Memory & Cognition
2010, 38 (3), 366-376
doi:10.3758/MC.38.3.366

Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it

Yana Weinstein and Henry L. Roediger III
Washington University, St. Louis, Missouri

We examined the effect of three variables (test list structure, report option, and framing) on retrospective bias in global evaluations of test performance (postdictions). Participants answered general knowledge questions and estimated correctness of their performance after each block. The ordering of the questions within a block affected bias: Participants believed they had answered more questions correctly when questions were sorted from the easiest to the hardest than when the same questions were randomized or sorted from the hardest to the easiest. This bias was obtained on global postdictions but was not apparent on item-by-item ratings, pointing to a memory-based phenomenon. In addition, forcing participants to produce a response to every question increased performance without affecting evaluations. Finally, framing the evaluation question in terms of the number of questions answered incorrectly (rather than the number correctly answered) did not affect how positively participants evaluated their performance, but did render postdictions less accurate. Our results provide evidence that students' evaluations of performance after a test are prone to retrospective memory biases.

After taking an exam, students retrospectively evaluate their performance and use this evaluation to guide their expectation of the approximate grade they might achieve on that exam. Sometimes this expectation may be accurate, but in other cases students seem surprised by their scores. The factors that affect such postdictions on tests are the focus of this article. Much previous research has focused on metacomprehension (i.e., how well students think they have understood a text; e.g., Maki & Berry, 1984) and predictions of test performance (i.e., how well students think they are going to do on a test before they have taken it, on the basis of how well they know the material; e.g., Glenberg & Epstein, 1985). Less research has investigated the factors that specifically affect global postdictions (i.e., how well students think they have performed once they have taken a test). The benefits of accurate self-evaluation have been discussed elsewhere (e.g., Hacker, Bol, & Keener, 2008), and include improved self-efficacy and more appropriate study behavior. In addition to these benefits, consider also the decision students have to make after taking exams for which they have the option to cancel their scores. For instance, in 2006–2007, 26.3% of students taking the Law School Admission Test (LSAT) canceled their scores and resat the test at least once (*LSAT Repeater Data*, n.d.); there are extensive discussions by students in online forums and even semiprofessional advice is available to help students decide whether to keep or cancel their scores (Ivey, 2005). Awareness of additional factors that bear on students' evaluations of performance following a test would make a valuable contribution to these discussions.

Hacker, Bol, Horgan, and Rakow (2000) showed that postdictions tend to be more accurate than predictions, and concluded that students are generally very accurate on judgments made after a test. Whereas predictions are made prospectively and are based on what students think they know, postdictions are made retrospectively and reflect the student's experience of the test (Hacker et al., 2008). In the absence of objective information about test difficulty, predictions are made entirely on the basis of internal states. Postdictions, on the other hand, may be more reliable insofar as they take test difficulty into account. Nevertheless, postdictions are susceptible to biases just like any metacognitive judgments (Nelson & Narens, 1990), although the sources of bias may be different from those guiding predictions. In addition to biases arising from inaccurate assessment of performance on individual questions (Lichtenstein, Fischhoff, & Phillips, 1982), postdictions are also susceptible to retrospective memory biases that arise from attempts to evaluate the test experience as a whole. To our knowledge, no studies have examined the latter. In the present article, we manipulate three possible sources of retrospective bias to examine their effects on postdictions. Below, we identify three factors that may affect postdictions: test list structure, framing of the postdictions, and report option. All three factors are manipulated in Experiment 1. In Experiments 2 and 3 we further examine the effect of test list structure on postdictions.

The primary focus of this article is test list structure. A given set of questions on an exam can be arranged in any number of ways. Typically, paper-and-pencil tests begin

Y. Weinstein, y.weinstein@wustl.edu

with easy questions and gradually progress to the more difficult items. Standardized tests that adapt to the test-taker's level, on the other hand, may start with more difficult items to gauge ability and then oscillate between easy and difficult items in a way that may appear random to the test taker. Might these different ways of arranging items in a test bias students with respect to their performance evaluation at the end of the test? To our knowledge, this question has only been addressed with respect to the initial item on a computer adaptive test. Although there is no evidence that the difficulty of the initial item affects performance on the test as a whole (Lunz, Bergstrom, & Gershon, 1994), educators have worried that reactions to the initial question may affect test-takers' perceptions of the test (Mills & Stocking, 1996). The one article that directly tested this hypothesis did not find evidence for any influence of the initial question on subsequent performance evaluations (Tonidandel, Quiñones, & Adams, 2002). However, performance evaluations were not directly comparable to scores (i.e., to evaluate their performance, participants responded to qualitative items such as "I did well on this test"), and only the difficulty of the initial question was controlled. In the present article, we directly compared performance with evaluations by asking participants to estimate the number of questions they had answered correctly. We also manipulated the ordering of the questions throughout the test, so questions were either arranged randomly, or arranged from the easiest to the most difficult, or the opposite. We were interested in whether manipulating question ordering would affect the relative difference between performance and evaluations.

The idea that test list structure alone could affect performance evaluations is based on the assumption that the performance evaluation relies on a memory of the test as a whole, and thus could be influenced by runs of easy or difficult questions. Studies that have demonstrated primacy and recency effects in other domains speak to this issue. On one hand, research into global evaluations of hedonic experiences has uncovered a recency effect, such that people are particularly sensitive to the portion of an experience that occurs right before it terminates (Kahneman, Wakker, & Sarin, 1997). For instance, patients undergoing a colonoscopy will remember the experience as having been less painful if the pain decreases toward the end of the procedure than if the event ends more painfully, even though the total amount of pain felt in the two versions of the procedure is identical (Redelmeier & Kahneman, 1996). If evaluations of performance on a test are subject to the same sorts of biases as are evaluations of hedonic experiences, postdictions should be higher when the test ends on a set of relatively easy questions. On the other hand, strong primacy effects have been observed in impression formation; for instance, people tend to focus on the first few adjectives used to describe another person when forming a judgment of their character (Anderson & Barrios, 1961). If this primacy effect holds for evaluations of test performance, participants should be more optimistic in their postdictions after taking a test that began with a run of easy questions. All three experiments in the present article seek to distinguish between these two competing hypotheses (i.e., primacy or recency effects in ordering of questions). Of course, another possibility is simply that the ordering of the questions would have no effect at all relative to the control condition in which questions were ordered randomly.

Another factor that has recently come to light as a potential source of bias in metacognitive judgments is the framing of the question used to elicit the judgment. Finn (2008) showed a framing effect on judgments of learning (JOLs)—that is, predictions of whether particular items would be remembered at test. Participants tend to be overconfident when making these judgments (e.g., Koriat, Lichtenstein, & Fischhoff, 1980). Finn manipulated the framing of the JOL question, so participants were either asked if they would remember (the standard positive framing of the question) or forget (negative framing) items on a later test. JOLs made in the forget frame showed less overconfidence than did those made in the standard remember frame. Using a hypothetical JOL situation with global judgments of forgetting, Koriat, Bjork, Sheffer, and Bar (2004) showed that participants were insensitive to the delay between study and test when they were asked how many words they would remember after various intervals. That is, participants who were asked how many words they might remember in 1 year gave the same response as others in a different group who were asked how many words they might remember on an immediate test. However, when participants made these judgments in a forget frame—being asked how many words they would forget rather than how many they would remember—they correctly predicted forgetting (Koriat et al., 2004, Experiment 7).

Confidence and postdiction judgments in psychology experiments are almost always framed positively; that is, participants are asked how confident they are that an answer is correct, and what number of questions they think they got correct at the end of the test. An alternative question that could be asked at the end of the test is one regarding the number of questions participants think they have gotten wrong. Would an estimate made in response to this negatively framed question be a direct transformation of that made in response to the standard question? Finn's (2008) results suggest that a negatively framed question could produce lower postdictions than the standard, positively framed question by drawing attention to participants' errors rather than their correct responses. In Experiment 1, we manipulated whether participants were asked to estimate the number of questions they had gotten correct or the number they had gotten wrong on an immediately preceding test to examine the effect of framing on postdictions and to determine whether participants engage in different strategies when making these two judgments.

The third factor we investigated with respect to performance evaluations was the criterion used to report answers. Tests vary as to whether they encourage quantity (i.e., answering as many questions as possible in order to maximize the chance of earning points for each question) or accuracy (i.e., only answering questions that one is sure to answer correctly, so as to avoid penalties for wrong answers). Most tests in educational settings use the former technique (with no penalty for guessing), whereas some standardized tests do penalize for guessing, so students usually have some

experience with both types of test. One way to manipulate this factor experimentally is by varying report option (free vs. forced; Koriat & Goldsmith, 1996). Under free report, participants can choose to answer only those questions for which they feel able to produce a potentially correct response. Under forced report, participants must produce an answer to every question. Although there is some evidence from work on free recall that participants can produce more items from a studied list when they are forced to write down more words than they would choose to under free report (Bousfield & Rosner, 1970; but see Roediger & Payne, 1985), Koriat and Goldsmith (1996) showed that forcing participants to produce an answer to every general knowledge question on a test did not increase the number of questions answered correctly in comparison with a free-report condition. Of course, whether performance under forced-report conditions will exceed that under free-report conditions will depend on the nature of the materials and participants' ability to generate plausible answers (Roediger, Srinivas, & Waddill, 1989). For example, research into response strategy on multiple-choice tests has shown that students often withhold correct answers, especially when penalized for errors (Higham, 2007).

Regardless of whether report option affects performance, one might expect it to independently influence performance evaluations. Winkielman, Schwartz, and Belli (1998) showed that participants asked to retrieve 12 childhood memories evaluated their ability to recall their childhoods less favorably than others who were asked to retrieve only 4 such memories, even though both groups were equally successful at retrieving the required number of memories. Winkielman et al. concluded that the relative difficulty of recalling the larger number of memories led those participants who were required to retrieve 12 memories to feel that their memories were poor. Applying these findings to our procedure, we might expect participants to feel unconfident about their performance when they are forced to produce more responses, because they know they are guessing. On the other hand, the act of producing an answer to every question may cause evaluations to be more optimistic if participants focus unduly on the sheer volume of their output and/or if they incorrectly assume that they successfully responded to more questions. This hypothesis was tested in Experiment 1 by a report option manipulation.

EXPERIMENT 1

The general design was the same for all three experiments: Participants generated answers to blocks of 50 general knowledge questions,[1] and after each block they were asked to evaluate their performance on that block. Experiment 1 was designed to examine all three factors identified above as potential sources of variability in retrospective bias. To investigate the effect of test list structure, we manipulated the ordering of questions in each block. Questions were either arranged from the easiest to the most difficult, or they were arranged in a random order. If a primacy effect occurs as in impression formation (Anderson & Barrios, 1961), ordered blocks should elicit more optimistic performance estimates than random blocks. If a recency effect occurs as in evaluations of hedonic experience (Kahneman et al., 1997), the opposite should be true because the most difficult items occurred at the end of the series in the ordered block.

To investigate framing effects, we asked participants in a between-subjects design to estimate how many questions they had gotten correct in each block of 50, or how many questions they had gotten wrong. If focusing on the number of incorrect responses is similar to focusing on forgetting (rather than remembering) when making a JOL (Finn, 2008), evaluations in the negative frame should be lower than those made in the positive frame.

Finally, to examine the effect of report option, we either gave participants the option to skip questions (free report) or else required them to answer every question in a block (forced report). If the effort of answering more questions reduces one's confidence in one's memory as shown by Winkielman et al. (1998), participants should be less optimistic about their performance in the forced-report than in the free-report condition. On the other hand, the sheer volume of responses may inflate confidence in the forced-report condition.

Note that all of the manipulations described above were designed to influence performance evaluations alone without affecting performance itself, and hence are predicted to have an effect on the difference between evaluations and performance (i.e., bias). However, it was also possible that performance could be affected, most notably by report option, where there is some evidence that forced guessing can lead to improvements in performance (Bousfield & Rosner, 1970; Erdelyi, Finks, & Feigin-Pfau, 1989; but see Roediger et al., 1989). In this case, we were interested in whether evaluations would accurately track these effects on performance.

In most research into confidence and evaluations of performance, postdictions are compared directly with performance to obtain estimates of bias. Typically, performance is subtracted from evaluations and the resulting values are compared with zero, so a positive value indicates overconfidence and a negative value indicates underconfidence (Dunlosky & Metcalfe, 2009, p. 49). However, most of this research has used multiple-choice testing, in which scoring is objective. In our study (as in any educational setting when subjective scoring occurs), performance could be affected by the criterion set by the scorers. Specifically, points could be awarded for correctly spelled answers only (strict criterion), or for a predetermined set of variants on the correct answer (more lenient criterion), or for any responses that deviated from the target answer but still indicated some knowledge of the tested fact, such as an author's first name instead of his last (lenient criterion). To anticipate this problem, and because our focus in the present article was on differences between conditions, we avoid comparing bias (the difference between estimates of performance and performance itself) with zero throughout the article. Any such comparison would be tied to the arbitrary scoring criterion. Instead, we always make comparisons of bias *between conditions*. We adjusted postdictions in all conditions for performance as scored using a relatively lenient criterion (as discussed in the Results section), and evaluate the effects

of the manipulated variables on these values (referred to as "bias"). In order to avoid confusion, we discuss performance evaluations in terms of optimism/pessimism rather than overconfidence/underconfidence; thus, participants in one condition may evaluate their performance more or less optimistically than those in another condition.

Method

Participants. Eighty Washington University undergraduates, ages 18 to 21, participated in the study, and were either assigned credit for fulfilling a course requirement or were financially reimbursed for their time.

Design. The experiment was a 2 (test list structure: easy–hard/randomized) × 2 (framing: positive/negative) × 2 (report option: free/forced) mixed design. Test list structure and framing were manipulated orthogonally between participants with 20 participants in each cell, whereas report option was manipulated within participants. We analyzed three dependent variables using this design: performance, bias in postdictions, and absolute error in postdictions. Performance refers to the number of questions answered correctly in each block of 50 questions, scored as described below. Bias in postdictions refers to the difference between postdictions (global estimates of performance for each block of 50 questions) and performance; more specifically, postdictions minus performance. Although higher values indicate overconfidence relative to performance and lower values indicate underconfidence, note that as stated above, these values are not compared with zero but instead are compared between conditions. Finally, absolute error in postdictions refers to the unsigned bias value (the absolute difference between performance and postdictions), and reflects error in evaluations of performance regardless of whether these resulted from undue optimism or pessimism.

Materials. Two sets of 50 general knowledge questions were selected from the Nelson and Narens (1980) norms. The blocks were designed by selecting a range of questions, from those that were easiest to answer (.9 probability of a correct response) to those that were most difficult to answer (.1 probability of a correct response), such that mean performance across each block would be roughly 50% according to the original norms. Each question could be answered by a single word or given name. An example of a difficult question is, "The general named Hannibal was from what city?"; a question of medium difficulty is, "What is the last name of the woman who began the profession of nursing?"; and an easy question is, "What was the name of Tarzan's girlfriend?" (The respective answers are *Carthage*, *Nightingale*, and *Jane*.)

Procedure. Test list structure was manipulated between participants, such that questions in both blocks were either ordered from the easiest to the hardest or randomized. Framing was also manipulated between participants, so that participants evaluated their performance by answering either a positively or negatively framed question. Finally, report option was manipulated within participants, so that each participant answered one block of questions under free-report conditions and the other under forced-report conditions. The presentation order of the two blocks of questions was fixed for all participants such that Question Set 1 always preceded Question Set 2, but half of the participants in each between-subjects condition answered the first set with free-report instructions and the second set with forced-report instructions, and the order of conditions was reversed for the rest of the participants.

Participants were tested individually or in small groups, and the task was fully computerized. Instructions stated that participants would be answering two blocks of 50 general knowledge questions at their own pace, and that their aim was to maximize the number of correct responses. Prior to the start of a block, participants were either instructed that they would be allowed to skip questions if they did not know the answer (free-report block), or that they should try to make a guess on each question (forced-report block). In the free-report block, when presented with a question, participants either typed an answer and pressed ENTER to continue to the next question, or they pressed ENTER without typing an answer to skip the question. In the forced-report block, if participants pressed ENTER without typing in an answer, they were alerted with a prompt requesting them to enter a response, and they could not continue to the next question until they had done so. Questions in each of the two blocks were either ordered from the easiest to the hardest question or presented in a fixed random order. This manipulation was not made explicit to participants. No mention was made of performance evaluation until participants had answered the first block of questions. At this point, they were either asked to estimate how many questions they thought they had gotten correct out of the 50 questions they had answered (positive framing) or how many questions they thought they had gotten wrong (negative framing). Participants made their evaluations by typing in a number from 0 to 50. After making their evaluation on the first block, participants were presented with instructions for the second block, at the end of which they made the same evaluation. At the end of the 30-min experiment, participants were presented with their scores and evaluations on each block. For the purposes of computerizing this feedback, the program calculated the number of correct responses automatically and thus only exact spellings were accepted as correct responses. However, as discussed below, we then scored the data more leniently for the purpose of analyses.

Results

The basic design for all analyses reported below was a 2 × 2 × 2 mixed-design ANOVA with report option (free/forced) as the within-subjects variable, and test list structure (easy–hard/random) and framing (positive/negative) as between-subjects variables. As described above, there were three dependent measures: performance, bias in postdictions, and absolute error in postdictions. We report analyses for each of these measures in turn. Postdictions, performance, and the difference between the two (i.e., bias) are shown broken down by test list structure and report option in Table 1. Framing did not have an effect on any of these measures and thus data were collapsed across positive and negative framing conditions. Framing did, however, affect the third measure, absolute error. These results are not presented in Table 1, in the interest of conciseness, but are discussed below.

Scoring. Participants' responses were initially scored as described above: One point was given for each answer that was identical to the correct answer. In addition, two blind judges scored responses and assigned a point for any misspelled answers. Examples of misspellings include "copenhaagen" for "copenhagen," "pompei" for "pompeii," and "spudnik" for "sputnik." (Capitalization was not taken into consideration since the program was set up to allow online lowercase characters.) In addition, responses

Table 1
Mean Postdictions, Performance, and Bias (i.e., the Difference Between the Two) by Test List Structure and Report Option in Experiment 1

	Performance		Postdictions		Bias (Difference)	
Condition	*M*	*SE*	*M*	*SE*	*M*	*SE*
Random						
Free	51.2	2.3	48.0	2.8	−3.2	2.0
Forced	53.6	2.5	46.0	3.1	−7.6	2.5
Easy–Hard						
Free	46.4	2.7	48.9	3.4	2.5	1.8
Forced	49.2	2.5	47.8	3.3	−1.4	2.3

were scored as correct when two letters were accidentally switched due to mistyping, and when the last letter was omitted, which would have resulted from participants hitting the ENTER key too quickly. Finally, answers that were given in the plural instead of the required singular, and vice versa, were allowed. Such errors accounted for an increase in performance of 4% (2 questions) for each block of 50 questions. There were no significant differences between conditions in the number of misspellings, apart from the fact that one of the two question sets elicited a larger number of misspellings than the other. Naturally, since assignment of question sets to report option conditions was counterbalanced between participants, this difference had no systematic effect on any comparisons of interest. Note that all analyses in this and subsequent experiments were also carried out using the strict criterion whereby points were only assigned for answers identical to the correct response, and these analyses yielded the same pattern of results as those with the more lenient criterion in all cases. The scoring criterion does not affect any of our conclusions because all conclusions refer to comparisons between conditions, which persisted regardless of scoring criteria.

Performance. Mean performance in terms of the percentage of correct responses in each block is summarized in the far left column of Table 1, broken down by test list structure and report option. In accordance with the Nelson and Narens (1980) norms, participants answered approximately half of all questions correctly in each block. Although no differences in performance between conditions were expected, the mixed-design ANOVA revealed a significant effect of report option such that participants achieved 2.5% higher scores when they were forced to respond to each question ($M = 51.3\%$, $SEM = 1.8$) than when they were allowed to skip questions [$M = 48.8\%$, $SEM = 1.8$; $F(1,76) = 7.95$, $MS_e = 32.8$, $p = .006$, $\eta_p^2 = .10$].[2] Even though there was an apparent (4.6%) difference in performance between the easy–hard and random conditions, this difference did not approach significance ($p = .19$); recall that this variable was manipulated between participants. Framing did not have any effect on performance, and there were no significant interactions.

Bias in postdictions. Bias (the difference between postdictions and performance) is shown in the far right column of Table 1. Bias was calculated by subtracting actual performance from estimated performance (postdictions, reproduced in the middle column of Table 1) for each participant in each condition. As mentioned above, the comparison of these values with zero depends on the scoring criterion, so such comparisons are not made. Instead, bias data were compared between conditions for the manipulated variables (framing, test list structure, and report option) to examine their effect on bias. A mixed-design ANOVA revealed a significant effect of the within-subjects variable report option, such that participants were relatively more optimistic about their performance in the free-report block than in the forced-report block [$F(1,76) = 5.07$, $MS_e = 131.0$, $p = .03$, $\eta_p^2 = .06$]. Furthermore, the analysis also revealed a significant effect of the between-subjects variable of test list structure, such that participants in the easy–hard condition were more optimistic about their predicted performance relative to their actual performance than were participants in the random condition [$F(1,76) = 5.63$, $MS_e = 249.6$, $p = .02$, $\eta_p^2 = .07$]. There were no other significant main effects or interactions.

Absolute error in postdictions. Finally, we also calculated the absolute (unsigned) difference between performance and postdictions of that performance. We did not include these data in Table 1 for conciseness, and also because absolute error did not differ between conditions in subsequent experiments. A mixed-design ANOVA revealed a significant effect of report option, such that participants were less accurate in evaluating their performance when they were forced to provide an answer to each question (absolute error of $M = 12.6\%$) than when they were allowed to skip questions (error of $M = 10.0\%$) [$F(1,76) = 4.80$, $MS_e = 55.3$, $p = .03$, $\eta_p^2 = .06$]. Framing also had a marginally significant effect on the accuracy of performance estimates: Estimates made with negative framing (i.e., in answer to the prompt "How many questions do you think you got wrong?") were made less accurately (error of $M = 12.6\%$) than were estimates made with the standard positive framing (error of $M = 10.0\%$) [$F(1,76) = 3.00$, $MS_e = 91.9$, $p = .09$, $\eta_p^2 = .04$]. No interactions reached significance.[3]

Discussion

Three variables were examined for their potential effect on retrospective assessment of test performance in Experiment 1: report option, framing, and test list structure. First, bias was affected by the ordering of the questions within each block. Participants rated their performance as best when questions were ordered from the easiest to the hardest relative to when questions were randomized. This pattern of data suggests that performance evaluations made at the end of a test are susceptible to memory-based distortions, and conceptually replicates the primacy effects found in impression formation (Anderson & Barrios, 1961): The run of easy questions at the start of the block created a more positive impression of the experience and led to higher estimates of performance without an analogous improvement in actual performance.

Second, although differences in performance as a function of report option were not expected on the basis of previous work by Koriat and Goldsmith (1996), participants did perform better when they were forced to respond to every question. However, this outcome was not reflected in their estimates: Estimates for the free and forced blocks did not differ. As a result, participants were less optimistic on the forced-report block than on the free-report block. In addition, the fact that participants were less accurate in postdicting their performance on the forced block (as evidenced by a higher absolute error in postdictions on this block) indicates that the act of producing guesses impaired their ability to evaluate their performance. Although guessing led to more correct answers being given, participants were not aware of this benefit. This result fits in with previous work on retrieval effort showing that the act of retrieving more items can harm self-perceptions with regard to the quality of one's memory, even if retrieval was successful (Winkielman et al., 1998).

Finally, unlike Finn's (2008) and Koriat et al.'s (2004) findings with JOLs, framing of the postdiction question did not affect bias; contrary to our prediction, participants who made postdictions in the negative framing were not more pessimistic than those who made postdictions in the positive framing. However, participants who evaluated their performance in terms of the number of questions they got wrong (negative framing) were less accurate in their evaluations in terms of absolute error than were participants who evaluated their performance in terms of the number of questions they got correct (positive framing). This effect can be explained by the fact that the latter judgment is the one most often used to evaluate performance, and thus is one that participants would have been better practiced in making. Note that whereas in previous research negative framing has been shown to reduce inaccuracies in JOLs, this is true only insofar as it served to decrease overconfidence. Since participants were not overconfident in our study even when asked to evaluate performance using the standard positive framing (at least, not with performance scored with the lenient criterion), the negative framing could provide no such benefit.

Two questions relating to test list structure cannot be answered by Experiment 1. First, although we demonstrated a primacy effect whereby an initial run of easy questions resulted in more optimism than random ordering, we did not have the reverse condition in which questions were arranged from difficult to easy. If the primacy effect holds, this condition should result in pessimistic evaluations relative to the random condition; Experiment 2 included this condition to test this hypothesis. Second, from the data reported so far, it is not clear by what process the ordering of the questions shifts bias. Does answering easy questions at the beginning of the block lead to a halo effect of optimism, so that participants might be more optimistic regarding their responses, even to more difficult questions later in the block, or does the bias arise only when global memory judgments are made at the end of the block? Experiment 3 was designed to distinguish between these possibilities.

EXPERIMENT 2

Experiment 2 was designed to further explore the effect of test list structure on bias. The first objective of this experiment was to replicate the effect obtained in Experiment 1 in a within-subjects design. A further objective was to determine whether ordering questions from the hardest to the easiest would have an effect on bias that was opposite to that found in the easy–hard ordering condition. If the primacy effect obtained in Experiment 1 is robust, participants should be most optimistic in the easy–hard ordering condition, and most pessimistic in the hard–easy ordering condition, with the random condition between the two.

Method

Participants. Thirty Washington University undergraduates, ages 18 to 21, participated in the study, and were either assigned credits or were financially reimbursed for their time.

Design. The experiment employed a within-subjects design with test list structure (easy–hard/hard–easy/random) as the only manipulated variable. As in Experiment 1, we analyzed three dependent measures: performance, bias in postdictions, and absolute error in postdictions.

Materials. Three new sets of 50 general knowledge questions were selected from the Nelson and Narens (1980) norms according to the same criteria as those in Experiment 1.

Procedure. The procedure was identical to that of Experiment 1 except that participants answered and made postdictions on three blocks of 50 questions, and they were allowed to skip questions to which they did not know the answer on all blocks (free report). The experiment took 30 min to complete. Participants answered one block of questions in each ordering condition. The presentation order of the three sets of questions was fixed for all participants such that Question Set 1 always preceded Question Set 2, and so on, but the order of the three test list structure conditions was counterbalanced across participants using three different presentation orders so that each of the three test list structure conditions was applied to the first, second, and third blocks for 10 participants each. Unlike in Experiment 1, in which the order of questions in the random condition was fixed across participants, in this case a new random order was created for each participant for the random block.

Results

The basic design for all analyses reported below was a repeated measures ANOVA with test list structure (easy–hard/hard–easy/random) as the within-subjects variable. As in Experiment 1, there were three dependent measures: performance, bias in postdictions, and absolute error in postdictions. We report analyses for each of these measures in turn. Postdictions, performance, and the difference between the two (i.e., bias) are shown, broken down by test list structure, in Table 2.

Scoring. Performance was scored by the same criteria as in Experiment 1. Misspellings and other typing errors accounted for 3.1% of all responses, and their occurrence did not vary between test list structure conditions, although as in Experiment 1, one of the three question sets elicited a larger number of misspellings. Once again, because an equal number of participants was assigned each set of questions under each test list structure condition, this was not a concern.

Performance. Mean performance in terms of the percentage of correct responses in each block is summarized in the far left column of Table 2, broken down by test list structure. Participants correctly reported 43.2% of the answers across all three blocks of 50 questions, and there was a marginally significant effect of test list structure on performance [$F(2,58) = 3.11$, $MS_e = 25.4$, $p = .05$, $\eta_p^2 = .10$]. In particular, participants performed 3.2% better in the easy–hard condition than in the random condition [$t(29) = 2.47$, $SEM = 1.30$, $p = .02$] (not significant with Bonferroni correction).

Table 2
Mean Postdictions, Performance, and Bias by Test List Structure in Experiment 2

Test List Structure	Performance		Postdictions		Bias (Difference)	
	M	*SE*	*M*	*SE*	*M*	*SE*
Hard–Easy	43.5	3.3	35.9	3.3	−7.7	2.2
Random	41.5	2.8	37.7	3.3	−3.8	2.0
Easy–Hard	44.7	3.3	46.9	3.5	2.2	2.1

Bias in postdictions. Bias data (i.e., performance subtracted from postdictions) are shown in the far right column of Table 2 for each test list structure condition. Test list structure had a significant effect on bias [$F(2,58) = 8.67$, $MS_e = 85.6$, $p < .001$, $\eta_p^2 = .23$]. Participants were more optimistic about their performance in the easy–hard condition than in both the hard–easy [$t(29) = 4.01$, $SEM = 2.46$, $p < .001$] and random conditions [$t(29) = 2.09$, $SEM = 2.45$, $p = .008$]. Although the difference in bias between the hard–easy and random conditions was in the predicted direction—that is, participants were less optimistic about their performance when questions were ordered from the hardest to the easiest compared with when question order was randomized—this difference did not reach significance ($p = .15$).

Absolute error in postdictions. The absolute difference between performance and evaluations of performance was 9.7% across all three test list structure conditions and did not vary with test list structure. Participants were equally accurate in evaluating their performance in the three test list structure conditions.

Discussion

In Experiment 2, we replicated the effect of test list structure on bias in a within-subjects design. Specifically, we showed once again that arranging questions from the easiest to the hardest created an illusion of competence compared with a randomized question order and compared with arranging the questions from the hardest to the easiest. The latter condition also produced some pessimism compared with the randomized condition, although this negative primacy effect from an initial run of difficult questions was not reliable. It could be that the negative manipulation was less effective because the difficult questions were not as difficult as the easy questions were easy, but this does not seem to be the case. As a manipulation check, we looked at the number of questions participants answered correctly (or incorrectly) in the first 10 trials of the easy–hard (hard–easy) blocks, respectively. Whereas participants answered 79.6% of the first 10 questions correctly in the easy–hard condition, they answered only 11.7% of the first 10 questions correctly in the hard–easy condition (hence 88.3% of these questions were answered incorrectly). Looking at performance this way, it seems as though the hard–easy manipulation was, if anything, stronger than the easy–hard manipulation, with respect to primacy.

A pessimistic theory of the way participants make metacognitive judgments is that they anchor around a realistic value and only achieve metacognitive accuracy through common sense adjustments (Scheck, Meeter, & Nelson, 2004). When evaluating their performance on a test, for instance, students might anchor around a number such as 50%, expecting that an average test would lead to this level of performance. The monitoring hypothesis suggests that various factors could then push this value around; so, in our experiments, the experience of easy questions at the beginning may have served to adjust evaluations of performance upward. This anchoring and adjustment process could help explain why evaluations in the hard–easy condition were more similar to the randomized condition than were those in the easy–hard condition. If participants are used to taking tests in which questions are arranged from the easiest to the hardest (and many pencil-and-paper tests are structured in this way), this condition may thus be the one in which participants are relying most on their anchor value. The hard–easy and random conditions, on the other hand, may be experientially more similar, because in neither condition does difficulty build up throughout the test, as participants may expect it to. Furthermore, as questions get easier in the hard–easy condition, participants may interpret easier questions as being more difficult than they really are because of the expectation that questions should increase rather than decrease in difficulty as the test goes on. The fact that evaluations in the easy–hard condition are more optimistic than in the other two conditions could thus reflect adjustment down from the anchor for both the randomized and hard–easy conditions.

EXPERIMENT 3

So far, we have demonstrated that performance evaluations can be shifted simply by changing the order in which questions are arranged within a test, both when this variable is manipulated between participants (Experiment 1) and within participants (Experiment 2). Because the questions themselves remain exactly the same between conditions, it seems reasonable to assume that the effect of ordering the questions occurs at the time when participants are thinking back on the block of questions as a whole and making a global judgment of performance on the basis of their memory for that block. On the other hand, there was some evidence in Experiment 2 that participants in the within-subjects version of the task actually performed better when questions were arranged in order from easiest to hardest (even though we used free-report testing). It is possible that participants felt more motivated in this condition as a result of successfully producing an answer for a large proportion of questions at the start of the block. One question that lends itself easily to investigation is whether the ordering of the questions affects confidence on an item-by-item basis throughout the test, or whether these shifts in bias only occur retrospectively.

Experiment 3 was designed to establish whether the bias effects observed on a global level also appear as participants work through the questions. Participants answered blocks of questions that were either ordered from the easiest to the hardest or randomized, but also rated their confidence in each response on some blocks. If an initial run of easy questions results in increasingly optimistic evaluations of performance at the time the questions are answered, it should spill over into confidence judgments made on later questions. If, on the other hand, the shift in bias is purely a retrospective, memory-based phenomenon, there should be no difference in confidence ratings between conditions while taking the test.

The item-by-item confidence ratings also allowed us to look at both relative metacognitive accuracy, also known as *discrimination* (the relationship between confidence values and accuracy on each response; Lundeberg, Fox, & Punćochaŕ, 1994); and absolute metacognitive accuracy,

or *calibration* (the relationship between mean confidence ratings across all items and overall performance; Keren, 1991), and how these measures related to the global postdiction judgments made at the end of each block. To ensure that the act of making confidence judgments did not in itself affect postdictions, participants also completed one block in each of the two ordering conditions without item-by-item ratings. The inclusion of these blocks also permitted us to fulfill a secondary goal: to examine whether making confidence judgments on an item-by-item basis increases absolute accuracy in evaluations by making participants more aware of how they are performing throughout the test. If this is the case, the absolute (unsigned) difference between performance and postdictions should be lower for the blocks on which participants made item-by-item confidence judgments.

Method

Participants. Thirty-six Washington University undergraduates, ages 18 to 21, participated in the study, and were either assigned credits or financially reimbursed for their time.

Design. The experiment employed a 2 (test list structure: easy–hard/random) × 2 (confidence ratings: included/omitted) within-subjects design. As in previous experiments, we analyzed three dependent measures: performance, bias in postdictions, and absolute error in postdictions. In addition, for the two blocks on which participants made item-by-item confidence ratings, we analyzed calibration and discrimination. For calibration, we averaged confidence judgments across both correct and incorrect responses to produce one value per participant for each of the two test list structure conditions. This value was then compared with both performance and global postdictions. Thus, calibration amounts to the relationship between average confidence across items and overall performance. For discrimination, we calculated mean confidence in correct versus incorrect responses, and compared these between test list structure conditions.

Materials. Four sets of 50 general knowledge questions were selected from the Nelson and Narens (1980) norms according to the same criteria as were used in Experiments 1 and 2.

Procedure. Participants answered and evaluated their performance on four blocks of 50 questions with test list structure manipulated within participants, so that each participant completed two blocks of questions that were ordered from the easiest to the hardest question and two blocks of questions that were randomized afresh for every participant. In addition, participants made item-by-item confidence ratings on one each of the two types of blocks. Item-by-item ratings were made either on the first two or the last two blocks, and this factor was counterbalanced between participants. The order of test list structure conditions was also randomized between conditions, whereas the four question sets were always presented in the same sequence. As in Experiment 2, participants were allowed to skip questions they could not answer (free report). On confidence-rated blocks, after participants made a response, they were presented with a slider and asked to rate their confidence in their response on a scale from 0 to 100. Prior to being given feedback on their overall performance, participants completed a questionnaire about their exam preparation techniques. The experiment took 60 min to complete.

Results

The basic design for all analyses reported below was a repeated measures ANOVA with test list structure (easy–hard/random) as the within-subjects variable. As in previous experiments, there were three dependent measures: performance, bias in postdictions, and absolute error in postdictions. These measures did not differ between blocks that included and omitted item-by-item confidence ratings. Postdictions, performance, and the difference between the two (i.e., bias) are thus shown in Table 3, collapsed across these two types of blocks, but broken down by test list structure condition. In addition, calibration and discrimination are analyzed for blocks on which item-by-item confidence ratings were collected. We report analyses for each of these measures in turn.

Scoring. Performance was scored as in previous experiments, with on average 3.3% of responses per block accounted for by misspellings. The number of misspellings once again did not vary by condition, although some sets of questions elicited more such errors than others. Participants attempted to answer 56.8% of questions across blocks, and the number of questions attempted did not differ between conditions.

Performance. Performance (reported in the far left column of Table 3) was 40.9% across all four blocks and did not vary significantly by test list structure, although there was a numerical difference in performance in the same direction as that reported in Experiment 2, with performance 1.5% higher in the easy–hard condition ($p = .11$). Making item-by-item confidence ratings did not affect performance.

Bias in postdictions. The effect of test list structure on bias (reported in the far right column of Table 3) was the same as in previous experiments: Participants showed a 3.2% difference in bias between the easy–hard and random conditions [$F(1,35) = 5.01$, $MS_e = 74.6$, $p = .03$, $\eta_p^2 = .13$]. As in Experiments 1 and 2, participants evaluated their performance more optimistically in the easy–hard condition. Making item-by-item confidence ratings did not affect bias.

Absolute error in postdictions. The absolute error margin between estimates and performance was 10.2% across all blocks, and this did not vary systematically between conditions. As in previous experiments, test list structure did not affect accuracy in evaluations of performance. In addition, making item-by-item confidence judgments did not affect postdiction accuracy.

Calibration. For calibration, we were interested in the relationship between average confidence across items and overall performance. This analysis was performed on the two blocks on which participants made item-by-item ratings. Because participants only attempted, on average, 28 questions in each block of 50, there were not enough observations to break down accuracy by confidence and produce calibration curves or compute signal detection measures. Instead, we calculated bias in item-by-item confidence ratings by the same method as was used to calculate bias in global postdictions throughout the article, as

Table 3
Mean Postdictions, Performance, and Bias by Test List Structure in Experiment 3

Test List Structure	Performance		Postdictions		Bias (Difference)	
	M	*SE*	*M*	*SE*	*M*	*SE*
Random	40.1	1.8	37.3	2.5	−2.8	1.9
Easy–Hard	41.6	1.8	42.0	2.6	0.4	2.3

suggested by Dunlosky and Metcalfe (2009, p. 50). Confidence judgments made on each item were averaged to produce an overall confidence value for each participant in each block (mean confidence across all blocks of 71.7%). In addition, we also calculated the percentage of questions participants got correct, of those questions they actually attempted (M = 73.0%). We did this because it did not make sense to include those questions that participants did not attempt, and hence for which there were no confidence ratings collected. In order to calculate calibration, performance on attempted questions was subtracted from mean confidence for each block. Contrary to the global postdictions, participants made slightly more optimistic confidence ratings in the random condition (confidence ratings were on average 2.1% higher than performance) than in the easy–hard condition (where mean confidence ratings were 0.4% lower than performance). A repeated measures ANOVA with test list structure (easy–hard/random) revealed no significant difference between the easy–hard and random conditions on this calibration measure (p = .14). To compare confidence ratings with postdictions, two separate correlations were conducted for the easy–hard and random conditions. In neither condition did the correlation of mean confidence ratings and postdictions come close to being significant ($p > .5$).

Discrimination. For discrimination, mean confidence was compared for correct and incorrect responses in the easy–hard and random conditions. Participants assigned much higher confidence ratings to correct (overall M = 82.4) than to incorrect (overall M = 46.5) items [$F(1,35)$ = 213.14, MS_e = 85.0, $p < .001$, η_p^2 = .86]. However, these ratings did not differ between test list structure conditions, nor was there an interaction.

Discussion

Experiment 3 once again replicated the effect of test list structure on bias: Participants were more optimistic about their performance on blocks in which questions were ordered from the easiest to the hardest in comparison with randomized blocks. In addition, we showed that this bias in performance evaluations is probably a retrospective memory distortion rather than an affective bias that has its effect during the test itself. This was evidenced by the lack of difference in average confidence ratings assigned to answered questions in the two test list structure conditions during the test. Having to make confidence ratings for each response did not affect the global judgments. Interestingly, there was also no significant relationship between confidence ratings and the global performance evaluations. In theory, if item-by-item confidence judgments are reflected in people's evaluations of their performance, mean confidence and postdictions should be closely related. However, item-by-item ratings could be influenced by affective heuristics such as familiarity of the information presented in a question, regardless of the difficulty of the question itself, or the number of candidate answers that participants generated before selecting their best guess. Stankov and Crawford (1996) reported very low correlations between item-by-item confidence ratings and global postdictions, concluding that different processes underlie item-by-item confidence judgments and global postdictions of performance. Our data point to a similar conclusion.

GENERAL DISCUSSION

In three experiments, we investigated the effects of three different factors on bias in participants' evaluations of their performance after taking a test. First, we looked at the effect of test list structure. The bulk of the article focused on the surprising result that merely changing the order of questions at test can significantly affect bias. Ordering questions from easy to hard produced significantly more optimistic evaluations of performance than did random ordering, both between participants (Experiment 1) and within participants (Experiments 2 and 3). One potential explanation for the effect is the use of an affect heuristic (Slovic, Finucane, Peters, & MacGregor, 2002). That is, a run of easy questions at the start of the block may produce positive affect, which is then used as a heuristic basis for the more optimistic estimates. However, in Experiment 3 we demonstrated that this effect did not seem to be driven by an increase in item-by-item confidence or a decreased ability to distinguish between correct and incorrect responses. In other words, the effect appears to be a result of a retrospective memory bias, more specifically, a primacy bias, as demonstrated in impression formation research (Anderson & Barrios, 1961) and unlike the recency effect demonstrated with respect to hedonic experiences (Kahneman et al., 1997). This outcome makes sense, given that our procedure is closer in nature to the procedure of impression formation than to that of ongoing experience, because it involves discrete verbal events as opposed to a fluid experience.

Another possible interpretation described in the Discussion of Experiment 2 is the anchoring and adjustment heuristic (Scheck et al., 2004). This heuristic suggests that participants already have an idea of how they are going to score on a test even before they start, and then adjust up or down following their experience of the test. The structure of the test could be one factor in this adjustment process. This explanation fits in well with Hacker et al.'s (2000) finding that postdictions in a classroom setting were more accurate than predictions because students adjusted down from unrealistically high predictions after they had completed the test.

Second, we examined whether the framing of the evaluation question affected bias (Experiment 1). On the basis of previous work by Finn (2008), we predicted that participants would produce lower estimates when evaluating their performance in terms of the number of questions they answered incorrectly, as opposed to the more standard framing in terms of number of questions correct. However, no such effect occurred in Experiment 1. Instead, framing affected the absolute accuracy of these judgments (i.e., the unsigned difference between estimates and performance), so that participants were less accurate in evaluating their performance under the negative framing. As previously noted, one possible explanation of this result is that participants would be more familiar with the standard positive framing and thus more adept at making this metacognitive judgment. Another possibility is that the negative framing engages a larger variety of processes than does the positive framing,

and thus introduces additional noise into the judgments. For instance, participants may approach the task by estimating the number of questions they think they got correct and subtracting this number from the total number of questions (and this in itself would introduce additional noise due to errors in calculation). Alternatively, participants might think about the number of questions they left blank (at least, in the free-report condition) and anchor their evaluations to this number. The explanation that participants are engaging in a larger number of strategies in the negative framing condition is also supported by the higher variability in performance evaluations in this condition.

Third, participants were less optimistic about their performance when they were forced to produce an answer to every question than when they were allowed to leave some questions blank (Experiment 1). More specifically, although forced guessing led to higher scores, participants' evaluations did not track this pattern. The fact that forced guessing significantly improved performance on our task is somewhat surprising, given that this did not occur in Koriat and Goldsmith's (1996) experiments using a similar procedure. However, as Koriat and Goldsmith indicated in their discussion, whether a forced-report manipulation has an effect on performance depends on the guessing base rate (i.e., the likelihood that a guess will yield a correct response) for the questions involved (Erdelyi et al., 1989; Roediger et al., 1989). Because we used different general knowledge questions from those in Koriat and Goldsmith's experiments, it is likely that the guessing base rate differed between the two. Performance evaluations not only failed to track the gains of forced responding, but in fact were numerically higher in the free-report condition. To account for this outcome, we proposed that the act of producing responses to a larger number of questions paradoxically lowered participants' confidence in their memory, perhaps by a similar process to that in Winkielman et al.'s (1998) demonstration of negative self-evaluations of one's memory following effortful retrieval. Put another way, it may be that participants were implicitly using what Koriat and Goldsmith call "output-bound scoring" when assessing their performance, asking "What percentage of the responses I gave were correct?" (rather than "For what number of items did I provide a correct response?" akin to input-bound scoring). If participants were using the former heuristic question, then the lower evaluation of performance would be expected—with all the guessing in forced responding, accuracy did suffer using output- bound scoring.

Winkielman et al.'s (1998) results and the effect of forced report on evaluations described above point to an intriguing possible alternative explanation for the shift in bias that results from changes in question order. Although we have been assuming throughout most of the article that the increased optimism of evaluations in the easy–hard condition results from the run of easy questions at the start of the block, it could be that the bias is actually related to the number of questions participants attempt to answer (and hence, the amount of retrieval effort) rather than the number of questions participants answer correctly. In Experiment 2, under free report, participants produced an answer to 91.3% of the first 10 questions in the easy–hard condition, but they only attempted 36.7% of the first 10 questions in the hard–easy condition. These percentages were similar when these sets of 10 questions appeared at the end of the block. Thus, at the end of an easy–hard block, participants would be skipping many questions, and this decrease in amount of retrieval effort expended toward the end of the block could be driving the increased optimism in evaluations. However, the forced-report condition of Experiment 1 provides some evidence that this is not the case. In this condition, retrieval effort was presumably equated between test list structure conditions because participants were producing responses to almost every question, and yet participants still showed a 6.2% difference in bias between the randomized and easy–hard conditions. This outcome indicates that the initial run of easy questions is more likely to be responsible for the increased optimism of evaluations in the easy–hard condition than is reduced retrieval effort on harder questions at the end of the block.

A fruitful direction for future research would be to firmly establish the causes of shifts in bias due to test list structure. It would also be interesting to look at alternative test list structures such as U-shaped difficulty functions or tests that begin with an easy run of questions but then switch to randomized ordering. Tonidandel et al. (2002) have already shown that the difficulty level of the initial question is not a significant predictor of performance evaluations. Further research could establish whether evaluations of performance are only influenced by test list structure when it is maintained throughout the test, or whether different parts of the test (e.g., the beginning) influence evaluations more strongly.

To gauge interest in the issue of evaluating performance after taking tests, a questionnaire was administered to the 36 participants in Experiment 3. Although these participants (all Washington University in St. Louis undergraduates) reported to be generally quite accurate in evaluating their performance after a test ($M_{\text{rating}} = 5.4$ on a scale from 1 = *not at all accurate* to 7 = *extremely accurate*), they were also reasonably interested in learning tips to improve this accuracy ($M_{\text{rating}} = 4.4$ on a scale from 1 = *not at all interested* to 7 = *very interested*). Furthermore, 81% of respondents reported engaging in strategies in order to try to determine their performance on an exam, and 64% of respondents were able to recall a time when they were particularly surprised by a grade in comparison with their expectations. The evaluation of performance after a test thus appears to be an issue close to the hearts of undergraduate students, and the experiments reported here speak to this important issue. Teachers should know that if they provide a block of easy questions at the beginning of the test, they may have more students surprised that the outcome of the test results was below their expectations.

AUTHOR NOTE

Support for this research was provided by a James S. McDonnell Foundation 21st Century Science Initiative grant: Bridging Brain, Mind and Behavior/Collaborative Award. Thanks to Kristy Duprey for assistance with data collection. Correspondence concerning this article should be addressed to Y. Weinstein, Department of Psychology, Box 1125, Washington University, One Brookings Drive, St. Louis, MO 63130 (e-mail: y.weinstein@wustl.edu).

REFERENCES

ANDERSON, N. H., & BARRIOS, A. A. (1961). Primacy effects in personality impression formation. *Journal of Abnormal & Social Psychology*, **63**, 346-350. doi:10.1037/h0046719

BOL, L., & HACKER, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *Journal of Experimental Education*, **69**, 133-151.

BOUSFIELD, W. A., & ROSNER, S. R. (1970). Free vs. uninhibited recall. *Psychonomic Science*, **20**, 75-76.

DUNLOSKY, J., & METCALFE, J. (2009). *Metacognition.* Thousand Oaks, CA: Sage.

ERDELYI, M. H., FINKS, J., & FEIGIN-PFAU, M. B. (1989). The effect of response bias on recall performance, with some observations on processing bias. *Journal of Experimental Psychology: General*, **118**, 245-254. doi:10.1037/0096-3445.118.3.245

FINN, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition*, **36**, 813-821. doi:10.3758/MC.36.4.813

GIGERENZER, G., HOFFRAGE, U., & KLEINBÖLTING, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98**, 506-528. doi:10.1037/0033-295X.98.4.506

GLENBERG, A. M., & EPSTEIN, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 702-718. doi:10.1037/0278-7393.11.1-4.702

HACKER, D. J., BOL, L., HORGAN, D. D., & RAKOW, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, **92**, 160-170. doi:10.1037/0022-0663.92.1.160

HACKER, D. L., BOL, L., & KEENER, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory.* (pp. 429-455). New York: Psychology Press.

HIGHAM, P. A. (2007). No Special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General*, **136**, 1-22. doi:10.1037/0096-3445.136.1.1

IVEY, A. (2005). *The Ivey guide to law school admissions: Straight advice on essays, resumes, interviews, and more.* PA: Harvest Books.

KAHNEMAN, D., WAKKER, P. P., & SARIN, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, **112**, 375-405. doi:10.1162/003355397555235

KEREN, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, **77**, 217-273. doi:10.1016/0001-6918(91)90036-Y

KORIAT, A., BJORK, R. A., SHEFFER, L., & BAR, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, **133**, 643-656. doi:10.1037/0096-3445.133.4.643

KORIAT, A., & GOLDSMITH, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, **103**, 490-517. doi:10.1037/0033-295X.103.3.490

KORIAT, A., LICHTENSTEIN, S., & FISCHHOFF, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 107-118. doi:10.1037/0278-7393.6.2.107

LICHTENSTEIN, S., FISCHHOFF, B., & PHILLIPS, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge: Cambridge University Press.

LSAT Repeater Data (n.d.). Retrieved June 3, 2009, from www.lsac.org/pdfs/RepeaterData.pdf.

LUNDEBERG, M. A., FOX, P. W., & PUNĆOCHAŘ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, **86**, 114-121. doi:10.1037/0022-0663.86.1.114

LUNZ, M. E., BERGSTROM, B. A., & GERSHON, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research*, **21**, 623-634.

MAKI, R. H., & BERRY, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 663-679. doi:10.1037/0278-7393.10.4.663

MILLS, C. N., & STOCKING, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, **9**, 287-304. doi:10.1207/s15324818ame0904_1

NELSON, T. O., & NARENS, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning & Verbal Behavior*, **19**, 338-368. doi:10.1016/S0022-5371(80)90266-2

NELSON, T. O., & NARENS, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125-173). New York: Academic Press.

PRESSLEY, M., GHATALA, E. S., WOLOSHYN, V., & PIRIE, J. (1990). Sometimes adults miss the main ideas and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly*, **25**, 232-249. doi:10.2307/748004

REDELMEIER, D. A., & KAHNEMAN, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, **66**, 3-8. doi:10.1016/0304-3959(96)02994-6

ROEDIGER, H. L., III, & PAYNE, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition*, **13**, 1-7.

ROEDIGER, H. L., III, SRINIVAS, K., & WADDILL, P. (1989). How much does guessing influence recall? Comment on Erdelyi, Finks, and Feigin-Pfau. *Journal of Experimental Psychology: General*, **118**, 255-257. doi:10.1037/0096-3445.118.3.255

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. doi:10.1016/j.jml.2004.03.004

SLOVIC, P., FINUCANE, M., PETERS, E., & MACGREGOR, D. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397-420). New York: Cambridge University Press.

STANKOV, L., & CRAWFORD, J. D. (1996). Confidence judgments in studies of individual differences. *Personality & Individual Differences*, **21**, 971-986. doi:10.1016/S0191-8869(96)00130-4

TONIDANDEL, S., QUIÑONES, M. A., & ADAMS, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, **87**, 320-332. doi:10.1037/0021-9010.87.2.320

WINKIELMAN, P., SCHWARTZ, N., & BELLI, R. F. (1998). The role of ease of retrieval and attribution in memory judgments: Judging your memory as worse despite recalling more events. *Psychological Science*, **9**, 124-126. doi:10.1111/1467-9280.00022

NOTES

1. There is some evidence that students achieve lower accuracy in performance estimates on selection (i.e., multiple-choice) tests than on production (i.e., recall) tests (see Koriat & Goldsmith, 1996, for a discussion of the distinction between the two), probably due to the familiarity of distractors (Bol & Hacker, 2001; Pressley, Ghatala, Woloshyn, & Pirie, 1990). Because we were interested in factors that cause bias to differ between conditions rather than bias per se, we used production tests of general knowledge in all three experiments in order to avoid inflating confidence through distractor familiarity. An additional goal of using this type of test is to extend research into test-driven factors of performance evaluation, which has mostly focused on multiple-choice tests (e.g., Gigerenzer, Hoffrage, & Kleinbölting, 1991; Higham, 2007), to a new domain.

2. Although the forced-report condition was designed so that participants could not proceed on to the next question without entering a response, in theory participants could have avoided making an attempt at producing a response by entering a nonsense text string into the response box. In fact, such responses (as well as variants of "don't know") accounted for only 3% of all answers in the forced-report condition. Thus, the report option manipulation had a large effect on the number of questions attempted, shifting it from 67.4% of questions in the free-report condition to 97.0% in the forced-report condition.

3. All the analyses reported in all three experiments were also carried out with block order as an additional variable; although this variable sometimes interacted with the independent variables of interest, indicating that one set of questions was more difficult to answer than the other, the inclusion of this variable in the analyses did not compromise any of the reported main effects or conclusions.

(Manuscript received June 23, 2009;
revision accepted for publication September 27, 2009.)