

**Why Do People Show Minimal Knowledge Updating with Task Experience:
Inferential Deficit or Experimental Artifact?**

Christopher Hertzog, Jodi Price, Ailis Burpee, William J. Frentzel, Simeon Feldstein,

Georgia Institute of Technology

& John Dunlosky

Kent State University

11/27/2007

Running Head: Strategy Knowledge

Address Correspondence to:

Christopher Hertzog
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332-0170
(404) 894-6774
FAX: (404) 894-8905
E-mail: christopher.hertzog@psych.gatech.edu

Abstract

Students generally do not have highly accurate knowledge about strategy effectiveness for learning, such as that imagery is superior to rote repetition. During multiple study-test trials using both strategies, participants' predictions about performance on List 2 do not markedly differ for the two strategies, even though List 1 recall is substantially greater for imagery. Two experiments evaluated whether such deficits in knowledge updating about the strategy effects were due to an experimental artifact or to inaccurate inferences about the effects the strategies had on recall. Participants studied paired associates on two study-test trials—they were instructed to study half using imagery and half using rote repetition. Metacognitive judgments tapped the quality of inferential processes about the strategy effects during the List 1 test and tapped gains in knowledge about the strategies across lists. One artifactual explanation – noncompliance with strategy instructions -- was ruled out, whereas manipulations aimed at supporting the data available to inferential processes improved but did not fully repair knowledge updating.

Why Do People Show Minimal Knowledge Updating with Task Experience: Inferential Deficit or Experimental Artifact?

Effective cognition often requires choosing appropriate strategies to achieve desired outcomes (e.g., Crowley, Shrager, & Siegler, 1997; Lemaire & Siegler, 1995). Hence, knowledge about the effectiveness of processing strategies is an important aspect of self-regulation during learning (Hertzog & Dunlosky, 2004; Schneider & Pressley, 1997). This study examines whether people learn about the differential effects of strategies as they use them to study and retrieve new associations. A number of different strategies exist for forming new associations between unrelated word pairs (e.g., TICK-SPOON). One highly effective strategy is the use of interactive imagery, which leads to superior cued recall relative to using rote repetition -- simply repeating to-be-learned associations (Richardson, 1998). This imagery benefit occurs in large part because retrieving an imaginal mediator when cued at test is highly likely to result in recovery of the target word (e.g., Dunlosky, Hertzog, & Powell-Moman, 2005).

Many individuals lack pre-experimental knowledge that imagery is superior to rote repetition for associative learning (e.g., Dunlosky & Hertzog, 2000; Shaughnessy, 1981), possibly because memorizing paired associates is an uncommon activity in everyday life. An important question, then, is whether individuals can learn about the differential efficacy of these strategies through structured task experience. As we discuss below, recent evidence indicates that students do learn about interactive imagery's superiority across multiple study-test trials, but this learning is relative and incomplete, with poorly calibrated metacognitive judgments. This study sought to better understand why the accuracy of metacognitive judgments does not better reflect the difference in effectiveness of these two strategies.

Measuring Students' Learning about Strategy Effects through Task Experience

Gains in strategy knowledge have been measured by collecting metacognitive judgments during multiple study-test trials (e.g., Brigham & Pressley, 1988; Dunlosky & Hertzog, 2000; Matvey et al., 2002; Pressley, Levin, & Ghatala, 1984). Typically, students judge strategy effectiveness prior to practice to obtain baseline measures of pre-experimental strategy knowledge. Updated strategy knowledge has been measured by changes in judgments as students experience strategy use. In our previous research, participants have been asked to make *global differentiated predictions* (GPREDs), which involve predicting the percentage of items that will be recalled under each of the strategies prior to task experience, to assess pre-experimental knowledge. After studying each item, participants also make *judgments of learning* (JOLs), rating the likelihood of recalling a given item on the upcoming test. For an initial study list,¹ both types of predictions do not differ markedly for imagery and rote items, suggesting that many students do not believe that imagery is the superior strategy (Dunlosky & Hertzog, 2000).

Most important, gains in strategy knowledge after study and test experience could be manifested in increased sensitivity of second-list GPREDs to the different strategies used during encoding. For instance, updated strategy knowledge could be reflected in greater separation of GPREDs for imagery and rote made prior to List 2 (tapping newly acquired knowledge) than are made prior to List 1 (based on pre-experimental knowledge). Similarly, updated strategy knowledge can be reflected in the sensitivity of List 2 JOLs to encoding strategies (for details, see Dunlosky & Hertzog, 2000). If so, JOLs will be more accurate at predicting recall on List 2.

Our central focus is on the absolute accuracy of GPREDs and JOLs, because both measures markedly underestimate the effects of imagery on List 2 performance (Dunlosky & Hertzog, 2000). Moreover, we also included a questionnaire that tapped people's beliefs about the effectiveness of these strategies, confidence judgments, and global differentiated

postdictions. The questionnaire measure taps declarative knowledge about strategy effectiveness outside the actual performance context. Confidence judgments, made after each recall attempt during the test, rate the likelihood that each recall response is correct. They measure item-level performance monitoring. Global differentiated postdictions, made at the end of testing an entire list, estimate the number of items that were correctly recalled under each study strategy. These judgments require inferential mechanisms that translate item-level performance monitoring into estimates of success rates across the entire list for each study strategy.

Deficits in Knowledge Updating

Dunlosky and Hertzog (2000) instructed people to use interactive imagery and rote repetition for learning unrelated paired associate (PA) items. Between-participant correlations between predictions (GPREDs and JOLs) and recall performance were higher on the second list than on the first list, indicating that a gain in strategy knowledge had occurred.

Other evidence, however, indicated that participants' learning about strategy effectiveness through task experience was biased and deficient. Postdictions after the List 1 test underestimated the actual level of recall with the imagery strategy by 20%. Moreover, this effect carried over to List 2 GPREDs, which showed more than 15% underconfidence for subsequent imagery recall. Astonishingly, JOLs for the two strategies were not reliably different during List 1 study and did not show a pronounced separation during List 2, with underconfidence for imagery items similar to the GPREDs. Thus, although the between-participant correlations suggested that some knowledge updating occurred, the underestimation of GPREDs for imagery recall and lack of updating in JOLs appeared to reflect a failure of participants to learn about the magnitude of effects these encoding strategies have on recall (for similar effects in a different encoding paradigm, see Bieman-Copland & Charness, 1994, and Matvey et al., 2002).

Hypotheses for Failures to Learn About Strategy Effects with Task Experience

In this paper, we consider two hypotheses for why people's GPREDs and JOLs demonstrate limited updating after task experience. One hypothesis is that students have difficulties inferring how many pairs were recalled for each strategy during the test on List 1. This *inferential deficit* hypothesis indicates that poor updating largely results from students' deficient cognitive processing when making metacognitive judgments. In contrast, a second hypothesis is that the experimental method used to measure updating makes students who do learn about the absolute effects of the strategies *appear* to show incomplete updating. We first briefly consider the latter possibility, and then return to the inferential deficit hypothesis.

Lack of updating as an experimental artifact. To assess knowledge updating, Dunlosky and Hertzog (2000) instructed participants to use imagery for some pairs and rote repetition for others to create supervised strategy experience. However, basing analyses on instructed strategy may systematically distort the accuracy of JOLs as measures of knowledge updating because individuals do not always comply with instructions to use the strategies. For instance, university students report complying with instructions to form interactive images for PA items between 60% and 80% of the time, depending on task conditions, item characteristics, and the nature of instructions (Dunlosky & Hertzog, 2001).

Typical procedures require evaluating knowledge updating as a function of *instructed* strategy, because no information on actual compliance is available. Thus, noncompliance with strategy instructions could create a distorted picture of knowledge updating. Successful (or unsuccessful) strategy use is a cue that influences magnitudes of JOLs (Hertzog, Dunlosky, Robinson, & Kidder, 2003). Thus, participants in Dunlosky and Hertzog (2000) might have appeared underconfident in predicting imagery recall because they gave relatively low JOLs to

items for which they had been instructed to use imagery, but could not do so and switched to rote repetition. Likewise, JOLs for instructed rote items may have been inflated because participants gave high JOLs to rote-instructed items they had actually studied with imagery. Presumably, students' GPREDs can take non-compliance into account – in the sense that they know their own strategic behavior -- but the criterion measure of recall, aggregated over instructed strategy, would not. Hence, students' JOLs and GPREDs may accurately reflect updating based on actual strategy use, but conflating instructed and actual strategy use could distort estimated knowledge updating.

Lack of updating due to an inferential deficit. Performance monitoring (as tapped by confidence judgments) during the List 1 test measures the potential basis for inferences about how many items were recalled with each strategy. Cued recall confidence judgments for unrelated PA items are highly accurate (Dunlosky & Hertzog, 2000; Higham, 2002). However, accurate performance monitoring is not sufficient for gaining accurate strategy knowledge. Individuals also need to attribute successful item recall to the strategy used, associate specific recall outcomes with prior strategy use, and aggregate this information across multiple items to produce declarative knowledge about the base rate of strategy effects. Hence, performance monitoring could be highly accurate, yet inferences about contributions of different strategies to this overall level of performance could be inaccurate.

Note that this particular claim implicates conscious awareness as critical for inferencing processes in knowledge updating. Knowledge updating (and corresponding inferences about strategy effects) have been measured either by having participants explicitly rate strategy effectiveness or by having them make metacognitive judgments that largely promote explicit assessments of strategy effects. Thus, although strategic behavior as procedural knowledge may

be obtained implicitly, our interest is in declarative knowledge about strategies as measured by explicit tests.

To measure the outcome of inferential processes about performance monitoring for different strategies, participants made a global differentiated postdiction for each strategy. The inferential deficit hypothesis suggests that the limited updating measured by GPREDs and JOLs on List 2 results from deficiencies in these inferential processes, which would already be evident in inaccurate Trial 1 postdictions.

Overview of the Experiments

We evaluated the aforementioned hypotheses in two experiments. In Experiment 1, students made item-by-item strategy reports so that updating could be assessed by instructed strategy and by reported strategy use. If limited updating in predictions is an artifact of conflating instructed strategy with actual strategy use, then the List 2 GPREDs and JOLs should show (a) limited updating when analyzed by instructed strategy use but (b) more complete updating when analyzed by reported strategy use. Postdictions were also collected to test the inferential deficit hypothesis. To foreshadow, results from Experiment 1 ruled out the hypothesis of instructed strategy artifact and confirmed the inferential deficit hypothesis. Thus, we explored other implications of this hypothesis in Experiment 2. In particular, if difficulty with inferences about test outcomes on List 1 limits the degree of updating, making the inferences easier— for example, by restructuring the List 1 test —may yield more complete updating and de-bias List 2 predictions.

Finally, in both experiments, we included an encoding questionnaire, both before and after task experience, in which participants rated the effectiveness of different associative

learning strategies, including interactive imagery and rote repetition. These data allowed us to assess declarative strategy knowledge updating that might not be reflected in JOLs and GPREDs.

Experiment 1

Method

Design

The design was a 2 X 2 X 3 mixed factorial, with Trial (first vs. second study list), Strategy instructions (rote repetition vs. imagery), and Judgment group (no global undifferentiated postdiction [control], undifferentiated postdiction with no performance feedback, and undifferentiated postdiction, followed by performance feedback).² Trial and Strategy instructions were manipulated within-subjects while the postdiction was treated as randomly assigned between-subjects factor.

Participants

Two hundred fifty-three undergraduate students at the Georgia Institute of Technology participated for course credit. Random assignment resulted in 85 participants in the control group, 82 in the undifferentiated postdiction group, and 86 participants in the undifferentiated postdiction with feedback group.

Materials

The Personal Encoding Preferences (PEP) questionnaire was used to assess participants' subjective ratings of the effectiveness of the interactive and rote repetition strategies (see Hertzog & Dunlosky, 2004 for the full set of PEP items). The PEP presents a list of memory strategy definitions and examples of how they would be used. Participants are asked to make a rating of the effectiveness of each of the strategies on a Likert-type scale (1 = least effective; 10

= most effective) printed beneath each of the strategies. The full PEP was administered, but we report ratings only for the interactive imagery and rote repetition items.

The PEP I assessed pre-experimental strategy effectiveness beliefs. The PEP II assessed participants' strategy knowledge after the experimental task by directing them to imagine themselves advising a friend about the effectiveness of each of the strategies. Otherwise, the two versions were identical. The different instructions were created to avoid demand characteristics associated with administering the identical questionnaire twice in a single session.

One hundred and six word pairs consisting of relatively frequent, concrete nouns were constructed for this study. All items consisted of associatively unrelated nouns, validated by checking against the University of South Florida norms (Nelson, McEvoy, & Schreiber, 1999). The instructions along with the experimental task were programmed in HyperCard on Macintosh computers. All responses were entered and recorded on the computer keyboards.

Procedure

Each of the experimental groups received two lists of 50 PA items, which participants studied and on which they were tested. The list of 100 PA items was randomly divided into two halves for each participant. Each half of the list (i.e., each 50-item list) was used for a trial of the experiment. For each list, half of the items were randomly assigned to one of the two strategies (rote repetition or interactive imagery). Throughout each of the study-test trials participants made all of their metacognitive judgments at their own pace. The PEP I was given prior to the study phase for List 1; the PEP II was given immediately after testing of List 2 ended.

Instructions. Before beginning the experiment, participants were given detailed instructions regarding the experimental task, including descriptions of the definitions of rote repetition and interactive imagery. These descriptions included specific examples of how these

strategies would be used. No normative information was provided regarding the differential effectiveness of the two strategies. Participants were also given detailed instructions regarding the metacognitive judgments that they would be making. They were next informed about the strategy report screen. Participants were encouraged to use the instructed strategy, but with the acknowledgment that one cannot always comply, were told to report the strategy (if any) they had actually used. After instructions, a practice study phase with 6 items ensued in order to familiarize them with the PA task and with making JOLs. No recall practice was provided.

Global Differentiated Predictions. After instructions and practice, participants were prompted to make GPREDs of their performance with each strategy prior to beginning the study phase of the experiment. They were told that they would soon be studying a list of 50 items, half of which would be studied with interactive imagery and half with rote repetition. They made two GPREDs, one for rote repetition and one for interactive imagery, estimating the percentage of items that they would correctly recall for each of the strategies. The computer program required participants to enter a prediction between 0 and 100%. The order of making these predictions (i.e., repetition first or imagery first) was counterbalanced across participants. Due to a programming error, the instruction screen for the rote repetition pre-study predictions in the undifferentiated postdiction with feedback group erroneously labeled the predictions as imagery. Hence, for 86 of the participants, the first rote repetition prediction was treated as missing data.

Study. After making their global predictions, participants pressed the return key to begin studying items. First, a screen appeared that notified the participant of the strategy with which they should study the subsequent word pair (“Imagery” or “Repetition”). This screen remained visible for approximately 2 s. Immediately thereafter, the to-be-studied word pair appeared in the center of the screen for 6 s. To avoid any possible memory problems, the instructed strategy was

displayed at the top of the screen. The order of presenting the 50 PA items and the order of instructing strategies was randomized for each participant.

Immediately after PA item presentation, the JOL screen appeared. Participants were asked to enter any number between 0 and 100 that indicated their confidence in recalling the target word of the just-presented word pair when given the cue word in about 10 minutes. A strategy report followed the JOL. Participants reported their strategy by entering a number corresponding to: (1) repetition, (2) imagery, (3) a combination of both imagery and repetition, (4) a strategy other than imagery or repetition, (5) no strategy, or (6) did not have time to use a strategy. In the current research and in our previous work (e.g., Dunlosky & Hertzog, 1998; 2001), the “other” category is rarely endorsed.

Test. At test, participants were presented with the cue words from the PA items. They were instructed to enter the target word. Responses were scored as correct if the first three letters of the typed response matched the first three letters of the correct target word. (None of the targets shared the same first three letters.) The consistency between hand scoring and computer scoring is nearly perfect, so the latter was preferred for ease of data processing. Following their target response participants were prompted to make a confidence judgment. Participants chose any number between 0 and 100 that reflected their percentage confidence that their response was correct.

Global postdictions. After the recall test, individuals made the undifferentiated postdiction, estimating the total percentage of PA items from the entire list they had just recalled. Next, participants made differentiated postdictions. Participants were asked to estimate the percentage of PA items they had correctly recalled for each of the two strategies, imagery and

rote repetition. For both types of postdictions, individuals were required to enter a number between 0 and 100%.

List 2. Following the global differentiated postdictions made after List 1, List 2 began immediately following the same sequence outlined above (but without practice). List 2 used a different set of 50 PA items, but otherwise the procedure was identical.

Results

Recall Performance

As reported in Table 1, recall with reported imagery use was reliably greater than recall with reported rote repetition use, $F(1, 249) = 981.82, p < .001$, partial $\eta^2 = .798$. Results were similar with instructed strategy use as the dependent variable, but produced a reliable Strategy X Trial interaction for instructed strategy, $F(1, 250) = 7.42, p < .01$, partial $\eta^2 = .029$, that was not reliable for reported strategy, $F < 1$. Most important, regardless of whether analyses were analyzed by instructed or reported strategy use, strategy had a major impact on recall.

Compliance with Strategy Instructions

Reported compliance with strategy instructions was good but far from perfect. Reported compliance with imagery instructions on List 1 and List 2 was 83% ($SD = 15.2$) and 84% ($SD = 17.4$), respectively; for rote repetition, the corresponding values were 67% ($SD = 24.4$) and 63% ($SD = 26.8$). Compliance was reliably higher for imagery, $F(1, 250) = 160.37, p < .001$, partial $\eta^2 = .391$. Examination of the percentage of time each strategy was reported used on each list indicated individuals shifted away from rote repetition use (.38 in List 1 versus .36 in List 2) to greater imagery use on List 2 (.55 versus .57), without any corresponding change in the other strategy report categories (i.e., other, no strategy, ran out of time to use a strategy), which collectively represented a static .07 percent of the strategy reports on both List 1 and List 2. This

shift away from rote repetition to imagery across lists produced a reliable Strategy X Trial interaction for the compliance variable, $F(1, 250) = 13.37, p < .001$, partial $\eta^2 = .051$, with individuals showing greater tendency to comply with imagery instructions, but lesser compliance with repetition instructions across lists. The lack of perfect compliance opened the possibility that different conclusions could arise about knowledge updating when analyzing prediction accuracy by reported versus instructed strategy use.

Prediction Accuracy

Global differentiated predictions. Table 2 reports the absolute accuracy of participants' GPREDs, computed as GPRED minus recall, as a function of reported and instructed strategies (for the constituent values that comprise these difference scores, see Table 1).³ Because the two types of strategy classification are based on rearrangements of the same recall data, it was not possible to conduct direct statistical tests of variables with reported vs. instructed strategies as a within-subjects factor (one cannot assume local independence of errors). Instead, here and in the rest of the paper, we compare results from separate analyses of both dependent measures, either aggregated on the basis of instructed strategy or on the basis of reported strategy.

When absolute accuracy of individuals' imagery and rote repetition predictions was examined as a function of strategy use, the results varied slightly depending on which strategy classification was used. Mean accuracy of imagery and rote repetition predictions tended to be better when GPRED absolute accuracy was based on instructed strategies rather than reported strategies. The effect size associated with significant Strategy X Trial interactions was lower for reported strategy, $F(1, 164) = 19.03, p < .01$, partial $\eta^2 = .104$, relative to instructed strategy $F(1, 165) = 39.88, p < .01$, partial $\eta^2 = .195$, respectively, indicating greater discrepancy in absolute accuracy for imagery versus rote across lists for analyses organized by instructed strategy. Most

relevant, as evident from inspection of Table 2, the absolute accuracy of predictions for imagery items actually worsened over lists for both reported and instructed strategies, $F(1, 250) = 3.16, p < .001$, partial $\eta^2 = .050$ and $F(1, 250) = 9.12, p < .01$, partial $\eta^2 = .035$, respectively, indicating that underconfidence for imagery items is not due to conflating instructed with actual strategy use. By contrast, the accuracy of predictions for rote items was almost perfect for List 2 when sorted by reported strategy but less accurate when analyzed by instructed strategy, $t(251) = 11.7, p < .01$.

JOLs. Table 2 also reports the absolute accuracy of JOLs, computed as mean JOL minus recall. For reported strategy use, a reliable Strategy X Trial interaction $F(1, 249) = 9.53, p < .01$, partial $\eta^2 = .037$, reflected participants exhibiting underconfidence for imagery items on both Lists 1 and 2, but overconfidence on List 1 and underconfidence on List 2 for rote items. A similar pattern of results was observed when the absolute accuracy of JOLs was examined as a function of instructed strategy type, $F(1, 250) = 34.47, p < .01$, partial $\eta^2 = .121$. Moreover, imagery JOLs actually became *less* accurate with task experience for both reported (List 1 mean = -16.67, $SE = 1.66$ vs. List 2 = -22.95, $SE = 1.47$), $F(1, 250) = 25.18, p < .01$, partial $\eta^2 = .09$, and instructed strategies (List 1 = -16.96, $SE = 1.68$ vs. List 2 = -21.65, $SE = 1.43$), $F(1, 250) = 13.46, p < .01$, partial $\eta^2 = .051$.

Performance Monitoring

Although analyzing by reported strategy use uncovered updating in one circumstance (GPREDs for rote items), GPREDs and JOLs for imagery items were still highly underconfident. This underconfidence may be localized to either faulty performance monitoring (as demonstrated by poor confidence judgment accuracy) and/or to faulty inferential processes (as demonstrated by poor postdiction accuracy).

Confidence judgments. Absolute confidence judgment accuracy, computed as mean confidence judgment minus recall, was nearly perfect (Table 3). For reported strategies, the mean difference between confidence judgments and recall was less than 3%. Only the main effect of Trial was reliable, $F(1, 249) = 8.27, p < .01$, partial $\eta^2 = .032$, with absolute accuracy improving from List 1 ($M = 2.76, SE = 0.43$) to List 2 ($M = 1.59, SE = 0.44$). Participants were slightly overconfident for both lists, but confidence judgment accuracy was excellent.

Global differentiated postdictions. Table 3 also reports absolute accuracy for the postdictions, computed as postdiction minus recall. A robust main effect of Strategy, $F(1, 249) = 136.64, p < .001$, partial $\eta^2 = .354$, showed that absolute accuracy was far worse for reported imagery items than for reported rote items (marginal $M = -15.2, SE = 1.0$, versus marginal $M = -1.7, SE = 1.0$). Absolute accuracy was reliably worse for List 2, $F(1, 249) = 20.99, p < .001$, partial $\eta^2 = .078$, and the loss of accuracy was greater for imagery items, reflected in the reliable Strategy X Trial interaction $F(1, 249) = 5.58, p < .05$, partial $\eta^2 = .022$. Using instructed strategy to organize the data reduced the interaction to nonsignificance, $F < 1$, with the difference in absolute accuracy being only 2.9 for instructed strategy versus 13.5 for reported strategy. Perhaps most notable was the underconfidence in List 1 postdictions for imagery items immediately after first experience with test outcomes, and following highly accurate confidence judgments.

Strategy Effectiveness Ratings

The PEP questionnaire demonstrated updated knowledge of differential strategy effectiveness, reflected in a robust Strategy X Trial interaction, $F(1, 250) = 286.64, p < .001$, partial $\eta^2 = .534$. As shown in Figure 1, imagery was rated as superior to rote repetition before the task, $t(252) = 3.47, p < .001$, but this difference was far greater after task experience, $t(252) =$

23.70, $p < .001$. Ratings of rote repetition effectiveness decreased after experience, whereas ratings of imagery effectiveness increased. Hence, after using both strategies individuals did gain declarative knowledge about the superiority of interactive imagery over rote repetition.

Discussion

The pattern of results from Experiment 1 supports a number of conclusions. First, although using reported strategy to analyze results improved the absolute accuracy of GPREDs for rote repetition items for List 2, the general patterns of limited knowledge updating observed by Dunlosky and Hertzog (2000) were replicated irrespective of whether analyses were sorted by instructed or reported strategy. Thus, limited knowledge updating cannot be explained by the experimental artifact of conflating instructed and actual strategy use. Second, confidence judgment accuracy was exceptional, whereas postdictions for imagery items made after List 1 recall showed substantial underconfidence. Like others, we found that cued recall performance monitoring is highly accurate (e.g., Higham, 2002). Although both judgments require post-recall evaluations of performance, they differ in a critical respect. When making differentiated postdictions, individuals must access strategy information and recall outcomes for multiple items, whereas confidence judgments involve judging the likelihood that a single recall response is correct. Hence confidence judgments do not necessarily require either (a) a strategy assessment or (b) aggregating information across multiple items. Therefore, these outcomes are consistent with faulty inferential processes involved in estimating the number of imagery items recalled after testing has been completed.

Experiment 2

Experiment 2 was conducted to evaluate whether absolute accuracy of the GPREDs, differentiated postdictions, and JOLs could be de-biased. We focused on the hypothesis that

individuals do not access accurate information about recall-strategy associations after test (made available by accurate confidence judgments), leading to inaccurate inferences about the magnitude of the strategy effects. This deficiency could result from two non-exclusive mechanisms: (a) from source misattributions or resource limitations when constructing global postdictions, and/or (b) because participants do not actively attend at test to how well they are performing with each strategy, perhaps because doing so was not an explicitly stated task goal. The first account places the locus of poor inferencing with deficiencies in the cognitive system, whereas the second indicates that participants are able to make accurate inferences but perform the task in a manner that prevents it.

Deficiency Accounts

To investigate knowledge updating in Experiment 1 and elsewhere (e.g., Bieman-Copland & Charness, 1994; Dunlosky & Hertzog, 2000), participants study a long list of items in a single study session. At test they need to associate memory for original encoding strategy with the recall outcomes to obtain well-calibrated information about the number of responses recalled with each strategy. When making postdictions after the test they must access this information in a manner that preserves these differences, and accurately infer (construct) the strategy-differentiated performance estimate.

This process resembles making source judgments, in which individuals must not only recall studied information but also the source where they obtained it. Individuals are prone to make misattributions about source in such experiments unless they encode relevant, distinctive information during study and explicitly attempt to retrieve that information at test (e.g., Johnson & Raye, 2000). Individuals may not retrieve information about which strategy had been used at the time of recall. Another limitation on PA knowledge updating could be that tracking the

strategy-outcome associations exceeds the processing capacity of individuals who are focused primarily on retrieving targets (e.g., Bieman-Copland & Charness, 1994). Students may have difficulties holding the number of recall successes for both strategies in memory while concurrently recalling targets, monitoring recall success and failures, and updating strategy success after each recall attempt. If so, accurate information about overall strategy success (e.g., the number of items correctly recalled per strategy) will not be available when participants later make postdictions and GPREDs. Finally, these associations may be accurately formed at test and available at the time of the judgment, but they may not be accessed in a way that affords accurate global postdictions. For example, explicitly searching for source-differentiating information is needed to avoid source misattributions (e.g., Johnson & Raye, 2000).

The first two influences place the locus of the effect at deficient association of strategy with recall outcome at the time of item testing. If individuals have difficulties accurately associating recall outcomes with the strategies that generated them, then structuring the PA learning experience to afford easier categorization of outcomes by encoding strategy should sidestep both deficiencies and hence improve the absolute accuracy of both List 1 postdictions and List 2 GPREDs. Thus, we used the strategy reports of participants to create blocks of PA recall that were homogeneous with respect to reported strategy (imagery or rote repetition). At the start of each block, participants were told the strategy they reported using for those items. This procedure created concentrated recall experience with a strategy while removing the requirement to recall the original strategy and associate it with a recall outcome. We hypothesized that blocked PA testing would enhance knowledge updating about the absolute effects of the strategies and hence lead to improved judgment accuracy.

Motivated Goal Attainment Account

In previous research on knowledge updating, individuals are not told that their goal is to learn about strategy effectiveness. Individuals often construct explicit performance goals to motivate themselves to reach a desired outcome (Gollwitzer, 1999). In our task, forming the explicit goal of learning about differences in strategy effectiveness may cause individuals to monitor performance with the two strategies more analytically. For instance, having an explicit goal of learning about strategies could make individuals more likely to explicitly form strategy-recall outcome associations during test. Goal attainment is optimized when individuals plan actions in advance to achieve their goals. One way to plan is to form implementation intentions. According to Gollwitzer (1999), implementation intentions are subordinate to goal intentions and specify when, where, and how responses will lead to goal attainment. By doing so, a person commits to performing specific behaviors when specific situations arise. Experimental induction of implementation intentions improves prospective memory performance (Chasteen, Park, & Schwarz, 2001). Hence it may improve knowledge updating as well.

We motivated participants to learn about strategy effectiveness by giving them goal and implementation intention instructions before beginning the task. If poor updating is related to a lack of specific goals to learn about strategy effectiveness, then goal instructions will enhance the absolute accuracy of List 1 differentiated postdictions, leading to more accurate List 2 GPREDs and JOLs.

Method

Design

The design of the experiment was a 3 X 2 X 2 X 2 mixed factorial, with Strategy instruction (interactive imagery versus rote repetition) and Trial (first vs. second study list) as

within-subjects variables and Goal instructions and Test (blocked vs. random testing) as between-subjects variables. The three levels of goal instructions were control (no additional instructions), goal intention, and goal plus implementation intention.

Participants

The sample consisted of 126 undergraduate students at the Georgia Institute of Technology who were given extra credit for their participation. Students were randomly assigned to the six between-subject design cells (i.e., control-blocked, control-random, goal intention-blocked, goal intention-random, implementation intention-blocked, and implementation intention-random), resulting in groups of 24, 20, 20, 20, 21, and 21 persons, respectively.

Materials

The same demographic information was collected from participants as in Experiment 1. The two PEP questionnaires again assessed participants' beliefs about the effectiveness of different strategies both before and after gaining task experience.

The 124 word pairs again consisted of relatively frequent, concrete nouns selected from the University of South Florida norms to have no prior association (Nelson et al., 1999). Four of these word pairs were used during practice while the other 120 pairs were randomly divided into two lists of 60 word pairs. The experimental task was programmed in Visual Basic programming language and run on PC desktop computers.

Procedure

Participants in the control group received the same instructions as Experiment 1. In the goal intention group, participants were instructed both before and after the practice trials that the goal of the experiment was to learn which strategy works best for them. They were told:

The goal of this experiment is for you to learn which strategy, rote repetition or interactive imagery, works best for you. Specifically, the goal is for you to

accurately estimate the probability that you will remember a word after having studied it with interactive imagery or after studying it with rote repetition.

Following the presentation of the goal intention instructions, participants were asked to answer the written question: “What is the goal of this experiment?” If they answered correctly they were allowed to continue with the computer task. An answer was considered correct if it included in any form of the statement, “The goal of this experiment is for me to learn which strategy, rote repetition or interactive imagery, works best for me.” If participants did not answer correctly, they were asked to be more specific. If still unable to answer correctly, they were presented with a copy of the goal intention statement, asked to reread it, and asked to answer again. Only articulating a correct goal summarization enabled them to proceed.

The implementation intention group was presented with the same instructions mentioned above for the goal intention group. In addition, the following set of instructions was used to establish an implementation intention:

To help you achieve this goal, when you are being asked to recall the second word from each word pair, you should: Think about which strategy (Rote Repetition or Imagery) you used to study each word pair, and keep count of how many word pairs you are correctly recalling for each strategy.

Following the presentation of the implementation intention instructions after the practice trials had been presented, participants were asked by the experimenter, “To help obtain this goal, when you are studying a word pair, what should you do?” The same query procedure was used for this group; participants only proceeded when they could correctly repeat the instructions.

The Test factor had two levels: blocked and random. In both cases, participants were tested on 40 of the 60 PA items studied. Testing was conducted on a reduced set of items to maximize the probability of filling the test-set with 20 items actually studied with each strategy. Reported strategy use was used to select items for testing. The 40 items tested were targeted to

include 20 items the participant reported studying with interactive imagery and 20 items the participant reported studying with rote repetition. If a participant used a strategy more than 20 times, 20 items of that set were randomly selected for testing.

The blocked testing group consisted of four blocks of 10 PA items, 2 blocks for each strategy. At the start of each block, participants were informed which strategy they had reported using to encode the items. To ensure that blocked testing for List 1 items was presented with accurate strategy information, only participants that reported using imagery and rote repetition at least 20 times each were included in the analysis. Twenty-one individuals were excluded from the analyses for not complying sufficiently with instructed strategies in List 1, leaving 105 participants in the analyses. In the random testing group, the 40 items were tested in a random order. The random group also imposed brief delays between blocks of 10 randomly selected items to match the procedure in the blocked group.

Predictions and postdictions. Procedures were identical to Experiment 1. Participants were told that the experiment investigated people's learning and their assessment of their own learning. They were also informed that after studying 60 PA items they would be tested on 40 items when cued by the first word. GPREDs, JOLs, and postdictions were collected; given that confidence judgments were highly accurate in Experiment 1 and other studies, they were dropped from this experiment.

Results

Surprisingly, goal instructions had minimal impact on all of the dependent measures, and most important, it did not reliably interact with strategy and trials. Thus, we collapsed across goal instructions for the subsequent analyses.

Recall Performance

Table 4 reports recall data for both random and blocked testing. As expected, recall was greater after interactive imagery ($M = .67, SE = .02$) than after rote repetition ($M = .26, SE = .02$), $F(1, 99) = 428.25, p < .001$, partial $\eta^2 = .812$. This difference interacted with Test, $F(1, 99) = 5.06, p < .05$, partial $\eta^2 = .049$. Imagery recall was reliably higher in the blocked group ($M = .71, SE = .03$, versus $M = .64, SE = .03$), whereas repetition recall was similar in the two groups.

Global-differentiated Predictions

Table 4 shows that blocking at test resulted in a greater separation of List 2 GPREDs for the imagery versus repetition strategies, which indicated enhanced gains in knowledge about the relative effects of the strategies. Table 5 shows that these changes in GPREDs were also reflected in their absolute accuracy (computed as GPRED minus recall).

Table 5 reports the absolute accuracy of GPREDs, which decreased across the two lists, $F(1, 99) = 71.44, p < .001$, partial $\eta^2 = .419$. List 1 predictions had a mean absolute accuracy of 7.0 ($SE = 2.2$) compared to a mean absolute accuracy of -11.4 ($SE = 1.6$) for List 2. The absolute accuracy for interactive imagery predictions ($M = -13.7, SE = 2.0$) was reliably different than for rote repetition predictions ($M = 9.4, SE = 1.7$), $F(1, 99) = 152.92, p < .001$, partial $\eta^2 = .607$.

Most important, the three way interaction between Trial, Strategy, and Test was significant, $F(1, 99) = 6.58, p < .05$, partial $\eta^2 = .062$. To follow up this reliable three-way interaction, a 2 (Trial) X 2 (random vs. blocked) ANOVA was conducted separately for imagery items and rote repetition items. A main effect occurred for Trial for both rote repetition items, $F(1,103) = 120.70, p < .001$, partial $\eta^2 = .540$, and for imagery items $F(1,103) = 7.33, p < .01$, partial $\eta^2 = .066$. Absolute accuracy improved for rote repetition items and declined for imagery items across lists. However, the Trial by Test interaction was only reliable for rote repetition items, $F(1,103) = 5.59, p < .05$, partial $\eta^2 = .051$, but not for imagery items, $F < 1$. The random

group improved their accuracy for rote repetition items across lists to a greater extent than the blocked group, but both groups showed similar declines for imagery items across lists.

Judgments of Learning

JOL magnitudes are presented in Table 4. Absolute JOL accuracy, calculated as mean JOL minus recall for each strategy, is reported in Table 5. Absolute accuracy decreased across Trials, $F(1, 99) = 40.08, p < .001$, partial $\eta^2 = .288$, showing a greater underestimation of performance on List 2 (marginal $M = -16.6, SE = 1.6$) than on List 1 (marginal $M = -7.3, SE = 2.0$). Absolute accuracy was worse for items studied under interactive imagery than items studied using rote repetition, $F(1, 99) = 198.55, p < .001$, partial $\eta^2 = .667$. This difference increased at List 2, resulting in a reliable interaction, $F(1, 99) = 8.62, p < .01$, partial $\eta^2 = .080$. For List 1, mean JOLs underestimated imagery strategy recall ($M = -22.0, SE = 2.5$), but overestimated rote repetition recall ($M = 7.4, SE = 2.1$). At List 2, the underestimation of imagery increased (marginal $M = -28.3, SE = 2.1$) and rote repetition recall was also underestimated (marginal $M = -4.9, SE = 1.6$). Most critical, blocking at test did not influence absolute JOL accuracy, $F(1, 99) = 1.07, p > .05$, partial $\eta^2 = .011$, even though it did have some effect on JOL magnitudes.

Global Differentiated Postdictions

Postdiction magnitudes are provided in Table 4. Table 6 reports the absolute accuracy of these differentiated postdictions, computed as postdiction minus recall for each strategy. As can be seen in Table 6, the random testing group replicated findings from Experiment 1, with substantial underestimation of imagery recall. Blocked testing improved postdiction accuracy on average, $F(1, 99) = 9.99, p < .01$, partial $\eta^2 = .092$, but this was complicated by a reliable Strategy X Test interaction, $F(1, 99) = 31.41, p < .001$, partial $\eta^2 = .241$. Blocked testing substantially reduced the underestimation of imagery recall, but did not improve estimation of

repetition recall. There was also a reliable Test X Trial interaction, $F(1, 99) = 5.66, p < .05$, partial $\eta^2 = .054$, with accuracy degrading from List 1 to List 2 in the control condition but not in the blocked condition. No other effects were reliable.

Strategy Effectiveness Ratings

As in Experiment 1, the effectiveness ratings provided definitive evidence of knowledge updating in a robust Strategy X Trial interaction, $F(1, 99) = 204.63, p < .001$, partial $\eta^2 = .674$ (Figure 2). The trend for the blocked testing manipulation to affect participants' ratings was not reliable, $F(1, 99) = 3.40, p > .05$, partial $\eta^2 = .033$. Separation of the strategy ratings after task experience was 2.1 *SDs* in the random condition, 3.0 *SDs* in the blocked condition. Participants learned that imagery was the superior strategy regardless of how they had been tested.

Discussion

Organizing recall tests into homogeneous blocks of reported strategy use (either imagery or rote repetition) and informing participants which strategy was being tested repaired much of the poor absolute accuracy of postdictions for imagery-studied items. The beneficial effect of blocking supports the hypothesis that limited resources (and/or poor strategy source monitoring) were partially responsible for participants' problems in postdicting imagery recall. However, such benefits were not accompanied by improved absolute accuracy for the List 2 imagery GPREDs. We consider reasons for such inaccuracy in the General Discussion.

Instructing participants to set a goal of learning about strategy effectiveness had a minimal impact on judgment accuracy and knowledge updating, which suggests that an explicit intention to learn about strategy effectiveness cannot easily overcome the resource constraints imposed by the randomized testing order. It also suggests, indirectly, that either individuals spontaneously seek to learn about differential strategy effectiveness, or that awareness and intent

are unimportant for accurate performance monitoring or for inferences about strategy effectiveness. Thus, this overall pattern of outcomes supports the inferential deficiency account (vs. goal orientation account) for why individuals show limited updating after task experience.

General Discussion

These experiments showed that individuals learn about the differential effectiveness of rote repetition and imagery with task experience. This learning is manifested in substantial changes in imagery and rote strategy effectiveness ratings (Figures 1 and 2). However, the magnitudes of metacognitive judgments, and their absolute accuracy, did not always reflect this knowledge updating. Although students learn that imagery is far more effective than rote repetition, this knowledge is not utilized when making judgments in a manner that achieves better absolute accuracy for these measures. Blocked testing ameliorates this tendency, especially for imagery postdictions, but does not eliminate the inaccuracies. In the remainder of this discussion, we briefly review the implications of these outcomes for understanding both successes and failures in knowledge updating with task experience.

Theoretical Implications

A motivating factor for this research was the fact that the absolute accuracy of imagery postdictions in this task is poor when items are tested in a random testing order. Thus, performance evaluation processes are not routinely accurate when differentiated outcomes must be associated at test with the encoding strategies that generated them. The present experiments indicate that one source of the poor accuracy is in the construction of List 1 postdictions, an inaccuracy that carries forward to the second study/test trial. Confidence judgments for List 1 item recall were well-calibrated and highly diagnostic of the accuracy of candidate item retrievals (see also Higham, 2002). Apparently, participants know when they are retrieving

correct versus incorrect responses during cued recall, at least for the unrelated PA items used in this study. The high accuracy of item-level performance monitoring contrasts the substantial underestimation of imagery recall by imagery postdictions. Thus, accurate item-level performance monitoring may be necessary for producing accurate postdictions, but it is certainly not sufficient.

Postdictions' underestimation of imagery recall was substantially repaired by blocking at testing. How should this effect be interpreted? The present experiments are not definitive as to the exact nature of this limitation. We favor the hypothesis that a resource limitation in accessing and compiling information about original strategy use, either at the time of recall or when the postdictions were made, limited the accuracy of the imagery postdiction.

First, consider the argument that individuals do not form the strategy-recall outcome association at the time each item is tested. Individuals may not retrieve the strategy initially used to form the new association between words at the time of the cued recall test, or may do so in a way that does not form an association between strategy and outcome. The fact that explicit goals to learn the associations by attending to and counting the instances of these conjunctions had no effect on absolute accuracy argues against this hypothesis. Furthermore, we conducted another experiment (not reported here) in which participants were explicitly told during List 1 recall which strategy had been instructed for each item as it was being tested (in a random testing order). This reminder had no effect on List 1 imagery postdiction underconfidence.

Thus, we currently favor the hypothesis that compiling strategy information about exactly how many items per strategy were recalled presents the main obstacle for accurate knowledge updating, and that this effect is constrained by the resource demands of engaging in cued retrieval search. The blocking effect detected in this study suggests that students may be able to

form more explicit representations of frequency of recall success with each strategy when given small homogeneous blocks of strategy trials. Although we argue that this effect is probably caused by a resource limitation during test, it could also be influenced by a general tendency to avoid tracking the base rate of success with each strategy, deferring this process until the time that the postdiction is requested by the experimenter.⁴ Arguing against this view is the fact that implementation intentions to actively track rates of success and failure had no effect, either when paired with random or blocked testing. We acknowledge, however, that further experiments are needed to test the limited resource hypothesis. For example, showing that the benefits of blocked testing for postdictions' absolute accuracy are eliminated by a secondary, attention demanding task during test would provide further evidence for the limited resource account.

To a degree, improvements in the accuracy of postdictions with blocking at test also carried over to the predictions made on List 2. In the blocked group only, the magnitude of GPREDs showed greater sensitivity to the strategies after task experience on List 2 than on List 1 (Table 4). Again, a major conclusion here, supported by the questionnaire data, is that people evidently were gaining knowledge about the *relative* effects of the strategies—i.e., that imagery is superior to rote repetition. Nevertheless, participants' List 2 GPREDs and JOLs still substantially underestimated imagery recall. Moreover, poor absolute accuracy was especially apparent in JOLs, for which blocking had no reliable effect on absolute accuracy. JOLs may be entrained more by on-line processing at encoding than by information about strategy effects (Hertzog et al., 2003; Koriat, 1997). The absolute accuracy of JOLs can be de-biased in other experimental contexts (e.g., Koriat & Bjork, 2006), so the failure to do so in this experiment is particularly telling.

Another potential explanation for the degree of underestimation of imagery recall is that the List 2 JOLs and GPREDs are partly anchored (e.g., Scheck, Meteer, & Nelson, 2004) by the *overall* poor recall performance on List 1 that individuals can report quite accurately, even without being informed they will be asked to do so (Devolder, Brigham, & Pressley, 1990). Given lack of access at the time of the postdictions to explicit information about base rate of strategy success, individuals probably do not attempt to retrieve all instances of success and failure with different strategies. Nor are they likely to be able to do so! Instead, individuals can estimate overall recall performance accurately, and then heuristically adjust imagery estimates up and rote estimates down from the anchor of overall postdicted performance. Participants adjust the List 2 predictions from this anchor (more after blocking), which reflects their new knowledge that imagery is superior to rote repetition, but this heuristic adjustment is not perfectly calibrated.⁵ The fact that blocked testing de-biases imagery postdictions without improving rote repetition postdiction accuracy is consistent with the operation of some kind of anchoring mechanism, combined with acquired strategy knowledge -- as opposed to a general improvement in estimation accuracy that should benefit both types of differentiated postdictions. We favor the anchoring hypothesis over a general reluctance to use extreme ratings because confidence judgments were highly accurate, necessitating high-confidence responses for successful cued recall.

Nevertheless, the present results serve as a reminder that one cannot assume that metacognitive judgments, like JOLs, will automatically and routinely reflect knowledge updating when it occurs. Past research using JOLs to infer that individuals do not learn about differential strategy effectiveness (e.g., Shaughnessy, 1981) need to be reinterpreted in light of our findings.

New Directions for Research on Knowledge Updating

Several questions arise that should be entertained in future research about knowledge updating. First, how durable is such newly gained knowledge when it is specific to a novel task such as PA learning? Second, will strategy knowledge gained with task experience transfer to ad lib strategy use when strategy use is unsupervised? Perhaps students who just learned that imagery is superior will spontaneously adopt it (over rote repetition) when learning a new list of PA items in the same context. Such a finding would reinforce the idea that knowledge reflected in the PEP questionnaires is sufficient to motivate greater imagery strategy use. Third, under what circumstances would strategy learning arise adventitiously, without conscious awareness, in tasks where metacognitive judgments were not inserted that focus explicit awareness on strategies and their benefits (e.g., Crowley et al., 1997)? Fourth, would an imagery preference gained in the PA task generalize to other contexts? Students must choose strategies for study. After structured experience with imagery superiority in the PA task, are they more likely to adopt imagery? To the extent that the knowledge gained is represented as specific to the task context, it may not generalize to new learning situations, either in or out of the laboratory.

Certainly, one should not conclude from our findings that people do not learn about actual magnitudes of strategic benefits. The PA task used here has the advantage of minimal pre-experimental exposure, a benefit for assessing new strategy knowledge. Yet it is likely that the kind of intensive, massed strategy experience in this task does not generalize to many situations in which people learn about new strategies, which can occur gradually over a series of repeated exposures and be accompanied by performance feedback (e.g., Crowley et al., 1997; Lovett & Schunn, 1999).

In sum, we have shown that the accuracy of inferences about strategy effectiveness can be improved through structured testing, even as we have demonstrated that such improvements

do not produce differences in rated strategy effectiveness outside the experimental task. These results challenge the validity of metacognitive judgments as measures of knowledge updating. Individuals learn about differential strategy effectiveness even as their metacognitive judgments fail to accurately reflect their newly acquired knowledge.

References

- Bieman-Copland, S., & Charness, N. (1994). Memory knowledge and memory monitoring in adulthood. *Psychology and Aging, 9*, 287-302.
- Brigham, M. C., & Pressley, M. (1988). Cognitive monitoring and strategy choice in younger and older adults. *Psychology and Aging, 3*, 249-257.
- Chasteen, A. L., Park, D. C., & Schwarz, N. (2001). Implementation intentions and facilitation of prospective memory. *Psychological Science, 12*, 457-461.
- Crowley, K. U., Shrager, J., & Siegler, R. S. (1997). Strategy discovery as a competitive negotiation between metacognitive and associative mechanisms. *Developmental Review, 17*, 462-489.
- Devolder, P. A., Brigham, M. C., & Pressley, M. (1990). Memory performance awareness in younger and older adults. *Psychology and Aging, 5*, 291-303.
- Dunlosky, J., & Hertzog, C. (1998). Aging and deficits in associative memory: What is the role of strategy production? *Psychology and Aging, 13*, 597-607.
- Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging, 15*, 462-474.
- Dunlosky, J., & Hertzog, C. (2001). Measuring strategy production during associative learning: The relative utility of concurrent versus retrospective reports. *Memory & Cognition, 29*, 247-253.
- Dunlosky, J., Hertzog, C., & Powell-Moman, A. (2005). The contribution of mediator-based deficiencies to age differences in associative learning. *Developmental Psychology, 41*, 389-400.

- Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*, 493-503.
- Hertzog, C., & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (pp. 215-251). San Diego: CA: Academic Press.
- Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 22-34.
- Hertzog, C., Price, J., & Dunlosky, J. (2007). Anchoring to prior recall and insufficient adjustment reduces predictive accuracy: An individual differences analysis. *Manuscript in preparation*.
- Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition, 30*, 67-80.
- Johnson, M. K., & Raye, C. L. (2000). Cognitive and brain mechanisms of false memories and beliefs. In D. L. Schacter & E. Scarry (Eds.), *Memory, brain, and belief* (pp. 35-86). Cambridge, MA: Harvard University Press.
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology, 126*, 349-370.
- Koriat, A., and Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learner's sensitivity to retrieval conditions at test. *Memory & Cognition, 34*, 959-972.

- Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology: General*, *124*, 83-97.
- Lovett, M. C., & Schunn, C. D. (1999). Task representations: Strategy variability, and base-rate neglect. *Journal of Experimental Psychology: General*, *128*, 107-130.
- Matvey, G., Dunlosky, J., Shaw, R. J., Parks, C., & Hertzog, C. (2002). Age-related equivalence and deficit in knowledge updating of cue effectiveness. *Psychology and Aging*, *17*, 589-597.
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1999). The University of South Florida word association, rhyme, and word fragment norms. <http://www.usf.edu/FreeAssociation/>.
- 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.
- Richardson, J. T. (1998). The availability and effectiveness of reported mediators in associative learning: A historical review and an experimental investigation. *Psychonomic Bulletin & Review*, *5*, 597-614.
- Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory & Language*, *51*, 71-79.
- Schneider, W., & Pressley, M. (1997). *Memory development between two and twenty* (2nd ed). Mahwah, NJ: Erlbaum.
- Shaughnessy, J. J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Verbal Learning and Verbal Behavior*, *20*, 216-230.

Footnotes

1. Henceforth the word “trial” will be used to generically refer to study-test trials or to the within-subjects factor. In contrast, the word “List” will be used to differentiate the source of PA items and/or participants’ metamemory judgments.
2. These groups were included to evaluate an anchoring hypothesis relevant to understanding individual differences in knowledge updating, which are reported in a separate paper (Hertzog, Price, & Dunlosky, 2007). Most important for present purposes, the manipulation did not affect any of our central dependent measures and hence we collapsed on this independent variable in our reported analyses. We do not discuss it further.
3. Although there are other methods for measuring the accuracy of metacognitive judgments, we shall limit ourselves to examining absolute accuracy (simple differences between mean judgments and mean recall) because this measure is most relevant to our current aims.
4. A reviewer suggested that the blocking effect could also be due to making the strategy information more salient during testing. Why salience would be greater for blocking than when providing specific information about strategy during test (as in the unpublished experiment we referred to earlier) is unclear, but it cannot be definitively ruled out on the basis of our data.
5. This anchoring hypothesis is consistent with a series of structural equation models conducted on the between-person correlations among these measures that are reported elsewhere (Hertzog et al., 2007).

Table 1. Magnitude of Recall, Predictions, and Postdictions in Experiment 1.

Measure	List 1				List 2			
	<u>Imagery</u>		<u>Repetition</u>		<u>Imagery</u>		<u>Repetition</u>	
	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>
	Instructed							
Recall	66	(1.4)	37	(1.4)	66	(1.5)	41	(1.5)
Global predictions	58	(1.7)	46	(1.5)	54	(2.0)	30	(1.6)
JOLs	49	(1.2)	41	(1.1)	45	(1.3)	33	(1.2)
Confidence judgments	68	(1.3)	40	(1.3)	68	(1.4)	43	(1.5)
Global postdictions	55	(1.7)	29	(1.5)	52	(1.7)	30	(1.5)
	Reported							
Recall	68	(1.4)	30	(1.4)	70	(1.4)	32	(1.4)
Global predictions	58	(1.7)	46	(1.5)	54	(2.0)	30	(1.6)
JOLs	51	(1.2)	37	(1.2)	47	(1.3)	28	(1.2)
Confidence judgments	70	(1.3)	33	(1.4)	71	(1.4)	34	(1.4)
Global postdictions	55	(1.7)	29	(1.5)	52	(1.7)	30	(1.5)

Note. All global judgments refer to global differentiated judgments. JOLs = judgments of learning; Confidence judgments were made for responses during PA recall; Imagery = items for which participants were either instructed to use imagery and/or reported using imagery to study the items; Repetition = items for which participants were either instructed to use rote repetition and/or reported using rote repetition to study the items. Entries are mean percentages (and standard errors) for the measures.

Table 2. Absolute Accuracy of Global Predictions and Judgments of Learning in Experiment 1.

Measure	List 1				List 2			
	<u>Imagery</u>		<u>Repetition</u>		<u>Imagery</u>		<u>Repetition</u>	
	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>
	Instructed							
Global predictions	-7.41	(2.1)	9.15	(2.0)	-13.35	(1.6)	-10.13	(1.5)
Judgments of Learning	-16.96	(1.7)	3.28	(1.5)	-21.66	(1.4)	-8.08	(1.3)
	Reported							
Global predictions	-9.52	(2.1)	16.31	(1.9)	-16.13	(1.6)	-0.51	(1.6)
Judgments of Learning	-16.58	(1.7)	6.83	(1.6)	-22.87	(1.5)	-3.87	(1.4)

Note. Entries are means (and standard errors) of individuals' difference scores between predictions (either global differentiated predictions or judgments of learning) and recall. Imagery = items for which participants were either instructed to use imagery and/or reported using imagery to study the items; Repetition = items for which participants were either instructed to use rote repetition and/or reported using rote repetition to study the items.

Table 3. Absolute Accuracy of Confidence Judgments and Global Postdictions Experiment 1.

Measure	List 1		List 2	
	<u>Imagery</u>	<u>Repetition</u>	<u>Imagery</u>	<u>Repetition</u>
	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)
	Instructed			
Confidence judgments	2.11 (0.5)	2.62 (0.4)	1.15 (0.5)	2.24 (0.4)
Global postdictions	-10.33 (1.2)	-7.88 (1.1)	-14.91(1.2)	-11.67 (1.1)
	Reported			
Confidence judgments	2.56 (0.5)	2.96 (0.5)	1.46 (0.5)	1.72 (0.6)
Global postdictions	-12.31 (1.2)	-0.76 (1.1)	-18.01 (1.2)	-2.64 (1.2)

Note. Entries are means (and standard errors) of individuals' difference scores between postdictions (either confidence judgments or global differentiated postdictions) and recall. Imagery = items for which participants were either instructed to use imagery and/or reported using imagery to study the items; Repetition = items for which participants were either instructed to use rote repetition and/or reported using rote repetition to study the items.

Table 4. Magnitude of Recall, Predictions and Postdictions in Experiment 2.

Measure	List 1				List 2			
	<u>Imagery</u>		<u>Repetition</u>		<u>Imagery</u>		<u>Repetition</u>	
	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>
	Random							
Recall	66	(3.1)	28	(2.5)	62	(3.2)	27	(2.6)
Global predictions	57	(2.5)	49	(2.7)	42	(3.5)	25	(2.2)
JOLs	46	(2.6)	35	(2.3)	36	(2.8)	23	(2.1)
Global postdictions	43	(3.9)	24	(2.4)	34	(3.8)	19	(2.3)
	Blocked							
Recall	70	(3.0)	25	(2.5)	71	(3.1)	26	(2.5)
Global predictions	60	(2.5)	52	(2.7)	56	(3.4)	18	(2.2)
JOLs	46	(2.5)	33	(2.3)	41	(2.8)	20	(2.0)
Global postdictions	61	(3.9)	16	(2.4)	65	(3.8)	16	(2.3)

Note. All global judgments refer to global differentiated judgments. JOLs = judgments of learning; Imagery = items for which participants reported using imagery to study the items; Repetition = items for which participants reported using rote repetition to study the items. Entries are mean percentages (and standard errors) for the measures. Blocked = blocked testing group; Random = random testing group.

Table 5. Absolute Accuracy of Global Differentiated Predictions and Judgments of Learning in Experiment 2.

Measure	<u>List 1</u>		<u>List 2</u>	
	<u>Imagery</u>	<u>Repetition</u>	<u>Imagery</u>	<u>Repetition</u>
	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)
	<u>Random</u>			
Global predictions	-9.26 (3.8)	20.91 (3.6)	-20.20 (3.3)	-2.07 (2.2)
Judgments of Learning	-19.45 (3.5)	6.52 (3.0)	-26.09 (3.0)	-4.09 (2.3)
	<u>Blocked</u>			
Global predictions	-10.04 (3.8)	26.50 (3.5)	-15.42 (3.2)	-7.87 (2.2)
Judgments of Learning	-24.63 (3.5)	8.33 (3.0)	-30.42 (3.0)	-5.74 (2.3)

Note. Entries are means (and standard errors) of individuals' difference scores between predictions (either judgments of learning or global differentiated predictions) and recall. Imagery = items for which participants were either instructed to use imagery and/or reported using imagery to study the items; Repetition = items for which participants were either instructed to use rote repetition and/or reported using rote repetition to study the items.

Table 6. Absolute Accuracy of Global Postdictions in Experiment 2.

Measure	List 1				List 2			
	<u>Imagery</u>		<u>Repetition</u>		<u>Imagery</u>		<u>Repetition</u>	
	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>	<i>M</i>	<i>(SE)</i>
Random	-23.04	(2.5)	-4.57	(1.9)	-27.72	(2.7)	-8.28	(2.0)
Blocked	-9.10	(2.5)	-8.83	(1.9)	-6.78	(2.7)	-9.80	(2.0)

Note. Entries are means (and standard errors) of individuals' difference scores global differentiated postdictions and recall. Imagery = items for which participants were either instructed to use imagery and/or reported using imagery to study the items; Repetition = items for which participants were either instructed to use rote repetition and/or reported using rote repetition to study the items.

Figure Captions

Figure 1. Experiment 1 strategy effectiveness questionnaire ratings for the interactive imagery and rote repetition strategies before and after task experience.

Figure 2. Experiment 2 strategy effectiveness questionnaire ratings for the interactive imagery and rote repetition strategies before and after task experience.



