Memory & Cognition
2003, 31 (2), 297-305

Predicting and postdicting the effects of word frequency on memory

AARON S. BENJAMIN
University of Illinois at Urbana-Champaign, Champaign, Illinois

In the experiments reported here, I replicate and extend recent results that reveal that judgments about the memorability of common and uncommon words differ qualitatively depending on whether they are made during study or elicited during a recognition test (Guttentag & Carroll, 1998). When assessing recognition ability for individual words, subjects *predict* superior performance for common words, but *postdict* better performance for uncommon words. This interaction suggests that subjects rely on different cues when making judgments during study than they do when making analogous judgments during the recognition test, and that the cues utilized during recognition lead judgments to be more accurate. The shift is then evident in later predictions: Subjects who make postdictions consequently correctly *predict* superior recognition performance for uncommon words on a subsequent study list. When subjects are asked to make later predictions about *recall* performance, however, having made postdictions on a test of recognition does not mislead subjects into predicting superior recall performance for uncommon words.

The experiments reported here address the question of whether subjects understand the effects of linguistic word frequency on their own recognition ability, and whether or not the time and situation in which the metacognitive judgment is elicited affect those judgments. These issues are relevant to three separate but related domains, each of which will be described here briefly and returned to later. First, cases in which people mispredict the effects of variables on memory are particularly instructive regarding the mental models they hold of how memory works and regarding the heuristics that they use in determining what makes events or stimuli memorable. Second, the effects of test trials on metacognitive accuracy can be revealed in such experiments; that is, we can examine whether judgments change over the course of trials in such a manner as to indicate that subjects are learning about the effects that the experimental variables have on their own memory performance. Finally, the ability of subjects to accurately judge the effects of word frequency on recognition memory turns out to be a watershed question for prominent theories of recognition memory, and in particular of the effects of word frequency on recognition. Because the reasoning underlying this latter issue is a bit more complex, and because a review of this issue provides an opportunity to review the extant work in this domain—most of which has been motivated by this particular question—I will provide a more in-depth review of this topic.

The author wishes to express gratitude to Sharyn Krueger for her assistance in data collection, to Dan Schacter for helpful early suggestions concerning Experiment 3, to Sameer Bawa for assistance in data organization and summarization, and to John Dunlosky and Robert Greene for their thoughtful reviews of and feedback on the manuscript. Correspondence should be addressed to A. Benjamin, Department of Psychology, 603 E. Daniel St., University of Illinois, Champaign, IL 61820 (e-mail: asbenjam@s.psych.uiuc.edu).

Less common words are more likely to be recognized than more common words after study, and also are less likely to be falsely recognized if they were not studied (Glanzer & Bowles, 1976; Gorman, 1961). This effect is an example of a *mirror effect* (Glanzer & Adams, 1985): The condition that elicits a higher hit rate also elicits a lower false-alarm rate. According to certain prominent explanations of this effect, a study exposure benefits low-frequency (LF) words by virtue of a more efficient encoding, owing perhaps to the distinctiveness of the event (e.g., Schulman, 1967) or greater orienting of resources devoted to attending and memorizing rarer events (Glanzer & Adams, 1990). In addition, unstudied LF words elicit a lower false-alarm rate either because they are simply less familiar than high-frequency (HF) words (Glanzer & Bowles, 1976) or because subjects know LF words to be more memorable than HF words and thus set a higher standard for the recognition of those items (Benjamin, Bjork, & Hirshman, 1998; Brown, Lewis, & Monk, 1977; Gentner & Collins, 1981).

The difficulty with this explanation is that when subjects are actually asked to make judgments about the recognizability of words, they typically predict superior recognition performance for HF words. This result obtained regardless of whether subjects predicted future performance during study (Begg, Duft, LaLonde, Melnick, & Sanvito, 1989), made judgments about probable recognition during a mock recognition test in which none of the items had been studied (Wixted, 1992), and even when such a mock test and prediction phase followed an actual recognition test for which HF and LF words had previously been studied (Greene & Thapar, 1994).

A clever series of experiments recently reported by Guttentag and Carroll (1998) suggested a resolution to this discrepancy, however. In their experiments, subjects made judgments about the memorability of each stimulus imme-

diately after making a recognition decision for that particular stimulus. That is, for each test item that they claimed not to remember, they submitted a "postdiction" about the memorability for that item. Guttentag and Carroll found that such a procedure reversed the typical word frequency judgment effect: Subjects provided higher ratings to LF than to HF words.

The first experiment of the present paper replicates this result, and does so in a procedure in which the prior results concerning *pre*dictions of word frequency (e.g., Begg et al., 1989) were also replicated. Guttentag and Carroll (1998) failed to replicate the result that HF words were accorded higher judgments when they were embedded in a mock recognition test (Greene & Thapar, 1994; Wixted, 1992). Experiment 1 shows that a shift from predicted HF superiority to predicted LF superiority can occur not only in a within-subjects design but even on a within-items basis. Experiments 2 and 3 addressed the consequences of making postdictions on future predictions about recognition (Experiment 2) and recall (Experiment 3).

EXPERIMENT 1

In this experiment, subjects studied a series of high- and low-frequency words and attempted to recognize them on a later test. During study, they were asked, for each word, to make a prediction of the likelihood of being able to recognize that word later. Then, for every item that they claimed not to recognize at test, subjects made postdictions of their belief that they *would* have recognized it if it had been studied.

The critical prediction is of a crossover interaction between word frequency and test phase on metacognitive judgments. Specifically, at time of study, subjects should predict higher rates of recognition for HF than LF words, thus replicating Begg et al. (1989); however, at test, subjects should postdict higher rates of recognition for LF than for HF words (as in Brown et al., 1977).

Method

Subjects. Fifty undergraduate students participated in order to partially fulfill course requirements. Eight were male and 42 were female. The mean age was 21.4 years, they had had an average of 15.2 years of education, and mean performance on the Mill-Hill test of vocabulary was 52%.

Design. The experiment employed a 2 (word frequency) $\times$ 2 (time of prediction) within-subjects design in which predictions were collected. In addition, rates of endorsement in recognition were gathered for HF and LF items that were or were not studied (also 2×2, within subjects).

Materials. The studied words were 4–8 letter nouns, verbs, and adjectives obtained from the compendium provided by Carroll, Davies, and Richman (1973). The 80 HF words ranked 100–270 on their scale, and the 80 LF words ranked 5,000–5,230. The study list consisted of 40 HF and 40 LF items randomly intermixed within the constraints that (1) each half of the study list contained an equal number of HF and LF items, and (2) no more than three items of a particular frequency appeared in a row.

The test list consisted of 160 items, 80 of which were the previously studied words and 80 (40 HF and 40 LF) of which were unstudied. Each quarter of the test list contained an equal number of HF and LF items, as well as an equal number of old and new items. Test lists were generated randomly subject to the constraints mentioned above. All presentation of stimuli and recording of responses were done on PC microcomputers programmed in QBASIC.

Procedure. Subjects were tested individually in a small, well-lit room. Prior to study, each subject read a set of instructions informing him/her that he/she would need to study the upcoming words for a recognition test and make predictions about their ability to recognize each word later. The nature of the recognition test was explained in some detail; in particular, it was emphasized that subjects would be making "yes" and "no" responses as to whether an item had been previously studied for some words that they had actually studied and some that they had not. Several examples were provided.

During the study phase, each word was presented for 2 sec and then removed from the screen. At that point subjects were prompted for the prediction on a scale of 1 to 9. During the entire study phase, a scale at the bottom of the screen reminded them that "1" indicated "I am sure that I will NOT remember this word" and that "9" indicated "I am sure that I WILL remember this word," with all gradations in between. After subjects had entered their one-key prediction and the a 1-sec interval had elapsed, the next word appeared.

After subjects cycled through the entire study list, there was a short break (15 sec) before the instructions for recognition were presented. Those instructions reminded subjects about the nature of the recognition test and further informed them of the judgment that they needed to make whenever they responded negatively to the recognition inquiry. Subjects were told that they would need to decide if they *would* recognize this word if they *had* studied it. So if subjects saw the word *tincture* and believed that they had not seen it (as indicated by an "N" response on the recognition test), they then made a judgment on the subjective likelihood of recognizing it if they had studied it. As during study, a scale remained on the bottom of the screen reminding them of the anchors of the scale. Similar to the prediction scale, "1" indicated "I am sure I would NOT recognize this word" and "9" indicated "I am sure I WOULD recognize this word." Each test word remained on the screen while subjects made their recognition judgments (and postdictions, when necessary), which were self-paced. After both judgments were made, there was a 1-sec interval before the next word appeared.

Results

The results of all inferential statistics reported below and throughout this article are reliable at the $\alpha = .05$ level using two-tailed tests unless otherwise noted. Figure 1 shows the recognition performance for HF and LF words. As is typical, there was an interaction between word frequency and study status such that unstudied HF words were (falsely) recognized more frequently than unstudied LF words, but studied LF words were (correctly) recognized more frequently than studied HF words [$F(1,49) = 77.19$].

The results from the judgment task are presented in the top half of Figure 2. Critically, the interaction between word frequency and time of judgment was reliable [$F(1,49) = 34.66$]. There was a trend for HF words to be accorded higher predictions than LF words [$t(49) = 1.63$, $p < .05$, one-tailed], but LF words were accorded higher postdictions than HF words [$t(49) = 4.83$]. It should be noted that all of the differences discussed here are quite small, owing to the dramatically different ways in which the rating scales were used across subjects.

In the bottom half of Figure 2 are presented the judgment data specifically for missed items—those words that were studied and nonetheless rejected on the recognition

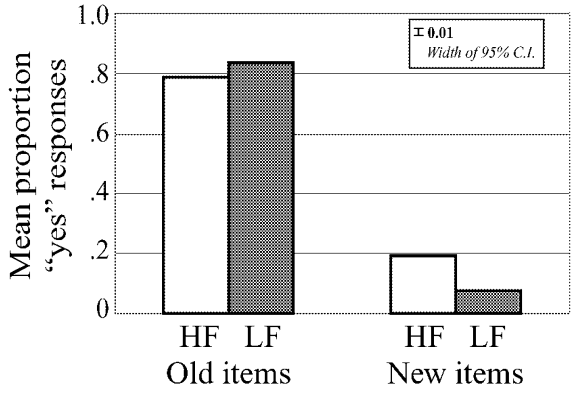

Figure 1. Recognition performance as a function of word frequency (Experiment 1). Error bars (top right corner of the figure) represent the 95% confidence interval based on within-subjects interaction variability. Note—This confidence interval is thus the one suggested for use by Loftus and Masson (1994), but adapted for use in a multifactor design. The error term represents the variability in the highest order interaction in the design, namely A × B × S, and, because the design is entirely within-subjects, it is thus scaled by a criterion *t* value based on (*n* − 1) degrees of freedom. The error bars are not placed on the bars representing the means themselves, however, because the interaction variability does not provide the appropriate error term for pairwise comparisons. Throughout this paper, for figures displaying data for which the interaction was the contrast of primary import (including all of the recognition performance figures), variability is depicted in this manner. For tests in which main effects were of primary interest, error bars are plotted on the data directly and represent the 95% confidence interval based on within-subjects variability in the factor of interest, exactly as suggested by Loftus and Masson.

test. There was an average of 15.2 such items per subject. A similar interaction to the one apparent for all rejected items obtained: HF words elicited higher judgments during prediction, but lower judgments during postdiction [$F(1,44) = 4.43$].[1] This result is particularly interesting because it demonstrates the effect on an entirely within-items basis. Shown in Table 1 are the mean Goodman-Kruskal gamma correlations between predictions and recognition accuracy for this and the following two experiments. None of the differences in gamma between conditions or experiments are reliable.

Discussion

The results of this experiment demonstrate that shifting the time of metacognitive judgment from study to test can have an effect on the way that such judgments are made. At study, the finding of Begg et al. (1989) was replicated: Subjects predicted higher future rates of recognition for HF than for LF words. This result is consistent with the view that subjects use ease of perception (Benjamin & Bjork, 1996), ease of conceptual processing (Begg et al., 1989), or familiarity (Metcalfe, Schwartz, & Joaquim, 1993; Reder & Ritter, 1992) as a basis for their metacognitive judgment.

The fact that this result reversed qualitatively when the judgments were made at test, replicating Guttentag and Carroll (1998), suggests a shift in the bases for the judgment. In particular, it seems that the suggestion of Brown et al. (1977) is consistent with the results: When viewing uncommon words, subjects accurately classify them as ones that they would be more likely to recognize. By the same token, perhaps, they realize the difficulty of mentally localizing a common word to the study list. Highly familiar words could have been seen in many places and many times prior to the list presentation, and subjects recognize the difficulty of picking out that one presentation from among the many they have experienced.

One possibility is thus that the act of making postdictions alerts subjects to the discrimination component of the recognition task. During study, the reliance on familiarity or fluency reveals an implicit assumption that prior experience with the stimulus should translate into memorability for that stimulus. This assumption would be correct if the subjects had been given a recall test (as evidenced by the superior recall of HF words), but it is incorrect on recognition, where the burden for the subject is on the discrimination of stimuli rather than the generation of previously seen words.

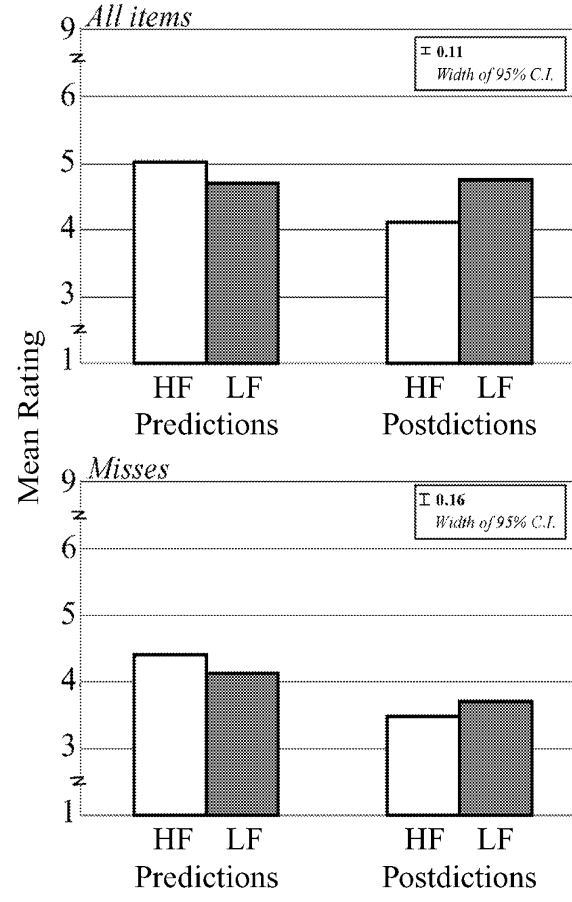

Figure 2. Mean ratings as a function of word frequency and time of judgment for all rated items (top half) and misses (bottom half; Experiment 1).

Table 1
Mean Goodman-Kruskal Gamma Correlations Between Predictions and Accuracy for All Items and Low-Frequency (LF) and High-Frequency (HF) Items Selectively

	Gamma				
		LF		HF	
	All	UC	C	UC	C
Experiment 1	.26	.28	.29	.24	.22
Experiment 2					
No-postdiction condition					
Test 1	.27	.30	.33	.22	.23
Test 2	.32	.31	.31	.26	.26
Postdiction condition					
Test 1	.37	.38	.38	.33	.29
Test 2	.47	.22	.19	.45	.32
Experiment 3					
Control condition (recall)	.37	.31		.41	
Postdiction condition					
Test 1	.36	.40	.37	.36	.36
Test 2 (recall)	.37	.41		.35	

Note—Uncorrected scores are the means of all subjects excluding those conditions in which gamma was undefined. Like any correlation measure, gamma is undefined when variability in either variable is 0. In the present experiments, this occasionally occurred when the hit rate for a condition was 1. Because this situation arose more often for low-frequency than high-frequency words [e.g., for 16 vs. 5 subjects on Test 1 in Experiment 2], there is some subject-selection contamination in the uncorrected scores. The corrected scores are also biased by this artifact, but the frequency conditions are biased equally by dropping all subjects for whom gamma was undefined for either frequency condition. Because perfect and zero scores were so rare when collapsed across frequency, and on the tests of recall, the correlations involving all recognition items, as well as predictions and recall, are not corrected in such a manner. Corrected scores are the means for only those subjects for whom gamma was defined for both low- and high-frequency words. UC, uncorrected score; C, corrected score.

If the simple act of making postdictions actually had the effect of informing subjects' mental models of the source evaluation component of the recognition task, then it would be expected that they should be able to incorporate such knowledge into future predictions about recognition. Another possibility is that the even simpler act of participating in the test of recognition—rather than having to engage in any kind of explicit metacognitive activity—is enough to reveal to subjects their misconceptions about word frequency effects in recognition. Experiment 2 addressed this question.

EXPERIMENT 2

In this experiment, subjects cycled through two study–test procedures similar to the one described in Experiment 1. One group simply replicated the Experiment 1 procedure twice, going through two consecutive study-prediction and test-postdiction phases. The other group went through the same procedure with one exception: During the first study–test cycle, they only made *pre*dictions about recognition. Thus, by the time they encountered the second study list, they had had a test for the first list, but no opportunity to make postdictions during that test. If the act of engaging in explicit metacognition during the test is crucial to informing subjects of their misappreciation of word frequency effects, then only the predictions from the first group should reveal a sensitivity to actual word frequency effects in their second-phase predictions for recognition. If the recognition test is sufficient to make subjects aware of the role of word frequency in recognition, then the second group should show similar sensitivity to this factor.

Method

Subjects. Seventy undergraduates (20 males and 50 females) participated in the experiment for course credit.

Design. Each subject participated in two study–test phases, and recognition performance was collected in each, yielding a 2 (word frequency) × 2 (Test 1 or 2) matrix of recognition performance. Presence of the postdiction stage during Phase 1 was manipulated between subjects, making the overall judgment design mixed in nature. In the no-postdiction condition, judgments were made twice at study and once at test, during Phase 2, yielding a 2 (word frequency) × 3 (time of judgment) design. In the postdiction condition, judgments were made at all four stages, yielding a 2 (word frequency) × 2 (Study Phase 1 or 2) × 2 (Test Phase 1 or 2) design.

Materials. The same word pool was used as in Experiment 1. The only difference between the two experiments was that each study and test list from Experiment 1 was broken into two equally sized lists for Experiment 2. Thus, each study list contained a total of 40 items, half from each frequency condition, and each test contained 80 items, half of which were studied. All counterbalancing details and ratios were the same as in Experiment 1.

Procedure. All details of presentation and timing were the same as in Experiment 1 with the following exceptions. First, following the first study–test phase and a 1-min distractor interval, the second study–test phase occurred. Second, for subjects in the no-postdiction condition, the recognition test during Phase 1 took place without any opportunity for the subjects to make metacognitive judgments.

Results

Recognition performance is presented in Figure 3. In each of the study–test phases, and for each of the between-subjects conditions, the standard pattern of LF superiority obtained. The interactions were all reliable [no-postdiction: Fs(1,35) = 24.04, 45.32 (Test 1, Test 2); postdiction: Fs(1,33) = 33.15, 37.06 (Test 1, Test 2)]. There were no effects of postdiction or test number on recognition, nor did either of these variables interact with word frequency or old/new status on the recognition test.

Performance on the metacognitive tasks from Phase 1 are presented in the left half of Figure 4. Subjects in the postdiction condition replicated Experiment 1 in Phase 1: HF words elicited higher predictions during study [$t(33)$ = 2.03], but lower postdictions at test [$t(33)$ = 2.85, interaction: $F(1,33)$ = 20.41]. In addition, subjects in the no-postdiction condition showed the typical effect of word frequency at study, with HF words being accorded higher judgments [$t(35)$ = 2.57].

The right half of Figure 4 shows the judgment data from the second phase of the experiment. The postdiction group showed a reversal of the typical pattern apparent at study: LF words were accorded higher predictions [$t(33)$ = 2.09]. No differences in Phase 2 predictions for the no-postdiction group were apparent. The simple interaction

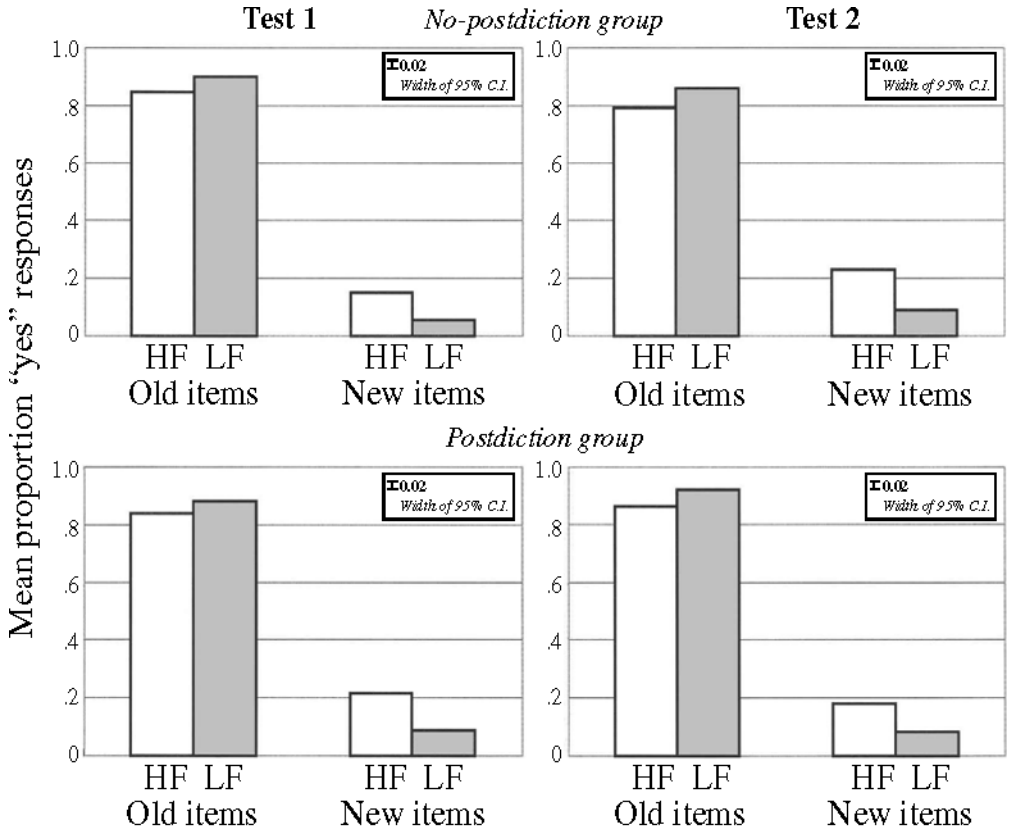

Figure 3. Recognition performance as a function of word frequency, prior study status, postdiction/control condition, and study/test phase (Experiment 2).

between prediction/no-prediction and word frequency on Phase 2 predictions was marginally reliable [$F(1,68) = 2.79, p < .09$]. Both groups attributed higher ratings to LF than to HF words during the Phase 2 test [no-postdiction group, $t(35) = 2.42$; postdiction group, $t(33) = 4.03$].

Discussion

The results from Experiment 2 show that the act of making postdictions can actually benefit future prediction performance: Those subjects that were given an opportunity to make metacognitive judgments during the recognition test of the first phase rectified their judgments during study of the second phase and correctly predicted superior recognition for LF items. The case for the group that had the opportunity to engage in recognition, but not the metacognitive judgments, is less clear. In the second phase, there was no evidence that they continued to incorrectly predict superior HF recognition, nor that they corrected their predictions. It is apparent, however, that the act of making postdictions confers a benefit in terms of later predictive capacity, even when compared with a group that went through the same recognition procedure.

The evidence for the benefit of making postdictions on the correlation between predictions and performance is shown in Table 1. Although the mean correlations rise from Test 1 to Test 2 in both conditions, and more in the postdiction than in the no-postdiction condition, none of these effects are statistically reliable. Of course, word frequency plays only a minor role in determining recognition performance, relative to the myriad of other word-related and subject-related factors not under experimental control (in this experiment, R^2 between the binary variables of word frequency and the probability of a positive response for studied items was less than .01). So, even if their predictions "improved" in the sense that they became additionally sensitive to word frequency—and otherwise remained the same with respect to the other variables that were incorporated into the judgment—this experimental design would have very little power to detect such a small effect.

EXPERIMENT 3

Although it seems that subjects in Experiment 2 learned something about what they were likely to remember, it is not apparent exactly what that something is. One likely possibility is that, through observing their own performance, they have noticed that uncommon words are more memorable than they had originally thought (or, similarly, that common words are less memorable). In that case, the

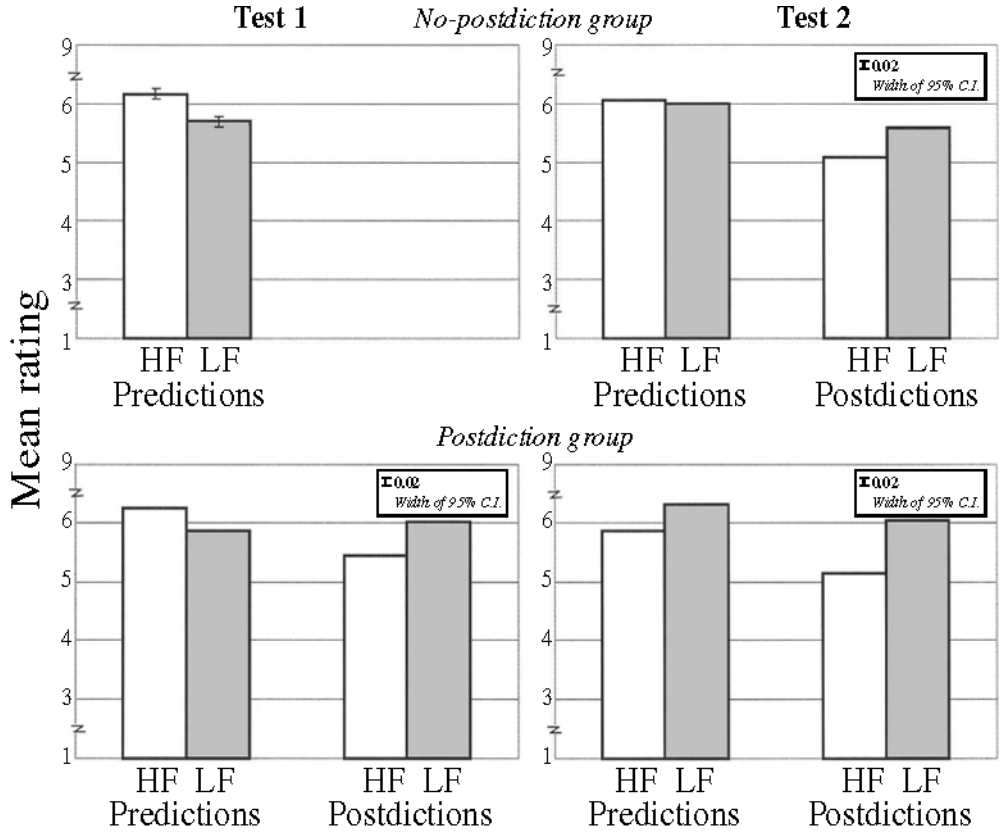

Figure 4. Mean ratings as a function of word frequency, time of judgment, postdiction/control condition, and study/test phase (Experiment 2).

interpretation of the prior results is not that they have had any particular insight about recognition as a mnemonic task, but rather that they have now been convinced that uncommon words are simply more memorable than common ones. If this construal is accurate, then it should be possible to elicit a new metacognitive error from subjects, one that they would not have made prior to the experiment. Specifically, if they are administered a test of recall after postdicting recognition performance, they should now mispredict superior recall of uncommon words.

If, however, the act of engaging in metacognitive reflection during the recognition test has illuminated the fact that uncommon words are more *discriminable*—rather than more memorable—then subjects should be able to correctly predict the superiority of HF word recall, much as they would have had they had no intervention at all. This pattern of results should arise if two conditions are met: (1) that subjects learned that the advantage that LF words possess lies in their easier discriminability, and (2) that recognition performance, but not recall performance, hinges critically on discrimination.

To review, mispredictions of future recall following recognition (and postdictions) would indicate an erroneous generalization that uncommon words are more memorable than common words. Correct predictions of recall would indicate that subjects accurately realize the advantage that uncommon words are afforded in tests that emphasize discrimination. In this experiment, we compared two groups, both of whom predicted their future recall, but only one of which had a previous study–recognition phase in which they both pre- and postdicted their recognition performance.

Method

Subjects. Sixty-two undergraduates participated in the experiment for course credit.

Design. One control group of subjects (26) engaged in a single study–test cycle. For subjects in this group, there were two dependent variables (recall and mean prediction scores) and one independent variable (word frequency). The other group of subjects (36) participated in two study–test phases in a manner analogous to the postdiction group in Experiment 2. The first phase utilized a recognition test, and the second, a test of recall. Subjects in this group thus made judgments at three points: twice during prediction (once for recognition and once for recall) and once during recognition postdiction for Phase 1. The memory performance design matrix was thus 2 (word frequency) × 2 (Phase 1 or 2: recognition or recall). The prediction design matrix was 2 (word frequency) × 2 (Phase 1 or 2: recognition or recall), and postdictions for LF and HF words were gathered in Phase 1.

Materials. Phase 1 of Experiment 3 was identical to Phase 1 for the postdiction group in Experiment 2. In Phase 2, subjects again studied 40 items, half of which were HF and half of which were LF. For Phase 2 recall, subjects recorded the remembered items on a blank sheet of paper. Again, the details of counterbalancing were the same as in Experiment 1.

Procedure. As described above, subjects in the experimental group went through two study–test phases. The first was equivalent to the first study–test phase in Experiment 2. Prior to the second phase, subjects in the experimental group were informed of the basic nature of a test of free recall—namely, that they would be presented with a blank sheet of paper and asked to write down as many of the previously studied words as possible. They then studied the to-be-recalled list (again, 2 sec/word) and, after the presentation of each item, made a judgment as to the probability of recalling that item on the upcoming test. After the study phase and a short distraction interval (30 sec), subjects were given a blank sheet of paper and asked to recall as many words from the prior list as possible. They were required to spend no less than 7 min and no more than 10 min on the recall portion of the task. The control group went through only a single study–test phase, which was equivalent in every detail to the second phase for the experimental (postdiction) group.

Results

Recognition performance in Phase 1 revealed a reliable mirror effect [$F(1,35) = 19.91$] and is shown in the top half of Figure 5. Recall performance, shown in the bottom half of Figure 5, was higher for HF than for LF words in both conditions, revealing the typical reversal of word frequency effects between recall and recognition [$t(25) = 2.67$ (no Phase 1 group); $t(35) = 2.81$ (postdiction group)].

Means for judgments during Phase 1 are shown in the top half of Figure 6. Predictions were higher for HF than for LF words [$t(35) = 1.60$, $p = .05$, one-tailed]. Postdictions were higher for LF than for HF words [$t(35) = 2.49$], and the interaction seen in the first two experiments between word frequency and time of judgment was replicated [$F(1,35) = 9.99$]. Predictions of recall (from Phase 2) are presented in the bottom half of Figure 6, and reveal that higher ratings were accorded to the HF than to the LF words in both the control condition [$t(25) = 2.07$] and the postdiction condition [$t(35) = 3.27$]. There was no interaction between frequency and control/postdiction condition ($F < 1$).

Discussion

The results from Experiment 3 reveal that the act of making postdictions of recognition performance during recognition does not affect later predictions of recall. This result is inconsistent with the view that subjects were incorrectly learning that uncommon words are more memorable than uncommon words during the recognition/postdiction phase of Experiments 1–3. Rather, it is consistent with the interpretation that they have learned something new about the nature of the task of *recognition*—namely, that it involves a large discrimination component that is easier for uncommon than for common words. The result that predictions of recall are not misled after postdictions of recognition provides the strongest evidence of this sophisticated learning about the nature of recognition.

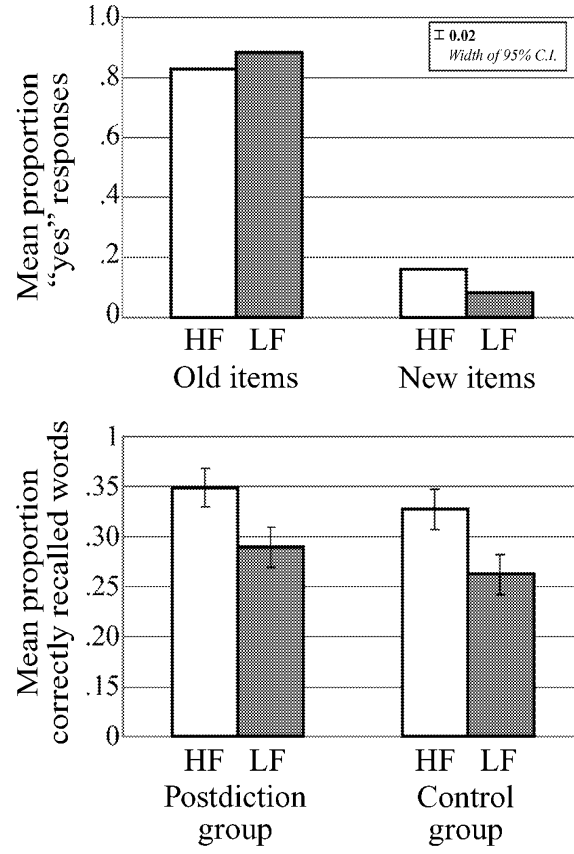

Figure 5. Recognition (top half) and recall (bottom half) performance as a function of word frequency (Experiment 3).

GENERAL DISCUSSION

The interpretation of the word frequency mirror effect in recognition described earlier hinges critically on the ability of subjects to accurately evaluate the greater recognizability of lower frequency words, and this view was cast into doubt by the result that subjects predicted greater recognition for HF words (e.g., Begg et al., 1989; Greene & Thapar, 1994; Wixted, 1992). However, as others (Guttentag & Carroll, 1998) have shown, and as I have replicated here, subjects do appreciate the greater recognizability of LF words during recognition itself. Since the mirror effect arises at the time of the recognition test, the present results render plausible an interpretation such as that suggested by Glanzer and Adams (1990).

Plausibility notwithstanding, the question remains as to whether the ability to consciously report the superiority of recognition of uncommon words actually underlies the word frequency mirror effect. Future theoretical endeavors that attribute differences in false-alarm rates to a conscious appreciation of the differential difficulty of recognizing HF and LF words must confront several problems. First, judgments of word frequency effects in recognition are miscalibrated even immediately after a recognition test on which a mirror effect obtains (Experiment 2; see also Greene & Thapar, 1994; Guttentag & Carroll, 1998; Wixted, 1992). Second, mirror effects obtain in frequency discrimination as well as recognition (Greene & Thapar, 1994), which suggests that the discrimination explanation of the effect (e.g., Glanzer & Bowles, 1976) may be incomplete.

More generally, the three experiments reported here have shown that judgments about the memorability of

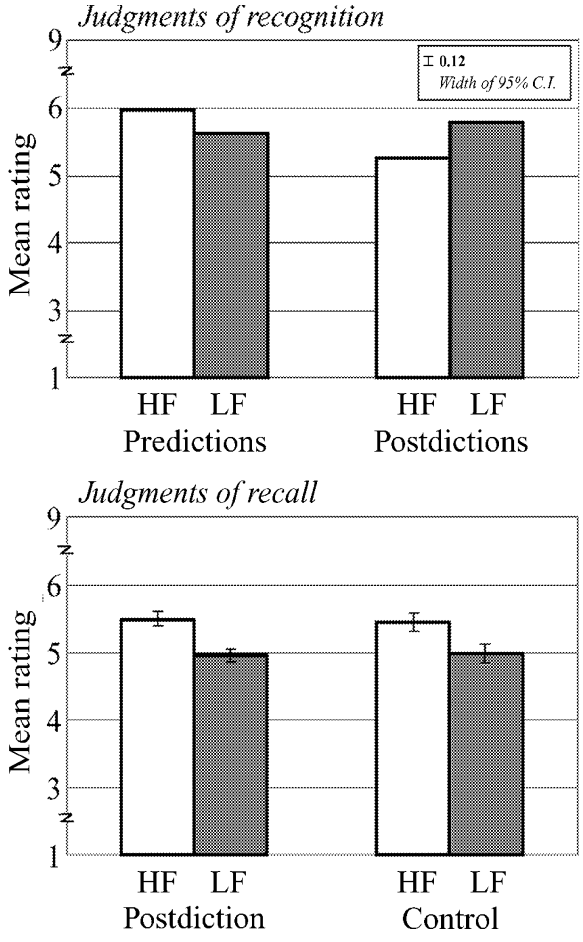

Figure 6. Top panel: Mean ratings of recognition as a function of word frequency and time of judgment (postdiction group). Bottom panel: Mean predictions of recall as a function of word frequency for the postdiction and control groups.

stimuli differed depending on whether those judgments were predictive or postdictive in nature. In particular, it has been proposed that the specific cues available for judgments vary from time to time, and situation to situation, and that one way in which humans can be trained to hone in on the ones that are maximally predictive is to ask them to make retrospective judgments about potential memorability during the criterion test (Experiments 1–3). The fact that subjects who do so learn to correctly predict the effects of word frequency on recognition during a later study session (Experiment 2) shows that the effects are not attributable to general differences between making predictions and postdictions, but likely reflect a correction of a particular misapprehension about word frequency effects that subjects hold in their implicit mental model of the operation of memory. This interpretation is also consistent with prior results showing that subjects do not correctly predict LF word recognition superiority even after engaging in a test of recognition (Greene & Thapar, 1994; Guttentag & Carroll, 1998).

However, the results do not suggest that the subjects simply learned that their initial beliefs about the greater memorability of HF words were incorrect. In fact, they revert to these beliefs when asked to make predictions of recall (Experiment 3). It thus appears that the acquired metacognitive knowledge is not that uncommon words are more memorable than common ones, but rather that recognition but not recall involves discrimination, and discrimination is easier for uncommon than for common words. This understanding allows subjects to correctly predict recognition (Experiment 2) and recall performance (Experiment 3), despite the opposing effects of word frequency on these different tasks.

These findings are also generally consistent with other work suggesting that experience with a task improves the accuracy with which subjects predict the relative efficacies of different orienting tasks during study (e.g., Bieman-Copland & Charness, 1994; Brigham & Pressley, 1988; Dunlosky & Hertzog, 2000). Although these experiments were not designed to and consequently did not reveal an overall improvement in prediction accuracy, as measured by correlations between predictions and performance, across study–test trials, I have argued that subjects do learn something about the nature of recognition over trials, and that this learning is revealed by the effects of word frequency on mean predictions of recognition. I have also claimed that the learning evident in the experiments reported here is of an analytic nature; that is, it is a conscious revision of the relative roles of word memorability and word discriminability that allows subjects to correctly predict the effects of word frequency on recognition. Note that this suggestion stands in contrast to the one presented earlier concerning the mediating role of factors such as ease of processing (Begg et al., 1989) on initial predictions of recognition, a process one might think of as nonanalytic (Koriat, 1997). Consistent with the interpretation provided here relating analytic processes to the improvement in predicting the memorial effects of word frequency, some reports indicate that predictions across multiple test trials improve in young adolescents but not children (Pressley & Ghatala, 1989).

The suggestion that predictions and postdictions of memory performance can be based on different cues is at the heart of the dissociations presented here and by others (e.g., Devolder, Brigham, & Pressley, 1990; Hertzog, Saylor, Fleece, & Dixon, 1996). As a final example of the importance of the distinction, and of the centrality of self-assessment in the improvement of metamnemonic accuracy, consider a recent result from Pritchard and Keenan (1999). In their experiment, mock jurors engaged in "deliberation" made predictions about their ability to recall trial-relevant information and then took a test of their knowledge. Those predictions were very poorly calibrated with actual knowledge, yet their postdictions after the test were closely related to their test performance.

This finding illustrates what may be a common scenario: poor self-assessment of one's own memory ability and, by extension, of the effects of different variables on one's memory. Yet a test coupled with a mandatory assessment of likely performance revealed these shortcomings to the jurors. Presumably, a conscious recognition of

their own failings—the metacognitive ones, as well as the mnemonic ones—would be necessary for them to improve their predictions in the future. No juror can retain all of the statements, evidence, and arguments that have been presented over the course of a lengthy trial, but we can hope that he/she might have the insight to review the transcripts when necessary. This may hold in cases in which the education of subjective experience is even more important than the improvement of memory performance (see Ghodsian, Bjork, & Benjamin, 1997).

REFERENCES

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

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

Benjamin, A. S., Bjork, R. A., & Hirshman, E. (1998). Predicting the future and reconstructing the past: A Bayesian characterization of the utility of subjective fluency. *Acta Psychologica*, **98**, 267-290.

Bieman-Copland, S., & Charness, N. (1994). Memory knowledge and memory monitoring in adulthood. *Psychology & Aging*, **9**, 287-302.

Brigham, M. C., & Pressley, M. (1988). Cognitive monitoring and strategy choice in younger and older adults. *Psychology & Aging*, **3**, 249-257.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency, and negative recognition. *Quarterly Journal of Experimental Psychology*, **29**, 461-473.

Carroll, J. B., Davies, P., & Richman, B. (1973). *Word frequency book*. New York: American Heritage.

Devolder, P. A., Brigham, M. C., & Pressley, M. (1990). Memory performance awareness in younger and older adults. *Psychology & Aging*, **5**, 291-303.

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology & Aging*, **15**, 462-474.

Gentner, D., & Collins, A. (1981). Studies of inference from lack of knowledge. *Memory & Cognition*, **9**, 434-443.

Ghodsian, D., Bjork, R. A., & Benjamin, A. S. (1997). Evaluating training during training: Obstacles and opportunities. In M. A. Quiñones & A. Ehrenstein (Eds.), *Training for twenty-first century technology: Applications of psychological research*. (pp. 63-88). Washington, DC: American Psychological Association.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, **13**, 8-20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 5-16.

Glanzer, M., & Bowles, N. (1976). Analysis of the word frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 21-31.

Gorman, A. N. (1961). Recognition memory for names as a function of abstractness and frequency. *Journal of Experimental Psychology*, **61**, 23-29.

Greene, R. L., & Thapar, A. (1994). Mirror effect in frequency discrimination. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 946-952.

Guttentag, R., & Carroll, D. (1998). Memorability judgments for high- and low-frequency words. *Memory & Cognition*, **26**, 951-958.

Hertzog, C., Saylor, L. L., Fleece, A. M., & Dixon, R. A. (1996). Metamemory and aging: Relations between predicted, actual, and perceived memory task performance. *Aging & Cognition*, **1**, 203-237.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, **126**, 349-370.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, **1**, 476-490.

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

Pressley, M., & Ghatala, E. S. (1989). Metacognitive benefits of taking a test for children and young adolescents. *Journal of Experimental Child Psychology*, **47**, 430-450.

Pritchard, M. E., & Keenan, J. M. (1999). Memory monitoring in mock jurors. *Journal of Experimental Psychology: Applied*, **5**, 152-168.

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, & Cognition*, **18**, 435-451.

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science*, **9**, 211-212.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 681-690.

NOTE

1. The degrees of freedom for this test are reduced because some subjects failed to contribute a score to one of the cells.

(Original manuscript received May 17, 2002;
revision accepted for publication October 2, 2002.)