion of the experiment, all participants were debriefed and given appropriate credit for their participation.

Results

An analysis of the results of Experiment 2A revealed that the act of prediction did not affect overall levels of recall, either during initial or final test. More importantly, this variable (predicting or not predicting) did not interact with time of test. Given that outcome, the following analyses are based on the pooled data sample from both Experiments 2A and 2B.

In Figure 2 are shown the effects of serial input position on immediate and delayed recall. The classic serial position curve is evident in the data from the immediate test, in which enhanced recall of both primacy and recency items is clear. Only primacy items enjoy a marked delayed recall advantage over the remainder of the items. In both cases, larger primacy effects are evident than were shown by Craik (1970). Such additional enhancement in the recall of primacy items may be due to the fact that, in the current experiment but not in Craik's study, participants were aware of the impending end-of-experiment recall task. Such awareness may also explain the absence of a negative recency effect in our data.

The presence of these qualitative relationships is borne out by an analysis in which we sum recall levels of primacy items (Items 1–4), middle items (Items 5–9), and recency items (Items 10–13). On the test of immediate recall, a main effect of input group (primacy, middle, or recency) was found, $F(2, 204) = 55.341$. Scheffé post hoc analysis revealed differences between primacy ($M = .77$) and middle ($M = .56$) groups, $F(1, 204) = 52.98$, as well as between middle and recency ($M = .70$) groups, $F(1, 204) = 21.265$. A main effect of input group was again found on the test of final recall, $F(2, 204) = 29.385$, but for these data, the differences were between primacy ($M = .53$) and middle ($M = .40$) groups—again, by a Scheffé test, $F(1, 204) = 19.400$—as well as primacy and recency ($M = .37$) groups, $F(1, 204) = 25.064$. These findings support the assumption described earlier that those factors supporting the recall of recency items during immediate testing fail to contribute to the same degree when testing is at a delay.

In Figure 3, final recall conditionalized upon immediate recall of an item is shown as a function of serial input position. Consistent with the notion that very easy or fluent retrieval enhances future recall of that item to a lesser degree than does a more labored retrieval, analysis reveals that recency items—the items that are typically output early and with ease—demonstrate reliably lower rates of final recall given initial recall than do primacy or middle items, $M = .54$, as compared to rates of .68 and .69 for primacy and middle, respectively, Omnibus $F(2, 204) = 19.948$; Scheffé

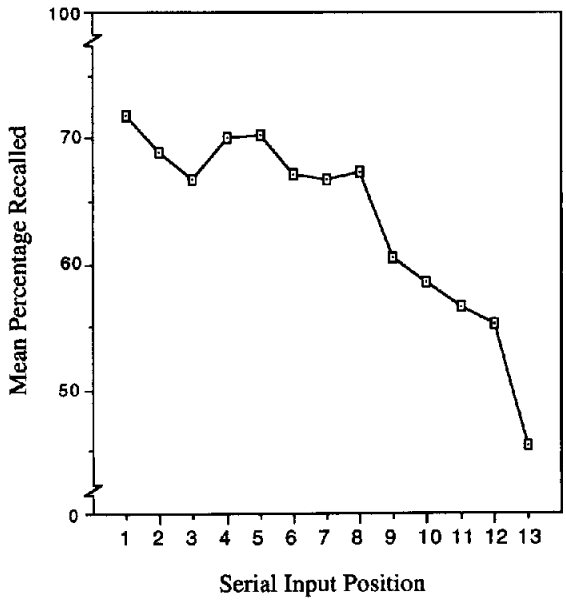

Figure 3. Mean final (delayed) recall conditionalized on immediate recall as a function of an item's serial input position (Experiments 2A and 2B).

$F_{[primacy,\ recency]}(1, 204) = 12.379$, $F_{[middle,\ recency]}(1, 204) = 17.588$.

Our data also support the premise that recency items tend to be output earlier in the retrieval process than other items. In Figure 4 we see average output position during recall as a function of original input position during study. The pattern in these data suggests earliest output for items that appear latest on the study list, next earliest output for primacy items, and latest output for items from the middle of the list. In fact, such a relationship is borne out by analysis: Recency items (items from Positions 10–13) have an average output position of 3.64, primacy items (Positions 1–4) have an average output position of 5.65, and items from the middle of the list (Positions 5–9) have an average output position of 6.49. There is a main effect of item type, $F(2, 204) = 74.795$, and Scheffé post hoc analysis reveals that all of these values are significantly different from one another.

For the purposes of normalizing across participants with respect to mean rates of recall, a quartile split analogous to the one described in Experiment 1 was performed. For each list immediately recalled by each participant, the output was split into four approximately equal groups, and mean rates of final recall given immediate recall and PFRs for items in that quartile were calculated. These rates were then averaged across all six study lists, yielding a total of eight scores per participant: four rates of final recall corresponding to the quartile split and four mean prediction values corresponding to those quartiles. Figure 5 represents these data in similar form to those presented from Experiment 1.

Friedman tests indicated main effects of quartile on both recall rates, $\chi^2(3, N = 103) = 30.081$, and predictions, $\chi^2(3, N = 103) = 46.949$. A Wilcoxon signed-ranks analysis revealed what is immediately evident in Figure 5: Recall rates are significantly lower in Quartile 1 ($M = .49$) than in any of the other quartiles (.62, .62, and .60, respectively), which are not different from one another. Also, PFRs are lower in Quartile 4 ($M = 42$) than any other quartile (52, 53, and 51, respectively), which do not significantly differ from one another.

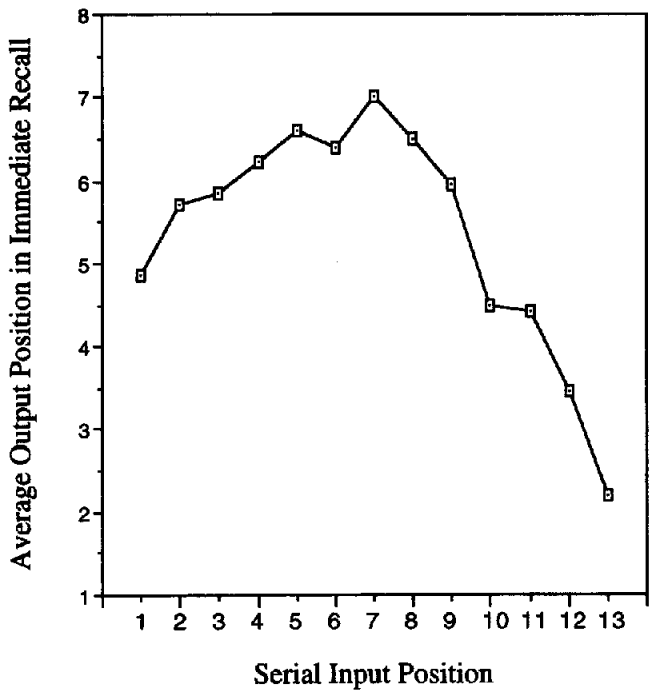

Figure 4. Mean output position during immediate test as a function of that item's serial input position (Experiments 2A and 2B).

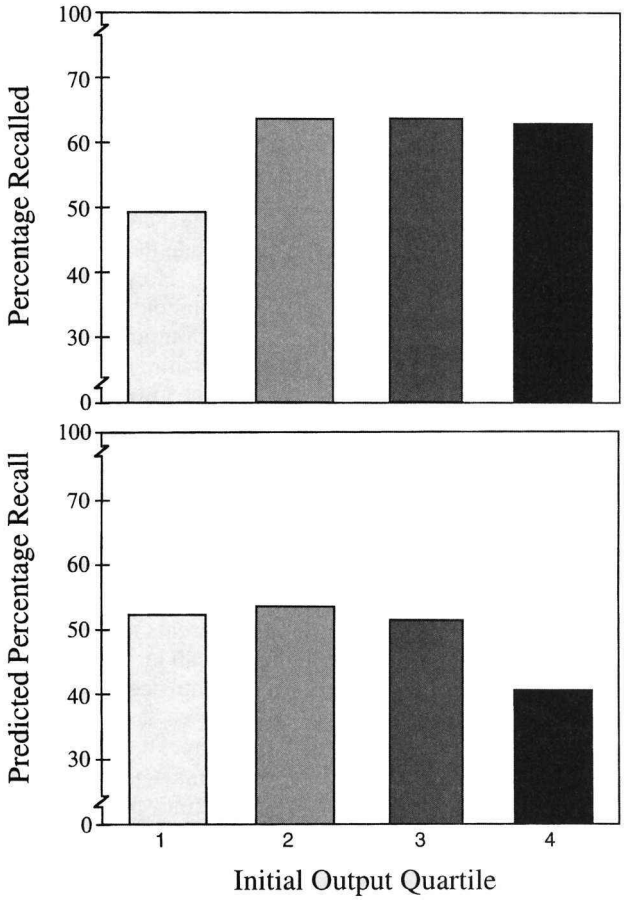

Figure 5. Mean final recall (top panel) and prediction of free recall (bottom panel) as a function of initial output quartile (Experiments 2A and 2B).

Discussion

In Experiments 2A and 2B, we replicated the effect that recency items suffer disproportionally relative to other items in their retrievability with the effects of time (e.g., Bjork, 1975; Craik, 1970). We have provided supporting evidence that this recency-to-primacy shift in our experiment derives from two sources: the fleeting nature of the representation supporting recall of recency items during immediate testing and the failure of such fluently retrieved items to benefit adequately from retrieval practice. However, our real concern was whether participants' predictions demonstrate sensitivity to these effects.

Quite clearly, they do not. It is evident in our data that items output with high RF—in other words, those items output in the initial portion of a participant's recall protocol, or Quartile 1—gain less in terms of future retrieval enhancement than do other items. This relationship can be seen by comparing the rates of final recall given initial recall as a function of initial recall output quartile. Items output in the first quartile at the time of the immediate test are re-recalled at the final test with a lower probability than items from the remainder of the initial output protocol. However, PFRs from this quartile do not reflect such a relationship: Participants predict equivalent recall for words output in the first quartile as words output in the second and third quartiles.

More strikingly, participants do predict lower recall for items output in the fourth quartile: those items output with the most difficulty, or least RF. However, such items do not suffer by virtue of their late recall. In this experiment, those items are recalled at an equivalent level as those in Quartiles 2–4; however, others have shown that such late-output items may indeed be recalled best at a final test (Bjork, 1970; Craik, 1970). In either case, it is clear that the assignment of low predictions to late-output items represents misled metacognition. In predicting from one free recall task to another, it is evident that participants fail to take into account factors that effectuate differences between the two tasks: namely, the effects of retrieval during the initial recall task and the disproportionate forgetting of recency items with a delay.

However, the interpretation of the results from Experiments 2A and 2B is limited in light of the lack of control over participants' output in each of the immediate recall attempts. Because different participants recalled different amounts, differences between groups were defined in terms of the quartile analysis discussed earlier. Although such analysis does circumvent certain problems with participant selection, it leaves open the possibility of a participant-by-item bias. That is, whereas an equal amount of each participant's data was distributed to each of the four quartiles (by virtue of the analysis discussed earlier), participants contributed different total amounts (across all four quartiles) on the basis of their overall level of performance in the initial phase of each task. It is unclear how aspects of the obtained results could be attributed to that difference, but Experiments 2A and 2B nonetheless have the property that the independent variable is defined by performance on the initial task, not by the experimenter. In Experiment 3, we take advantage of a task in which the predictor variable is manipulated by the experimenter.

Experiment 3

Experiment 3 also addressed the extent to which participants appreciate the transient nature of the effects of recency on retrieval fluency and the impact of the difficulty of initial recall on later recall performance. In this experiment, however, predictions are made for performance on a test of cued recall. This difference adds not only a degree of cross-task validity; it also makes our results more comparable to those reported in the JOL literature (e.g., Dunlosky & Nelson, 1992, 1994), in which cued-recall testing is the norm. The failure of participants to accurately predict their own performance on this task is especially striking when compared to the results of others (e.g., Groninger, 1979) who have demonstrated that JOLs are of considerable accuracy.

The general procedure for Experiment 3 is taken from Madigan and McCabe (1971). In their experiment, participants cycled through 50 short lists, each of which consisted of five paired-associate terms. After each list, they were tested on one of the preceding five pairs with a cued-recall test in which the left-hand member of the pair was presented and participants were to recall the appropriate right-hand term. After all 50 lists, they were tested on the same 50 pairs with which they had been tested during the previous phase of the experiment and in the same manner. Their results showed—consistent with the effects we have described here—that performance dropped dramatically (from approximately 100% to 0%) from first to second test for those items that had been studied fifth in the list of five. Items at the beginning of the list (Item 1) showed a far smaller drop in performance between the two tests (from about 40% to about 28%). Such results are consistent with the two factors we have emphasized here: Recency items are not privy to the same long-term retention as other items and their easy retrieval (high RF) during the initial recall phase does not support later recall on the final test.

In this experiment, participants made predictions as to the later retrievability of a test item immediately after being tested on that item. In that sense, the task was highly similar to the tasks used in Experiments 1, 2A, and 2B. Note, however, that the task in Experiment 3 differs from those employed in Experiments 1, 2A and 2B in that recall—on both the immediate and delayed test—was cued.

Another important difference is that the analysis of both cued-recall performance and predicted final cued-recall performance can be examined as a function of the initial probe position within the list. We do not have to define post hoc categories for each participant's data; they fall into one of five categories, corresponding to the position in the list of the item that was tested. Unlike the previous sets of results, for which nonparametric analysis was necessary, the linear model analysis of variance (ANOVA) is an appropriate analytic tool for such data.

Method

Participants. Fifty-four undergraduates from the University of California, Los Angeles, participated in partial fulfillment of course credit. There were 22 men and 32 women in the group.

Materials and apparatus. We used 500 words from the Kucera and Francis (1967) word compendium. All words were between three and seven letters and were of medium to high frequency. For each participant, a unique study metalist was generated by randomly pairing words with one another and then randomly assigning those pairs to positions within the 250-pair list. The presentation of words, the recall testing, and the predictive task were all executed with a personal computer.

Procedure and design. Participants cycled through their uniquely created list and were tested, as in Madigan and McCabe (1971), after every five words (constituting one list). The probe

position of the item on which they were tested was derived from a preset sequence, also uniquely generated for each participant, but constrained such that each set of 10 lists (of which there were five) contained two probe tests from each of the five positions. Furthermore, no 2 immediately successive lists were tested on the same probe position. Other than those two constraints, the pair tested from a given list was random.

After each immediate test, participants were required to make a prediction of their own future cued recall of that same pair on an end-of-experiment test. Participants were informed that the cues at the later test—namely, the initial word in the paired associate—would be the same as during the current test. These predictions, as before, ranged from 0 to 100. They were also told that they would be studying 50 short lists of five pairs each and that the final test would be administered immediately thereafter.

After all 50 lists, participants were retested (after about a 1 min delay) on the same pairs on which they had been tested during the course of the initial phase of the experiment. The test order of these 50 pairs was random subject to the following constraints: (a) Each group of 10 tests included 2 pairs from each of the five probe positions, (b) the two probes from a given position were never drawn from the same one fifth of the presentation sequence, and (c) the items corresponding to the five probe positions were all drawn from separate fifths of the presentation sequence. After the final test, participants were debriefed, given credit for their participation, and released.

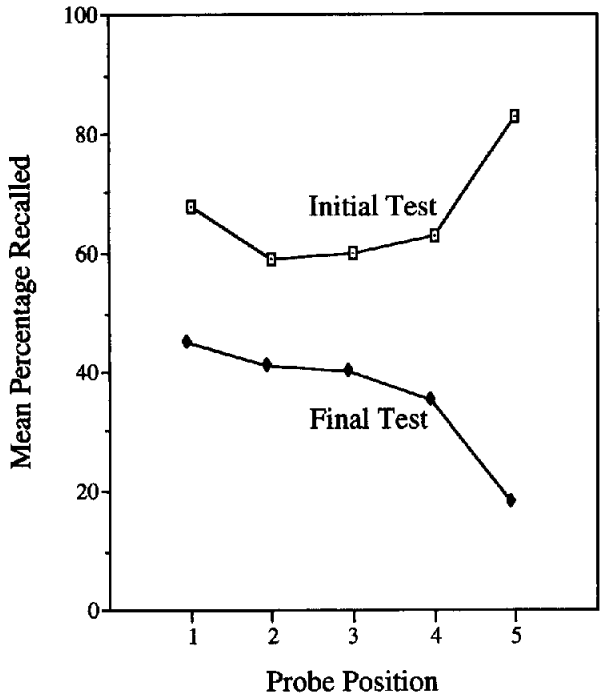

Figure 6. Mean immediate and delayed cued recall as a function of probe position (Experiment 3).

Results

Four participants were excluded from the analysis because they failed to contribute data to one of the five probe positions. One additional participant was excluded because she or he had a final recall score of zero.

In Figure 6, immediate cued-recall performance for the remaining 50 participants is shown as a function of the probe position of the test item from the just-studied list. The curve shows a pronounced recency effect. Madigan and McCabe (1971) used auditory presentation of items and consequently elicited a somewhat larger recency effect. (Such an interaction between serial position and input modality on rates of recall has been noted elsewhere; see, e.g., Murdock & Walker, 1969.)

There is a main effect of probe position on immediate cued recall, $F(4, 192) = 9.925$. The difference is limited to Probe Position 5 (the recency item), the value of which is reliably different from all other probe positions (by a Scheffé test). Rates of immediate recall for the other probe positions do not differ reliably from one another.

Figure 6 also shows cued-recall performance at the final test as a function of initial probe position. Here we see a pattern very different from the one evident in the immediate recall data: Recency items demonstrate markedly lower recall than items from other positions. The effect of probe position on recall is significant, $F(4, 192) = 10.6$, and again the difference is limited to the pairwise comparisons between Probe Position 5 and all other positions. However, unlike at the initial test, the difference is such that recall for Position 5 items is lower than recall for all other items.

For the purposes of examining the predictions, each participant's average prediction scores corresponding to the five probe positions were converted to rank scores. This transformation was made in order to parcel out the variance caused by individual differences in using the prediction scale. There were substantial differences between both the means and variances of participants' predictions; converting these scores to ranks helps to remove such individual differences. Thus, for each participant, the average predictions corresponding to the five probe positions were converted to the values 1–5, corresponding to the lowest and highest scores, respectively.

Gamma correlations are not presented in Table 1 for this experiment because the data describe a nonmonotonic function. The bottom panel of Figure 7 shows participants' predictions, by rank, as a function of the serial position of the pair being tested. Because these predictions are conditional upon initial recall, the top half of Figure 7 shows levels of final recall also conditionalized upon initial recall: the performance that the predictions are intended to describe. That is, the curves in the top and bottom graphs represent actual final cued recall and predictions of final cued recall for the exact same set of items. It is evident that both primacy (Serial Position 1) and recency (Serial Position 5) items were predicted by the participants to be more recallable on the final test than the middle pairs of a given list (Positions 2, 3, and 4). It is also clear, both from the top panel of Figure 7 and from the absolute levels of final cued recall seen in Figure 6, that recency items suffer disproportionately relative to other items on the delayed test.

A main effect of probe position on prediction values, $F(4, 192) = 17.444$, supports the claim that participants do incorporate information into their ratings that is related to serial position. Predictions for items in Serial Position 1 are

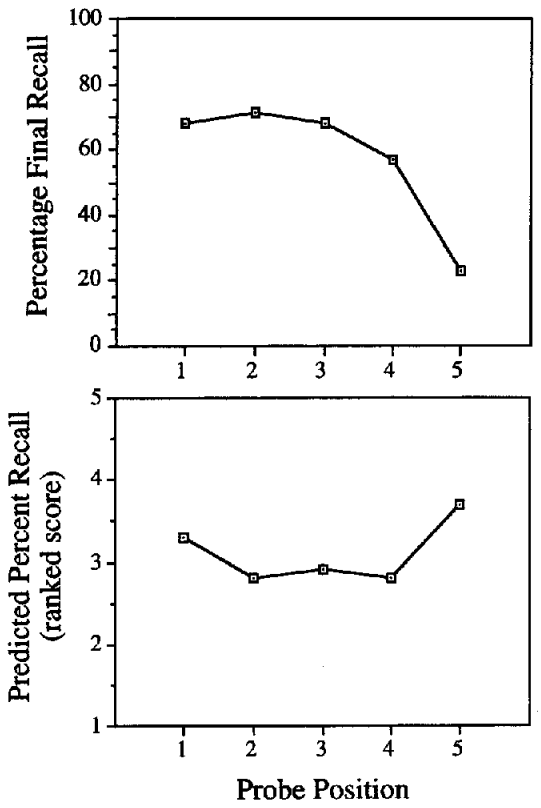

Figure 7. Mean rate of final cued recall conditionalized upon immediate recall (top panel) and mean predictions of future cued-recall performance (bottom panel) as a function of probe position (Experiment 3).

higher than predictions for items in Positions 2, 3, and 4 (by a Scheffé test). More important for our purposes, the predictions for items in Position 5 are also higher than those for items in Positions 2, 3, and 4. Predictions for items from Positions 2, 3, and 4 are not significantly different from one another.

Discussion

The results of Experiment 3 provide further support for the notion that RF at time of metamnemonic prediction is a potent source of information for judgments. Clearly, predictions of delayed recall that follow an immediate recall attempt reflect both long-lasting effects, such as stimulus–response associative strength, and more spurious, short-term effects. Differential predictions are not made for primacy and recency items, despite the fact that those two groups of items follow quite different courses of retention and consequently exhibit different rates of retrieval at a delay. These results, in conjunction with Experiments 2A and 2B, indicate that participants fail to appreciate the nature of serial position effects and further fail to understand the consequences of initial-retrieval difficulty in facilitating later retrieval.

General Discussion

The goal of our experiments was twofold: (a) to assess the degree to which fluency of initial retrieval is an index humans use in predicting their own future performance and (b) to examine whether humans' mental models of the functioning of their own memory would incorporate an understanding of the conditions under which initial fluent retrieval is and is not heuristically valid. Our general method for achieving that goal was to test whether predictions vary with RF at time of prediction, even when RF at that time does not predict future recall. Thus, in the three experimental scenarios that we have outlined, retrieval fluency is unconfounded from all factors at time of prediction that actually have predictive accuracy. To the degree that patterns of prediction vary with RF, the ascription of other causal agents to those patterns is limited to those that are entirely nondiagnostic and that correlate with current retrieval fluency. Such a constraint poses difficulties for theories of metacognition that posit trace access (Schwartz, 1994).

Our results are consistent with a view of metacognition that assumes that metacognitive processes are fundamentally inferential. In that sense, metacognition is simply a special case of cognition in general. We use the same skills of inference that we might use in evaluating when our children are likely to come home or when our car sounds like it needs a new muffler. When such skills are applied to the evaluation of our own performance, as opposed to our children's or our car's, we are engaging in metacognition. The notion that metacognition has such an inferential basis is strongly supported by current research (e.g., Koriat, 1993, 1995; Metcalfe, Schwartz, & Joaquim, 1993; Reder & Ritter, 1992; Schwartz, 1994; Schwartz & Metcalfe, 1992).

In each of the three experiments described here, we found a negative relationship between predictions of recall performance and actual recall performance. We have argued that such a relationship obtains because participants use current RF as a basis for such predictions, but our tasks were selected on the basis of RF being counterdiagnostic of later performance. Such negative relations between prediction and performance are rare but not unheard of in metacognitive research. Begg et al. (1989) found that participants gave higher predictions of later recognizability to high-frequency than low-frequency words, despite the superiority of low-frequency word recognition. Similarly, Koriat (1995) found that for those subsets of general information questions in which accessibility to an answer was high, but that the fluently-accessed answer was typically incorrect, participants often greatly overpredicted the likelihood of their being able to recognize the correct answer. Like our study, each of these failures to accurately predict performance can be viewed as a failure to apply the correct heuristic (or inference) given the task with which the participants were presented.

Good Versus Poor Learners

One somewhat neglected aspect of metacognitive skill is the degree to which individuals differ in their capacity for

accurate prediction. Such consideration is crucial if the results of metacognitive research are to be used in pedagogical applications. Poor self-monitoring capacity necessarily entails poor selection and execution of relevant control processes: If you do not know what you do not know, you cannot rectify your ignorance. Recognizing differences in the strategies that may be employed to the same metacognitive end allows evaluation of those strategies that are maximally appropriate for the task at hand. Concomitantly, such an understanding allows for evaluation and correction of poor strategy use by some students.

Maki and Berry (1984) provided evidence that metacognitive skill is to some degree correlated with cognitive skill in general. In their work, it was found that students who performed better on an exam in terms of performance were also more accurate in their predictions of what they did and did not know concerning the exam material. Data reported by Owings et al. (1980) support the notion that deficits in metamnemonic control process implementation underlie some apparent failures of memory. In their work, children who performed well on a test of memory allocated significantly more study time to those materials that were objectively difficult than to those that were easy; however, children that performed more poorly on the memory test did not differentially allocate their study time between easy and hard materials (for a review, see Nelson, 1996).

The results from the current experiments may provide an experimental analogue for an oft-seen real world phenomenon related to such test performance. Students often vehemently proclaim their perceived preparedness for an exam on which their performance is revealed to be quite inadequate. It may well be that they are indeed learning the material well by some criterion but that their assessments of learning are not suited to the task at hand. Imagine a student preparing for an upcoming psychology exam by preparing flash cards that have an experiment name, authors, and year on one side and the details of the appropriate research on the other. Further suppose that the student's gauge of learning is how quickly and easily she responds to the referential information of authors, and so forth, in recalling the details of the experiment. Such measures of speed, accuracy, or both do indicate something about how well the material has been learned.

However, there are two ways in which such study would not prepare our student well for the upcoming exam. First, the rate of response in the cuing procedure with which the student was preparing herself reflects little concerning the degree to which that information has been integrated, which, in turn, governs how fluidly that information can be assembled in the generation of a coherent essay. On a multiple-choice exam, however, such initial measures may indeed support accurate metacognitive prediction of future performance.

The second way in which such a study regimen can lead to faulty prediction reflects aspects of the predictive task itself, as in our experiment. Very simply, high RF during the flash-card studying episode does little to augment the future retrieval of information from that event. Low RF leaves a more lasting episodic impression that can foster performance on a future exam independent of the exam format. There are thus two factors at odds with one another in a predictive task such as the one described here: Whereas high RF does indicate better mastery of the to-be-retained information than low RF, it enhances future retrieval of that information to a lesser degree. Predictive tasks for which such indexes come into conflict thus place a burden on the metacognizer to assess aspects of the to-be-predicted task and to weigh those conflicting factors appropriately.

Maki and Berry's (1984) finding motivated a somewhat closer look at the results of Experiment 3, the design of which lended itself most aptly to such analysis. That is, in order to explore most appropriately the role of individual differences per se, we concentrated on those data for which the total number of item presentations per condition was experimenter controlled. Experiments 1, 2A, and 2B used a quasi-experimental design, making additional post hoc partitioning of the data a somewhat questionable exercise.

In Figure 8, immediate and delayed recall levels are again plotted as a function of the probe position of the pair being tested. The left panel shows the patterns for those subjects (high-recall participants) whose total immediate recall (across all probe positions) was above the median of 66%. The right panel shows the same results for those participants (low-recall participants) whose total immediate recall fell below the median. Note that the two patterns are highly similar: Both demonstrate primacy and recency effects at immediate test, but only primacy effects on the delayed test of cued recall. Such consistency supports the notion that those factors that motivate the differential retention of the primacy and recency items represent constraints inherent to human memory, rather than strategic selection of control processes. That is, the similarity between the patterns for high- and low-recall participants is consistent with the idea that some fundamental aspect of human memory, rather than some control process, drives the effects apparent in those patterns.

In Figure 9, however, we see that the pattern of predic-

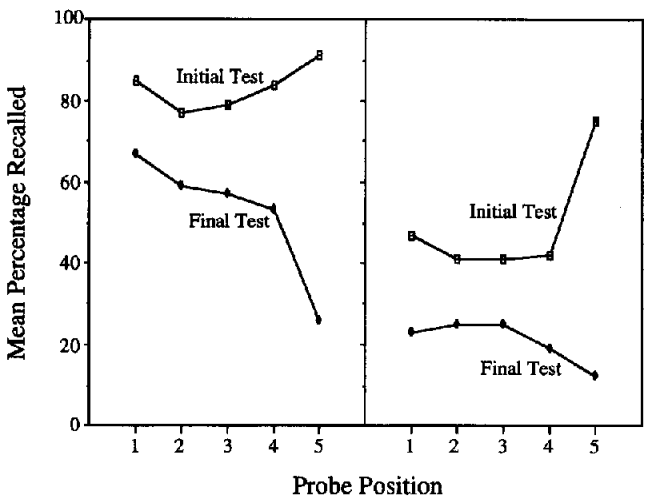

Figure 8. Mean rates of immediate and delayed cued recall as a function of probe position for high-recall participants (left panel) and low-recall participants (right panel; Experiment 3).

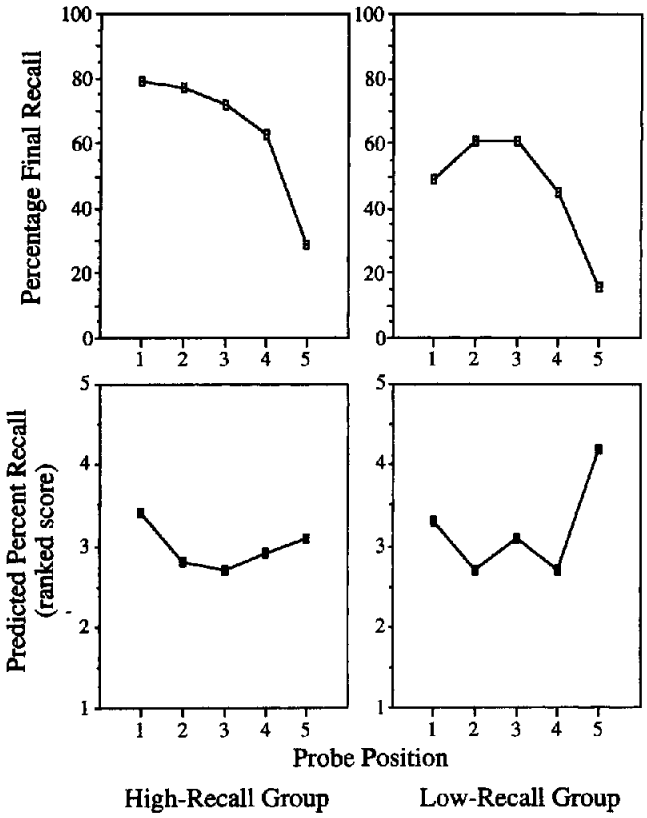

Figure 9. Mean rates of final cued recall conditionalized upon initial recall for high-recall participants (top left panel) and low-recall participants (top right panel). Mean predictions of future cued recall as a function of probe position for high-recall participants (bottom left panel) and low-recall participants (bottom right panel; Experiment 3).

tions of recall differs sharply between our two groups of participants. As in Figure 7, note that final recall conditionalized upon initial recall is plotted in the graphs immediately above the predictions graphs. Again, note that those items contributing to the recall data represented in the top panels are the exact same set that are contributing to the prediction data in the bottom panels.

High-recall participants do not provide elevated predictions for the recall of items from Position 5, but low-recall participants do. By comparing these patterns of prediction with the actual patterns of conditionalized final recall in the bottom panels and with the retention data presented in Figure 8, it is apparent that the predictions of high-recall participants are not misled seriously by elevated initial recency recall. However, the predictions by low-recall participants, particularly for the most recent item in each list, are more heavily influenced by level of initial recall. This finding is supportive of the notion that those participants that are better rememberers, either by means of superior intellectual capacity or better strategy selection, are also better metacognizers. Such participants apparently recognize, at least to some degree, the temporary nature of elevated recency-item retention. Those participants may then, in turn, use such metamnemonic information to guide the selection of which items on an exam might require further and more intensive study. It is a matter of future empirical and theoretical resolve whether better memory causes better metamemory, better metamemory causes better memory, or both derive from some more generalized intellectual ability.

Pedagogical Implications

One important implication of our findings is that we have identified a potentially educable component of metacognition. Although work in the domains of education research (e.g., Owings et al., 1980) and child development (e.g., Pressley, Levin, & Ghatala, 1984) has identified particular metacognitive failures to monitor learning, implement effective control processes, or both, there has been a dearth of research that has identified variables (such as feedback) that improve metamnemonic accuracy. In that sense—the recent advances in theories of metamemory notwithstanding—progress toward Nelson and Narens' (1994) goal of "*explain-[ing] (and eventually improv[ing]) the mnemonic behavior of a college student who is studying for and taking an examination*" has been lacking (p. 6, their italics).

We have drawn an analogy between one of our experimental paradigms and a situation in which students may find themselves and have suggested that the inability to appreciate certain differences between the ability to retrieve information in one way at one time and a potential retrieval at another time underlies some metamnemonic failures. We propose further that such failures to make correct metacognitive predictions derive from an oversimplified conceptualization of memory that participants hold. Benjamin and Bjork (1996) have argued that participants fail to appreciate (a) the multidimensional quality of memory, (b) the effects of time on memory, and (c) the effects of retrieval on memory. A rudimentary understanding of each of these notions has the potential to prevent the kinds of generalizations that mislead predictions in a manner akin to the three studies presented here.

Concluding Comments

We set out to demonstrate that retrieval fluency is an important index that humans rely on when making metacognitive judgments concerning future retention of information. It appears to be a component of a growing body of such indexes used in the formation of different metacognitive judgments. Our data support this idea that such metacognitive judgments are inferential in nature; in some sense, we make judgments of our own future performance subject to the same biases that might be implicated in our judgments of others' performance. To the extent that phenomenology plays a role in supporting metacognition, the aspects thereof that seem available for such judgments are limited, for example, to the speed or accuracy with which, or persistence by which, we retrieve information. We thus echo the claim of Miller (1962): "It is the *result* of thinking, not the process of thinking, that appears spontaneously in consciousness" (p. 56).

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81,* 126–131.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28,* 610–632.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1970, September). *Control processes and serial position effects in free recall.* Symposium on Memory, Mathematical Psychology Meetings, Miami Beach, FL.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 124–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Blake, M. (1973). Prediction of recognition when recall fails: Exploring the feeling-of-knowing phenomenon. *Journal of Verbal Learning and Verbal Behavior, 12,* 311–319.

Burke, D., 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.

Craik, F. I. M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior, 9,* 143–148.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20,* 374–380.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545–565.

Freedman, J. L., & Loftus, E. F. (1971). Retrieval of words from long-term memory. *Journal of Verbal Learning and Verbal Behavior, 10,* 107–115.

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

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732–764.

Groninger, L. D. (1979). Predicting recall: The "feeling-that-I-will-know" phenomenon. *American Journal of Psychology, 92,* 45–58.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56,* 208–216.

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning and Verbal Behavior, 6,* 685–691.

Jacoby, L. L., Bjork, R. A., & Kelley, C. M. (1994). Illusions of comprehension and competence. In D. Druckman & R. A. Bjork (Eds.), *Learning, remembering, believing: Enhancing team and individual performance* (pp. 57–80). Washington, DC: National Academy Press.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124,* 311–333.

Kucera, H., & Francis, W. N. (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. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Lee, V. A., Narens, L., & Nelson, T. O. (1993). *Subthreshold priming and the judgment of learning.* Manuscript submitted for publication.

Loftus, E. F., & Freedman, J. L. (1972). Effect of category-name frequency on the speed of naming an instance of the category. *Journal of Verbal Learning and Verbal Behavior, 11,* 343–347.

Madigan, S. A., & McCabe, L. (1971). Perfect recall and total forgetting: A problem for models of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 10,* 101–106.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 663–679.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122,* 47–60.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study time allocation? *Memory & Cognition, 18,* 196–204.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1263–1274.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff's amnesia. *Psychological Review, 100,* 3–22.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 851–861.

Miller, G. A. (1962). *Psychology: The science of mental life.* New York: Harper & Row.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64,* 482–488.

Murdock, B. B., & Walker, K. D. (1969). Modality effects in free recall. *Journal of Verbal Learning and Verbal Behavior, 8,* 665–676.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133.

Nelson, T. O. (1993). Judgments of learning and the allocation of study time. *Journal of Experimental Psychology: General, 122,* 269–273.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51,* 102–116.

Nelson, T. O., & Dunlosky, J. (1992). How shall we explain the delayed-judgment-of-learning effect? *Psychological Science, 3,* 317–318.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain" effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 476–486.

Nelson, T. O., & Narens, L. (1980a). A new technique for investigating the feeling of knowing. *Acta Psychologica, 46,* 69–80.

Nelson, T. O., & Narens, L. (1980b). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior, 19,* 338–368.

Nelson, T. O., & Narens, L. (1994). Why investigate metacogni-

tion? In J. Metcalfe and A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–26). Cambridge, MA: MIT Press.

Owings, R., Petersen, G., Bransford, J., Morris, C., & Stein, B. (1980). Spontaneous monitoring and regulation of learning: A comparison of successful and less successful fifth graders. *Journal of Educational Psychology, 72,* 250–256.

Postman, L., & Phillips, L. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology, 17,* 132–138.

Pressley, M., Levin, J. R., & Ghatala, E. S. (1984). Memory strategy monitoring in adults and children. *Journal of Verbal Learning and Verbal Behavior, 23,* 270–288.

Rabinowitz, J. C., Ackerman, B. P., Craik, F. I. M., & Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. *Journal of Gerontology, 37,* 688–695.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 435–451.

Reder, L. M., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 45–78). Hillsdale, NJ: Erlbaum.

Schacter, D. L., & Worling, J. R. (1985). Attribute information and the feeling-of-knowing. *Canadian Journal of Psychology, 39,* 467–475.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feeling of knowing. *Psychonomic Bulletin and Review, 1,* 357–375.

Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1074–1083.

Shimamura, A. P. (1994). The neuropsychology of metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 253–276). Cambridge, MA: MIT Press.

Siegler, R. S., Adolph, K. E., & Lemaire, P. (1996). Strategy choices across the life span. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 79–121). Hillsdale, NJ: Erlbaum.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3,* 315–316.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 382–403). New York: Academic Press.

Received August 7, 1996
Revision received February 11, 1997
Accepted April 2, 1997 ■