Mem Cogn
DOI 10.3758/s13421-012-0187-3

The effect of question order on evaluations of test performance: how does the bias evolve?

Yana Weinstein · Henry L. Roediger III

Abstract Weinstein and Roediger (Memory & Cognition 38:366–376, 2010) found that manipulating the order of questions on a general knowledge quiz resulted in differing evaluations of performance at the end of the quiz: Irrespective of their actual performance, participants were consistently more optimistic about their performance when questions were given in an easy-to-hard order. In the present experiment, the participants were stopped 10 times throughout a 100-item test and asked to evaluate their performance on the last 10 questions they had answered, as well as rating their impressions of the test so far and predicting their final performance. Arranging the questions from the easiest to the hardest produced more optimistic performance evaluations on each block than did an analogous hard–easy question order, even though performance on the two versions did not differ significantly as a function of question order. Furthermore, the ratings of item difficulty on each block of 10 questions were asymmetrical in the two conditions, with a higher sensitivity to increasing as compared to decreasing question difficulty. On the other hand, the item-by-item ratings and predictions remained unaffected by question order. Our findings are best explained by an anchoring interpretation, which suggests that students fail to adjust their evaluations of performance as the difficulty of the questions changes across the test.

Keywords Judgment · Memory · Metamemory

Weinstein and Roediger (2010) showed that when general knowledge questions on a cued recall test are arranged from the easiest to the most difficult, people tend to estimate their performance on the test more optimistically than when the same questions are arranged in a different order (either from the most difficult to the easiest or in a random order). This difference in evaluations of performance is observed despite the lack of a significant difference in performance as a function of question order. Previously, Dean (1973) found that ordering questions by difficulty affected judgments of test fairness, and other research has addressed the effects of ordering questions by topic (e.g., Pettijohn & Sacco, 2007). The impact of difficulty in question ordering on evaluations of performance has received little attention so far. An awareness of this bias could be crucial to students who have to make important decisions on the basis of evaluations of their performance on a test (e.g., when deciding whether to cancel their score on the Law School Admission Test).

Three broad candidate processes could account for the observed phenomenon that starting a test with easier questions leads to more optimistic evaluations of performance. The first explanation involves a reliance on the affect heuristic, whereby the difficulty of the questions at the beginning of a test sets the mood for how participants will later evaluate their performance (Slovic, Finucane, Peters, & MacGregor, 2002). Counter to this theory, Weinstein and Roediger (2010, Exp. 3) found that confidence ratings on individual questions were unaffected by the question order, whereas reliance on an affect heuristic should have elicited higher confidence ratings when easy questions appeared at the beginning of the test. One aim of the present study was to replicate the absence of a bias in item-by-item confidence ratings due to question order, thereby ruling out the affect heuristic explanation.

The primary aim of the experiment presented here, though, was to contrast two other explanations, which we cannot currently distinguish between by examining the data

Y. Weinstein (✉) · H. L. Roediger III
Department of Psychology, Washington University,
Box 1125, One Brookings Drive,
St. Louis, MO 63130, USA
e-mail: y.weinstein@wustl.edu

Published online: 24 February 2012

in Weinstein and Roediger (2010): anchoring and primacy. One possibility is that the difficulty of the questions at the beginning of a test sets an anchor that constrains participants in their evaluations of performance throughout the remainder of the test (Scheck, Meeter, & Nelson, 2004). This explanation is different from the affect heuristic explanation, in that it does not predict a bias in item-by-item confidence ratings. That is, according to the anchoring explanation, participants may form an impression of their performance on the test early on and may then fail to adjust this global impression as the difficulty of the questions changes, while still making unbiased evaluations of the likelihood of a correct individual response. The second explanation is that global evaluations of performance are affected by a primacy bias, whereby earlier questions are weighted more heavily than later questions in the overall evaluation of performance (Anderson & Barrios, 1961; Crano, 1977). The primacy explanation also makes no predictions about differences in item-by-item confidence ratings, but instead it refers to a greater weighting of initial items in evaluations of performance.

The anchoring and primacy explanations can both explain the bias in global postdictions demonstrated by Weinstein and Roediger (2010), so to distinguish between these two accounts we examined evaluations of performance on individual blocks of questions within tests that had different question orderings. The anchoring and primacy accounts make different predictions for this intermediate grain of performance judgments. According to the anchoring explanation (Scheck et al., 2004), we should expect to see more optimistic evaluations in block postdictions throughout the test when easy questions appear at the beginning. That is, because the first block of questions will produce performance near ceiling (10 out of 10 correct), participants should anchor to this level of performance and fail to adjust their estimates of performance on each block accordingly as the questions increase in difficulty. According to the primacy explanation, on the other hand (Anderson & Barrios, 1961; Crano, 1977), differences in optimism should only be apparent on the final evaluation of performance when the test as a whole is taken into account, because this judgment, unlike the block postdictions, can be influenced by the unequal weighting of different parts of the test.[1] That is, the extremely easy or difficult items at the beginning may be weighted more heavily in the global judgments.

A corollary question that we wanted to address with this experiment was how subjective perceptions of the test evolve over time. That is, when students are taking a test, at what point have they formed an impression of the test's difficulty level, and how is this impression updated as question difficulty gradually changes? In the experiment that we present, participants were asked a series of questions to evaluate their impressions of the test so far after each block of questions (in addition to making the postdictive performance judgment already mentioned above). Specifically, participants indicated how much they were enjoying the test and how difficult they were finding the test so far. If these questions are answered normatively, participants should take into account all questions answered up to that point. However, if participants do evaluate their cumulative experience of the test so far when they make the judgments, the subjective ratings should track cumulative accuracy. Cumulative accuracy will be steadily increasing from floor in the hard-to-easy ordering condition and steadily decreasing from ceiling in the easy-to-hard condition, eventually reaching the same value by the final block (because the overall difficulty of the test in the two ordering conditions was equivalent and normed to 50% accuracy). However, biases might affect these judgments, as discussed next.

One possibility, already discussed above with regard to the final postdictions, is that participants will weight the early blocks more heavily than subsequent blocks when making judgments that are supposed to reflect their experience of the test so far, and/or that they may anchor to their first impressions of the test and fail to adjust the ratings accordingly as the difficulty of the questions changes. If this is the case, by the end of the test participants in the easy–hard condition should still rate their enjoyment of the test as higher than should participants in the hard–easy condition, and they should also rate the test as less difficult. Anderson (1981) found anchoring/primacy effects in impression formation and suggested that they might be driven by a decline in attention from the beginning to the end of the list (see also Crano, 1977, and the attention decrement hypothesis).

On the other hand, it is also possible that participants might instead weight *more recent* blocks more heavily than early blocks in their judgments. This could occur because of a combination of recency effects (the tendency to recall more information from the end of an event; Glanzer & Cunitz, 1966) and the availability heuristic (Tversky & Kahneman, 1973), which states that people tend to make judgments based on the events that are easiest to recall.

A third possibility is that the two different question orders will have asymmetrical effects on the impressions of the test. That is, ratings in one question order condition might change more rapidly as a function of time as the test goes on. The literature on the negativity bias (e.g., Richey, McClelland, & Shimkunas, 1967; Wason, 1959) suggests that negative information has more of an impact on judgments than does equivalently positive information.

[1] Note that, due to the large number of questions on the test and to random noise in accuracy levels, questions within a block did not differ sufficiently in difficulty to produce within-block primacy effects. This issue is addressed more thoroughly in the Results; see, in particular, Fig. 1b.

Consequently, participants in the easy–hard question condition might be influenced more strongly by the questions getting harder than the participants in the hard–easy condition are influenced by the questions getting easier. As a result, participants in the hard–easy condition might show more anchoring and adjust their ratings more slowly than participants in the easy–hard condition.

Finally, in addition to making postdictions and subjective ratings of test difficulty and enjoyment, participants were asked to predict their performance on the test as a whole after each block of 10 questions. This required participants to project into the future and make a prediction of how the difficulty of the test would evolve. This measure allowed us to determine the extent to which participants were sensitive to the trend of increasing or decreasing question difficulty, as well as the extent to which they expected this trend to continue. If participants were perfectly able to extract the trend, their predictions should tend toward estimates of 50% accuracy.

Method

Participants

A group of 50 Washington University undergraduates 18 to 22 years of age (M = 20.04, SD = 1.13; 27 females, 23 males) participated in the study; they either were assigned credit for fulfilling a course requirement or were financially reimbursed for their time at the rate of $10/h. Of the participants, 25 were randomly assigned to the easy–hard ordering condition, and the other 25 were assigned to the hard–easy condition.

Design

We used a between-subjects design, with question order (easy–hard/hard–easy) as the only manipulated variable. The following dependent measures were analyzed, always by being compared between the easy–hard and hard–easy question order conditions: confidence ratings for each answered question; actual performance on each block of 10 questions (× 10 blocks); postdictions of performance on each block of 10 questions (× 10 blocks); ratings of enjoyment and difficulty made after each block of 10 questions (× 10 blocks); predictions of performance on the test as a whole made after each block of 10 questions (× 10 blocks); and global postdictions of performance made after the test.

Materials

The 100 general knowledge questions were taken from Weinstein and Roediger (2010, Exp. 1), where two sets of 50 questions had been extracted from the Nelson and Narens (1980) norms. For the present experiment, an updated version of the norms was used: The mean accuracy on each question was computed across all 80 participants in Weinstein and Roediger's Experiment 1 (regardless of their question order conditions). The mean accuracies for the two sets of norms (old and new) were highly positively correlated (r = .81, p < .001); however, the new norms were used in the experiment presented here because they were collected from the same population that was drawn from for the present study. The 100 questions ranged from extremely easy to answer (.99 probability of a correct response in the present population; an example of such a question was "What is the last name of the author who wrote *Romeo and Juliet*?") to those that were the most difficult to answer (0 probability of a correct response; i.e., not a single participant of the 80 in the original study could answer this question: "What is the name of the villainous people who lived underground in H. G. Wells's book *The Time Machine*?"). The mean performance according to the norms, across the whole set of 100 questions, was 50%.

Procedure

Question order was manipulated between subjects, with all 100 questions ordered either from the easiest to the hardest, or vice versa, for each participant. The procedure was taken directly from Weinstein and Roediger (2010), except that the test was lengthened (we used a single set of 100 questions per participant rather than only 50) and the participants were stopped after every block of 10 questions to make judgments.

Participants were told that they would be answering a set of 100 general knowledge questions and also evaluating their performance a few times throughout the test (they were not told of the exact placement of the prompts). The participants could skip questions that they did not know the answer to. When presented with a question, participants could either type an answer and press Enter, to make a confidence rating and continue to the next question, or press Enter without typing an answer, to skip the question (skipped questions did not require confidence ratings). The question order manipulation was not mentioned in the instructions.

Participants made judgments after every answered question, after every 10 questions, and after all 100 questions. After every question that they chose to answer, participants made a confidence rating on a scale from 0 to 100 by moving a slider with the mouse. The slider was labeled *0* and *100* at the extremes, but it did not display additional numeric information when participants moved the slider. The slider was recentered at 50 after every question, and it had to be moved after each answered question before participants were allowed to proceed to the next question. After every block of 10 general knowledge questions, participants answered four questions about their experience of the test. For the first question, participants picked a number from 1 to 10 to indicate how many of the last 10 questions they thought that they had gotten correct (0 was not included as an option in either question order condition, due to experimenter error). Participants then rated their enjoyment of the test ("How much are you enjoying this test?") and the difficulty of the test ("How difficult are you finding the test?") using sliders, with the extremes labeled either *not at all* and *very much* or *extremely easy* and *extremely difficult*, respectively for the two questions. Finally, participants were also asked to predict what their overall performance on the test would be once they had answered all of the questions, on a slider labeled from *0%* to *100%*. Note that while participants had been informed of the total number of questions on the test, they were not told how many questions they had completed so far at any point (other than at the end of the test). The participants answered the set of four evaluation questions 10 times, including after the final block (i.e., after general knowledge Questions 91–100). Immediately following this last block and the associated questions, the participants were told that they had answered all 100 questions on the test, and they were asked to indicate how many questions in total they thought that they had gotten correct. The participants entered their responses numerically into a text box.

Results

Performance (see Fig. 1a and b) is presented as a function of question difficulty, where each data point represents the same question(s) for every participant, regardless of the question order. Similarly, item-by-item confidence ratings and block postdictions (Fig. 2a and b) are presented as a function of block difficulty, where each data point represents the same block of 10 questions for every participant, regardless of the question order. This means that for the hard–easy condition, the data are not presented in the order in which participants experienced the blocks. These data were subjected to mixed 2 × 10 ANOVAs with question order (easy–hard/hard–easy) as the between-subjects variable and block difficulty (1–10) as the within-subjects variable.

Ratings of enjoyment and difficulty (Fig. 3a and b), as well as the predictions about overall performance that the participants made after each block (Fig. 3c), are presented as a function of block order, starting from the earliest block (which would be the easiest block for the easy–hard condition and the hardest for the hard–easy condition) and ending with the last block of questions that the participants answered, to demonstrate the evolution of these ratings. These data were subjected to mixed 2 × 10 ANOVAs with question order (easy–hard/hard–easy) as the between-subjects variable and block position (1–10) as the within-subjects variable. We also compared the global postdictions between conditions and correlated actual performance with the evaluations of performance.

In all cases, the comparisons are significant at the $p < .05$ level unless stated otherwise. Pairwise comparisons were performed only when the ANOVA identified a significant interaction between question order and block difficulty/position. In order to alert the reader to the independent variable (block difficulty or block position), we have used two different formats when plotting the figures. When the results are presented as a function of block difficulty, we have used line graphs; when the results are presented as a function of block position, bar graphs are used instead.

Performance

The lenient criterion described in Weinstein and Roediger (2010) was used to score all of the responses: Misspelled answers were scored as correct. Figure 1a shows performance on all 100 questions as a function of question order, with vertical lines delineating each block. In addition, Fig. 1b demonstrates that within any block of 10 questions, there was little relationship between serial position and accuracy, making it impossible for participants to extract a difficulty gradient within an individual block. Any bias in the evaluations of performance by blocks, then, can be assumed to arise from the cumulative experience of the test so far and not from anchoring or attention effects within an individual block.

For the analysis, accuracy was binned across each block of 10 questions and subjected to mixed 2 × 10 ANOVAs with question order (easy–hard/hard–easy) as the between-subjects variable and block difficulty (1–10) as the within-subjects variable. As expected, performance differed as a function of block difficulty, $F(6.32, 303.5) = 215.1$, $\eta_p^2 = .82$ (Greenhouse–Geisser corrected). Averaged across all blocks and items, the mean performance in the easy–hard condition was 53.3% ($SD = 15.3$), and in the hard–easy condition the mean was 49.6% ($SD = 13.0$). However, as in Weinstein and Roediger (2010), this numerical difference was not significant ($p = .37$), nor was there a Block Difficulty × Question Order interaction ($p = .89$).

Item-by-item confidence ratings

Figure 2a presents the mean item-by-item confidence ratings as a function of block difficulty and question order. The confidence ratings made for every answered question in each block were averaged to produce one value for each block. For unanswered questions, a confidence rating of

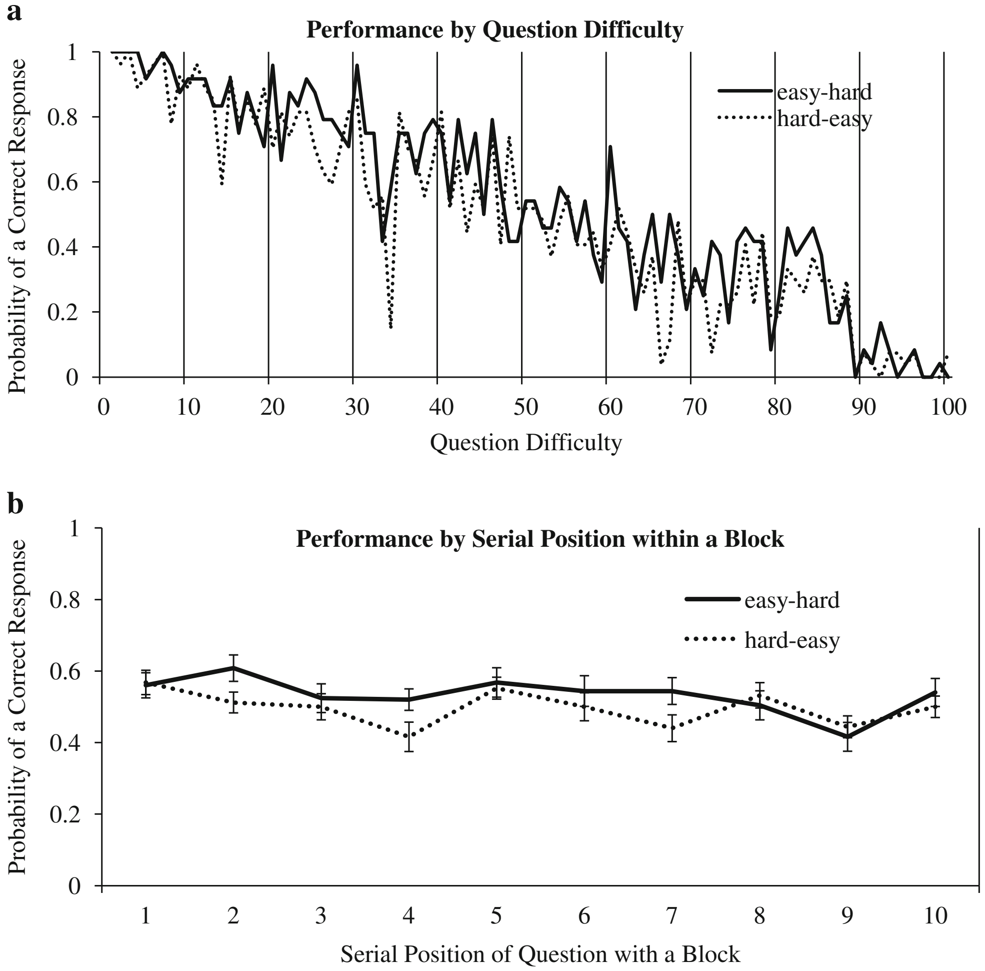


Fig. 1 **a** Probabilities of correct responses on each question, as a function of question order. **b** Probabilities of correct responses at each serial position within a block. To obtain these data, performance at each of the 10 serial order positions was averaged across all 10 blocks. Error bars in this and all subsequent figures represent *SEM*s

zero was assigned. Not surprisingly, item-by-item confidence ratings tracked accuracy, with a significant effect of block difficulty, $F(5.52, 265.1) = 262.7$, $\eta_p^2 = .85$ (Greenhouse–Geisser). More importantly, there was no hint of an effect of question order on the item-by-item confidence ratings, nor was there an interaction between question order and block difficulty (smallest $p = .72$).

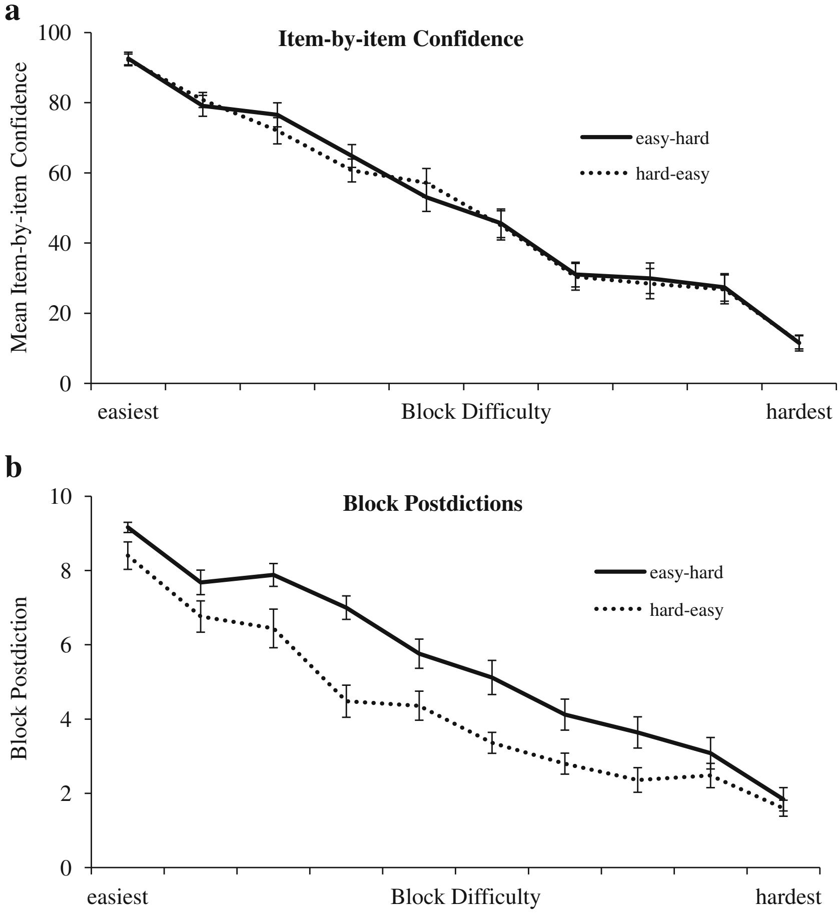
Block postdictions

Figure 2b presents postdictions of performance (i.e., how many of the previous 10 items the participants judged they had gotten correct) as a function of block difficulty and question order. Naturally, block postdictions decreased as question difficulty increased, $F(4.62, 221.9) = 163.8$, $\eta_p^2 = .77$ (Greenhouse–Geisser). Contrary to performance and the item-by-item confidence ratings, though, the block postdictions, matched with block difficulty, did differ as a function of question order: Participants in the easy–hard condition were more optimistic than participants in the hard–easy condition about their performance on the same questions, $F(1, 48) = 9.88$, $\eta_p^2 = .17$. There was also an interaction between block difficulty and question order, $F(4.62, 221.9) = 3.24$, $\eta_p^2 = .06$ (Greenhouse–Geisser). From Fig. 2b, it is clear that participants in the easy–hard condition were more optimistic about their performance on every block, but that this difference was diminished when performance was at ceiling (easiest blocks) or at floor (hardest blocks), producing the interaction. We chose three time points for follow-up comparisons, here and on all subsequent measures: Blocks 1, 5, and 10, with significance levels adjusted to $p = .017$ for multiple comparisons. Of those three blocks, only Block 5 showed a significant difference in the postdictions between question order conditions, $t(48) = 2.53$. The significant difference on Block 5 was replicated after adjusting evaluations of performance for actual performance, $t(48) = 3.49$.

Enjoyment and difficulty ratings

Questions that referred to the experience of the test so far are also presented temporally to show the evolution of the participants' impressions of the test over time. Figure 3a and b,

Fig. 2 **a** Mean item-by-item confidence ratings on each of the 10 blocks of 10 questions, as a function of question order. **b** Block postdictions (i.e., the number of questions participants thought that they had answered correctly on each of the 10 blocks of 10 questions) as a function of question order



respectively, show ratings of enjoyment and difficulty as a function of time and question order condition. Both enjoyment [$F(2.81, 134.7) = 22.4$, $\eta_p^2 = .32$, Greenhouse–Geisser] and difficulty [$F(3.24, 155.3) = 84.8$, $\eta_p^2 = .64$, Greenhouse–Geisser] showed the expected Time × Question Order interaction: Participants in the easy–hard condition gradually enjoyed the test less and less and found it more difficult, whereas participants in the hard–easy condition gradually enjoyed the test more and more and found it easier.

For the enjoyment ratings (Fig. 3a), neither time ($p = .34$) nor question order ($p = .35$) produced significant overall main effects, indicating that enjoyment of the test was symmetrical for the two conditions (higher when questions were easy and lower when questions were hard). Follow-up *t* tests showed that participants in the easy–hard condition were enjoying the test more than participants in the hard–easy condition on the first block, $t(48) = 4.00$; the two groups were enjoying it equally on the fifth block ($p = .22$); and participants in the easy-hard condition were enjoying it less than participants in the hard–easy condition by the last block, $t(48) = –2.34$.

For difficulty ratings (Fig. 3b), on the other hand, both time [$F(3.24, 155.3) = 7.21$, $\eta_p^2 = .13$, Greenhouse–Geisser] and question order [$F(1, 48) = 25.5$, $\eta_p^2 = .33$] produced significant main effects. Participants in the hard–easy condition rated the test as, on average, more difficult (mean cumulative difficulty = 73) than did participants in the easy–hard condition (mean cumulative difficulty = 49). Looking at the follow-up *t* tests for Blocks 1, 5, and 10, we see that participants in the hard–easy condition rated the test as significantly more difficult on both the first block [$t(48) = –15.7$] and the fifth block [$t(48) = –5.03$], but participants in the easy–hard condition rated the test as more difficult on the final block [$t(48) = 3.07$].

Predictions

Figure 3c presents predictions (made on a percentage scale from 0% to 100%) of overall performance on the whole test that were made after each block of questions, as a function of time and question order. There was a main effect of question order [$F(1, 48) = 49.3$, $\eta_p^2 = .51$] and an interaction between time and question order [$F(2.48, 118.9) = 62.4$, $\eta_p^2 = .57$, Greenhouse–Geisser], but no main effect of time ($p = .23$, Greenhouse–Geisser) on predictions. That is, participants in the easy–hard condition consistently predicted higher overall performance throughout most of the test than did participants in the hard–easy condition. This was the

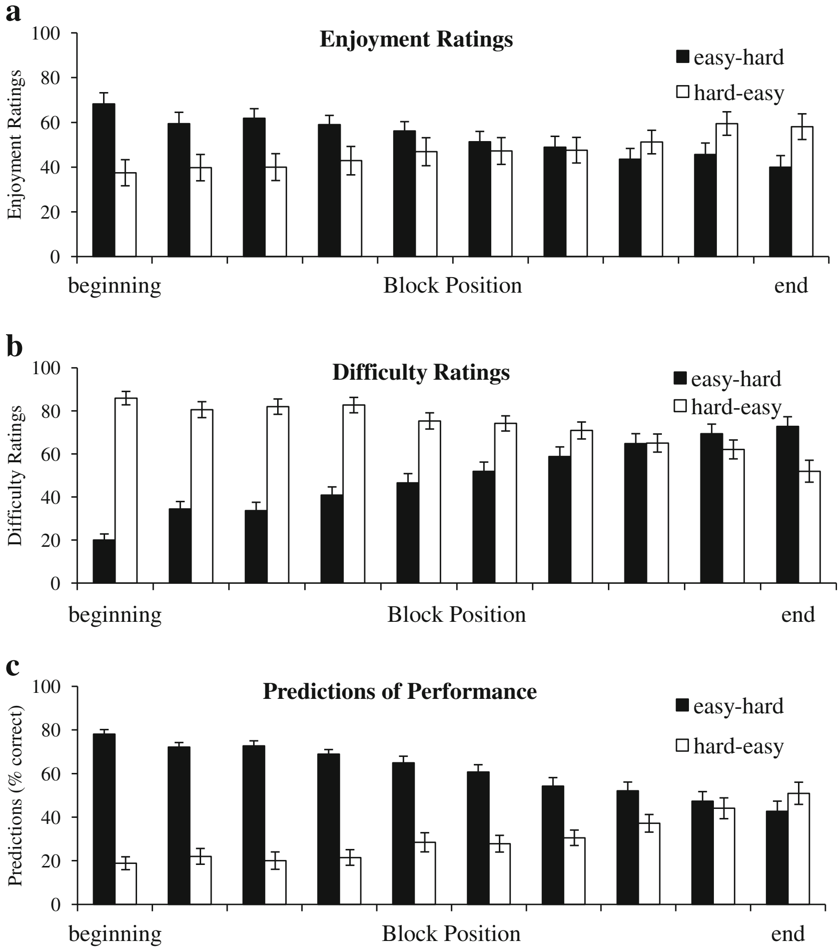


Fig. 3 **a** Ratings of cumulative enjoyment of the test so far, made after each block of 10 questions, as a function of time and question order condition. **b** Ratings of cumulative difficulty of the test so far, made after each block of 10 questions, as a function of time and question order condition. **c** Predictions of performance on the whole test made after each block of 10 questions, as a function of time and question order condition

case on both the first block [$t(48) = 16.6$] and the fifth block [$t(48) = 6.82$], but not on Block 10, where the predictions were not statistically different ($p = .24$). The predictions that participants made were thus in line with how they had performed on the test so far, since cumulative accuracy would follow the same pattern (initially low for the hard–easy condition and high for the easy–hard condition, meeting at 50% accuracy by the final block).

Final postdiction

The final estimate of the number of questions answered correctly was affected by question order: Participants in the easy–hard condition reported that they got $M = 55.5$ ($SD = 19.4$) questions correct, whereas participants in the hard–easy condition reported that they correctly answered only $M = 42.2$ ($SD = 17.3$) questions, $F(1, 48) = 6.50$, $\eta_p^2 = .12$. The same analysis performed on bias (i.e., final postdictions adjusted for actual performance) also revealed a significant difference between conditions, $F(1, 48) = 6.69$, $\eta_p^2 = .12$.

The final postdictions were also compared to the total numbers of questions that participants had estimated they answered correctly across the 10 individual block postdictions, by summing these 10 estimates. Comparing these judgments in a 2 × 2 mixed ANOVA with question order as the between-subjects variable and judgment (final postdiction or summed-block postdictions) as the within-subjects variable, there was only a main effect of question order, $F(1, 48) = 8.66$, $\eta_p^2 = .15$, and no effect of judgment nor any interaction between the two (ps > .74).

Correlations

Table 1 presents the correlation matrix between performance, item-by-item confidence ratings, block postdictions (summed across all 10 blocks), block predictions (summed across all 10 blocks), and final postdictions. Note that these correlations were performed across data from both the easy–hard and hard–easy ordering conditions. To account for multiple comparisons (a total of 10), the significance criterion was adjusted to $p < .005$. As can be seen from the table, all judgments except item-by-item confidence ratings were correlated with each other and with actual performance. Item-by-item confidence ratings were not significantly

Table 1 Pearson correlation matrix among performance and performance evaluations

	Performance	Item-by-Item Confidence	Summed Postdictions	Summed Predictions
Item-by-item confidence	.16			
Summed postdictions	.84*	.26		
Summed predictions	.47*	.35	.74*	
Final postdiction	.70*	.21	.84*	.60*

* $p < 0.005$

correlated with any of the other judgments. Note that here, item-by-item confidence ratings were only considered for answered questions, rather than replacing unanswered questions with a zero confidence rating, as doing so would necessarily create a correlation between confidence and accuracy. The best predictor of actual performance was the sum of the block postdictions across Blocks 1–10.

Discussion

The experiment we reported was designed to distinguish between three possible explanations of bias in performance evaluations due to question order: affect, anchoring, and primacy. The affect heuristic (Slovic et al., 2002) was an unlikely candidate explanation, because it predicts an effect of question order on evaluations of performance made at all grain sizes, including at the item-by-item level. In the present experiment, we replicated Weinstein and Roediger (2010, Exp. 3) with respect to item-by-item confidence ratings: We did not find any difference in item-by-item confidence ratings when comparing the easy–hard and hard–easy question orders.

Ruling out the affect heuristic explanation, we now turn to the two other proposed theories. The critical result that distinguishes between the anchoring and primacy explanations is the observed bias in postdictions at the block level. Every block produced higher performance evaluations in the easy–hard question order condition than in the hard–easy condition, most notably in the middle blocks, where performance was not at the extremes. In addition, after correcting block postdictions for actual performance, by Block 5 participants in the easy–hard condition were already evaluating their performance on each individual block more optimistically than participants in the hard–easy condition. This pattern of results was predicted by the anchoring explanation (Scheck et al., 2004), but not by the primacy explanation, which only predicted bias on the final, global postdiction (Anderson & Barrios, 1961; Crano, 1977).

Why would anchoring affect block postdictions but not item-by-item confidence ratings? According to Gigerenzer, Hoffrage, and Kleinbölting (1991; see also Sniezek, Paese, & Switzer, 1990, for a similar idea), people use different probabilistic information to make item-by-item and global judgments: In the case of item-by-item confidence, the content of the question itself is used to make the judgment, whereas in the case of global postdictions, other information, such as overall task difficulty and the subjective experience of the test as a whole, feeds into the decision. The Gigerenzer et al. model thus predicts that when making item-by-item confidence judgments, participants are able to ignore question order and focus on the content of each individual question, independent of its position in the test.

The dissociation between final postdictions and item-by-item confidence ratings was supported by a lack of correlation between the two measures. This lack of correlation between local and global evaluations of performance was reported by Stankov and Crawford (1996), although they contrasted their results with those of Schraw (1994), who reported relatively high correlations. The key difference between these two studies was that the "local" evaluations of performance examined by Schraw actually required participants to estimate performance across four questions, whereas Stankov and Crawford used true item-by-item measures. In that sense, our results are consistent with both Schraw (i.e., we found a correlation between global and block postdictions) and Stankov and Crawford (i.e., we did not find a correlation between global postdictions and item-by-item confidence ratings).

A novel aspect of the present experiment is the measures that we introduced to evaluate online impressions of the test: enjoyment ratings, difficulty ratings, and predictions of overall performance that required participants to anticipate the difficulty of the whole test after each block. These measures allowed us to examine whether increasing and decreasing question difficulty resulted in symmetrical impressions of the test. Whereas enjoyment of the test and predictions of overall performance did not show any asymmetry between the two question order conditions (i.e., enjoyment and performance predictions both increased steadily in the hard–easy condition and decreased steadily in the easy–hard condition), the difficulty ratings were asymmetrical, such that participants in the hard–easy condition found the test to be, on average, more difficult than did participants in the easy–hard condition. This was a result of participants in the hard–easy condition showing less sensitivity to a decrease in question difficulty, as compared with the analogous sensitivity of participants in the easy–hard condition to the increase in question difficulty that they experienced. The difficulty ratings made on each of the 10 blocks were thus made within a narrower range in the hard–easy condition than in the easy–hard condition. This is the first result from our question order paradigm suggesting that perceptions of a test may be affected to different extents when participants experience an increase versus a decrease in difficulty across the test. This observation is in line with findings that negative information has a stronger influence on judgments than does positive information (Richey et al., 1967; Wason, 1959). Interestingly, predictions of performance did not seem to be affected by any kind of bias and were relatively normative, tracking cumulative accuracy and tending toward the actual accuracy of 50% in both question order conditions.

In sum, the results of the present experiment shed light on the finding we reported in Weinstein and Roediger (2010) that arranging questions on a test from the easiest to the hardest produces more optimistic estimates of performance than do other question orders. Contrary to the idea that bias arises due to a retrospective memory bias (i.e., primacy), the results of the experiment reported here suggest that bias is driven by anchoring, which results in consistently optimistic evaluations of performance throughout a test that begins with easier questions, at least when participants are asked to estimate performance across a number of questions. This conclusion can be bolstered further by the difficulty ratings we collected on every block: Whereas enjoyment ratings were fairly symmetrical and did not differ in the middle blocks, difficulty ratings still showed a divergence between question order conditions by the time that participants reached the middle of the test, revealing a tendency to anchor difficulty to the beginning of the test, especially when the test began with difficult questions.

It would be helpful for students and instructors to be aware of the possible biasing effect of question order. For instance, a practice test that consists of questions ordered with increasing difficulty could give students an overly optimistic impression of their level of knowledge. Even though question order does not seem to significantly affect performance, the finding that it changes evaluations of performance is important in and of itself. In fact, much effort is being dedicated to the creation of interventions designed to de-bias student evaluations of performance (e.g., Nietfeld, Cao, & Osborne, 2005), not least because accurate self-monitoring can lead to better performance in the future (Thiede, Anderson, & Therriault, 2003). Examining the processes that lead to biases, as we do here, is a necessary step toward this aim.

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 Kelsey Schwantes for assistance with the data collection.

References

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

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

Crano, W. D. (1977). Primacy versus recency in retention of information and opinion change. *Journal of Social Psychology, 101,* 87–96.

Dean, M. L. (1973). The impact of exam question order effects on student evaluations. *Journal of Psychology, 85,* 245–248.

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

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5,* 351–360. doi:10.1016/S0022-5371(66)80044-0

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 and Verbal Behavior, 19,* 338–368. doi:10.1016/S0022-5371(80)90266-2

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the classroom. *Journal of Experimental Education, 74,* 7–28.

Pettijohn, T. F., II, & Sacco, M. F. (2007). Multiple-choice exam question order influences on student performance, completion time, and perceptions. *Journal of Instructional Psychology, 34,* 142–149.

Richey, M. H., McClelland, L., & Shimkunas, A. (1967). Relative influence of positive and negative information. *Journal of Social Psychology, 97,* 233–241.

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51,* 71–79. doi:10.1016/j.jml.2004.03.004

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19,* 143–154.

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.

Sniezek, J. A., Paese, P. W., & Switzer, F. S., III. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes, 46,* 264–282. doi:10.1016/0749-5978(90)90032-5

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences, 21,* 971–986.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95,* 66–73.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232. doi:10.1016/0010-0285(73)90033-9

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology, 11,* 92–107. doi:10.1080/17470215908416296

Weinstein, Y., & Roediger, H. L., III. (2010). Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it. *Memory & Cognition, 38,* 366–376. doi:10.3758/MC.38.3.366