Mem Cogn (2014) 42:1325–1332
DOI 10.3758/s13421-014-0439-5

Impression formation of tests: retrospective judgments of performance are higher when easier questions come first

Abigail Jackson · Robert L. Greene

Published online: 28 June 2014

Abstract Four experiments are reported on the importance of retrospective judgments of performance (postdictions) on tests. Participants answered general knowledge questions and estimated how many questions they answered correctly. They gave higher postdictions when easy questions preceded difficult questions. This was true when time to answer each question was equalized and constrained, when participants were instructed not to write answers, and when questions were presented in a multiple-choice format. Results are consistent with the notion that first impressions predominate in overall perception of test difficulty.

Keywords Impression formation · Metacognition · Primacy effect · Testing

First impressions can be particularly influential in shaping our view of a person, a notion common enough to have its own aphorisms: "First impressions are lasting impressions" and "You only get one chance to make a first impression." Indeed, Jane Austen's early version of *Pride and Prejudice* was entitled *First Impressions* (Fergus, 1997); a central theme in the novel revolves around characters making initial judgments about each other, some warranted and some unwarranted, but all difficult to change. The importance of early impressions in person evaluation has been a common theme in social–psychological research on person perception since Solomon Asch's influential work in the 1940s. Asch (1946) reasoned that we form impressions of people globally. In one of a series of experiments, Asch looked specifically at order effects in impression formation. Participants were read a list of traits, both favorable and unfavorable, describing a hypothetical person. Asch left the content of the list the same, but manipulated the order of the traits between participants. After hearing the list, participants were asked about their impressions of the person. Participants who had heard the list with the positive traits first rated the person more favorably than did participants who had heard the negative traits first. Forgas (2011) notes: "The disproportionate influence of first impressions is one of the most robust and reliable effects distorting such [impression formation] judgments (Asch, 1946; Crano, 1977)" (p. 427). Researchers in the decades since Asch's research have built on his work by studying the impact of a variety of different variables, such as affect (Forgas, 2011), mental fatigue (Webster, Richter, & Kruglanski, 1996), and need for cognition (Ahlering & Parker, 1989), on primacy in impression formation.

The importance of first impressions presumably does not apply only to people, but to situations as well. For example, impressions of cognitive tasks should play an important role in metamemory—that is, knowledge about memory and attempts to monitor and control learning and retrieval (Dunlosky & Thiede, 2013). Much research on metamemory has focused on prospective monitoring, those judgments pertaining to future performance. However, an important component of metacognitive judgment is retrospective, that is, for example, an evaluation of how well one has performed on a test, in the absence of feedback. Retrospective judgments may play a role in students' decisions about whether to drop a class or to cancel a standardized test score. Such retrospective evaluations should be based in part on impressions of the overall difficulty of a test. The focus of this article is the influence of the order of test questions, in terms of difficulty, on self-evaluations of performance. Questions on a test can be arranged in different ways, and in many cases are not arranged randomly. Some tests, such as the computer-adaptive testing currently used for the GRE, may begin with a question of medium difficulty followed by harder or easier questions,

A. Jackson (✉) · R. L. Greene
Department of Psychological Sciences, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA
e-mail: abigail.jackson@case.edu

depending on the test taker's performance. For paper-and-pencil tests in school settings, some institutions, such as the Center for Teaching Excellence at Cornell University, give instructors the advice to present easier questions at the beginning of a test (Devine & Yaghlian, n.d.). This may be an attempt to prevent students from becoming discouraged early in the test, or to prevent students from spending too much time on more difficult questions while neglecting easier questions.

Weinstein and Roediger (2010) showed that question order can influence retrospective judgments of test performance: On tests containing a mixture of easy and hard questions, participants gave higher estimates of performance on tests that began with easy questions than they did on tests that began with difficult questions or on tests that had a random arrangement of questions. Bias occurred only after participants had taken the entire test, and not on an item-by-item basis. Although retrospective judgments were higher for the tests beginning with the easier questions, actual performance on the tests did not significantly differ. Weinstein and Roediger (2012) replicated their previous findings of bias in postdictions made at the global level but not at the item level, and in addition found bias in postdictions made after blocks of every 10 questions. (Participants made confidence ratings after each individual question for that one question, after every 10 questions for those 10 questions, and after the test for the entire test.)

In their 2010 article, Weinstein and Roediger ruled out an affect heuristic behind the retrospective bias, whereby the difficulty of the questions at the beginning of the test would set the mood for later postdictions. An affect heuristic predicts that item-by-item ratings will be influenced by overall test list structure, but they found no such bias at the item level. With their 2012 article, Weinstein and Roediger were able to rule out a primacy heuristic as the source of bias, whereby the questions at the beginning of the test would be given more weight in overall performance evaluation. A primacy heuristic predicts bias on only the final, global postdictions and not on the judgments made after each block. Because bias was found at the block level (after every group of 10 questions), Weinstein and Roediger (2012) concluded that their results support an anchoring heuristic, whereby the difficulty level of early questions casts an anchor that constrains performance evaluations throughout the test.

Four experiments are reported here that replicate and extend the discovery by Weinstein and Roediger (2010, 2012) that retrospective judgments of test performance are influenced by question order. Perhaps the most significant difference between our design and the original is in test format. The tasks used by Weinstein and Roediger were fully computerized, with participants viewing test questions one at a time. Although this procedure provides the experimenter a great amount of control over the testing situation, many testing situations in the real world are not conducted in such a manner. Rather, many tests in classrooms are given in paper-and-pencil format; students are responsible for their own pacing and attention throughout the test. Our first experiment is a replication of Weinstein and Roediger (2010), specifically, their second experiment, which successfully extends their computer procedure to paper-and-pencil testing. Experiments 2, 3, and 4 are similar in design to our first, but with changes made to the procedure in an attempt to further investigate the processes responsible for the retrospective bias phenomenon. Experiment 2 included an instruction for participants to spend an equal and limited amount of time on each question, in an attempt to rule out the possibility that the bias in our testing situation was caused by participants reflecting on earlier questions for a longer time than they spent on questions presented later in the test. For Experiment 3, participants were instructed to refrain from actually providing written answers to the questions. This was done to discern the difference, if any, between participants forming an overall impression of the test and participants remembering what they physically wrote on the test. Experiment 4 used the same questions as the preceding experiments but presented questions in a multiple-choice format to determine if results could be extended beyond a free-response design.

Experiment 1

The experiments shared a similar general design. Experiment 1 was designed to replicate the findings of Weinstein and Roediger (2010), specifically, their second experiment, on retrospective test bias, whereby presenting easy items at the beginning of a test increased participants' estimations of their performance. A few changes were made to Weinstein and Roediger's original design in its translation to our location, but the general design remained intact. Weinstein and Roediger tested participants individually or in small groups and employed computerized testing. We tested participants concurrently, using paper tests, a method and format common in many classroom testing situations. Participants were given three sets of general knowledge questions, varying in difficulty, which were taken from the same normed list (Nelson & Narens, 1980) used by Weinstein and Roediger.[1] The order of difficulty of the questions was manipulated to create three conditions: tests with questions that were ordered from easiest

[1] An updated normed list based on the Nelson and Narens questions has since been published (Tauber, Dunlosky, Rawson, Rhodes, & Sitzman, 2013), but was not yet available at the time of the present research. The exact probabilities of recall have changed somewhat between the two lists. However, since we are interested in the relative ease and difficulty of the questions, which have not changed dramatically between the lists, we have no reason to suspect that conclusions based on the original norms are invalid.

to most difficult, questions that were ordered from most difficult to easiest, and a random-order condition. Each participant was exposed to all three conditions. At the end of each set, participants were asked to estimate their performance on that block of questions. The manipulation of question order was intended to have an impact on participants' evaluations of their performance without having an impact on performance itself.

Method

Participants Thirty-three Case Western Reserve University students from introductory psychology classes participated in the study, in partial fulfillment of a course requirement.

Design The study used a within-subjects design, with test list structure (easy to hard, random, and hard to easy) as the manipulated variable. We analyzed performance, postdictions, and bias in postdictions.

Materials Three sets of 50 general knowledge questions were selected from the Nelson and Narens (1980) norms. The 75 easiest and 75 most difficult questions were selected. Each question could be answered in one word. One of the easy questions and one of the hard questions were eliminated because they were no longer accurate; they were replaced with the next easiest and next hardest questions, respectively. An example of an easy question is, "What is the capital of France?" An example of a hard question is, "What is the last name of the 21st U.S. president?" (The answers are *Paris* and *Arthur*, respectively.) Questions were given in the form of paper tests, with participants writing their answers below each question.

Procedure Participants were tested as one large group. They were given three different tests of 50 questions each. Test list structure was manipulated within participants so that each participant was exposed to all three of the different question orders (easy to hard, random, and hard to easy). The manipulation of test list structure was not made explicit to participants. Questions were assigned to a particular trial so that all tests given in each trial contained the same questions, with only the test list structure manipulated. Presentation order was counterbalanced across participants such that participants were divided into three groups (n = 11 per group), each receiving a different test list structure on each of the three trials. Participants were given 10 min to complete each test. They were told that the aim was to provide as many correct responses as possible. Because report option (free vs. forced) was previously found by Weinstein and Roediger (2010) to have no significant effect on bias in postdictions, participants were instructed that they could either make a guess or not respond to questions for which they were not sure of the answer. They were asked not to consult anyone else or any devices for the answers. After each test, participants were asked to estimate how many questions out of the previous 50 they believed they had answered correctly, and to write this number on the back of the test. After the participants finished writing their estimates, the tests were collected and the next trial began.

Results

A repeated-measures analysis of variance (ANOVA) was used, with test list structure (easy to hard, random, and hard to easy) as the within-subjects variable. There were three dependent measures: postdictions (i.e., retrospective estimates of number of questions answered correctly), performance (number of questions answered correctly), and bias in postdictions (see Table 1). All tests used a .05 significance criterion unless otherwise noted.

Postdictions Mean postdictions—estimated number of answers correct out of 50—are presented in Table 1. As in Weinstein and Roediger's (2010) experiment, test list structure had a significant effect on postdictions [$F(2,64) = 5.86$, $MSe = 17.68$, $\eta_p^2 = .16$]. Pairwise comparisons revealed that the easy-to-hard condition elicited significantly higher postdictions than both the random condition [$t(32) = 2.56$] and the hard-to-easy condition [$t(32) = 3.13$]. There was no significant difference between the random and hard-to-easy conditions [$t(32) = 0.26$].

Performance Mean performance, in terms of number of answers correct out of 50, is presented in the middle column of Table 1. One point was given for each answer identical to the correct answer. One point was also given for answers that were misspelled, had transposed letters, or had grammatical errors, but were otherwise correct. Participants answered approximately 31 % of questions correctly under each test list structure. No significant difference was found between conditions [$F(2, 64) = 0.19$, $MSe = 5.89$, $\eta_p^2 = .01$].

Bias in postdictions Bias in postdictions was calculated by subtracting performance from postdictions for each participant, giving a measure of optimism or pessimism. Mean bias data are shown in the far right column of Table 1 for each test

Table 1 Mean postdictions, performance, and bias by test list structure in Experiment 1

	Postdictions		Performance		Bias (Difference)	
Test List Structure	*M*	*SE*	*M*	*SE*	*M*	*SE*
Easy to hard	20.03	1.34	15.67	1.08	4.36	0.71
Random	17.09	1.18	15.30	1.00	1.79	0.76
Hard to easy	16.85	1.13	15.42	1.03	1.42	0.65

list structure condition. One observation worth noting immediately is that participants tended to be optimistic, with postdictions significantly exceeding performance in all three conditions: easy to hard [$t(32) = 6.15$], random [$t(32) = 2.34$], and hard to easy [$t(32) = 2.18$].

Test list structure had a significant effect on bias in postdictions [$F(2, 64) = 6.18$, $MSe = 13.72$, $\eta_p^2 = .16$]. Participants were more biased in the easy-to-hard condition than they were in the random condition [$t(32) = 2.40$] or in the hard-to-easy condition [$t(32) = 3.19$]. No significant difference in bias was found between the random and hard-to-easy conditions [$t(32) = 0.52$], though the difference is in the predicted direction: Bias is highest for the easy-to-hard condition and lowest in the hard-to-easy condition, with random falling in the middle, but much closer to hard-to-easy.

In addition, we calculated the absolute (unsigned) difference between postdictions and performance, a variable included in Weinstein and Roediger (2010) to reflect the amount of error in evaluations of performance without regard to the direction (optimism or pessimism) of the bias. There was a significant effect of test list structure [$F(2, 64) = 4.51$, $MSe = 6.29$] on absolute error in postdictions. The only significant pairwise comparison was between the easy-to-hard and hard-to-easy conditions [$t(32) = 2.82$], with greater absolute error occurring in the easy-to-hard condition ($M = 4.85$) than in the hard-to-easy condition ($M = 3.00$). The random condition ($M = 3.79$) did not differ significantly from either the easy-to-hard [$t(32) = 1.77$] or the hard-to-easy [$t(32) = 1.33$] conditions. Participants were both more optimistic and less accurate in evaluations of their performance when easier questions preceded more difficult questions than when questions were ordered randomly or in descending order of difficulty.

Experiment 2

Experiment 2 was designed to explore the effect of a time constraint on retrospective bias. It is possible that questions that appeared first made more of an impact on postdictions than did later questions because participants may have spent the entire testing time reflecting on them, making earlier questions more memorable. In Experiment 2, participants were instructed to spend an equal amount of time on each question and to focus on only one question at a time.

Method

Participants Twenty-one Case Western Reserve University students from introductory psychology classes participated in the study, in partial fulfillment of a course requirement.

Design As in Experiment 1, Experiment 2 used a within-subjects design with test list structure as the manipulated variable. We analyzed performance, postdictions, and bias in postdictions.

Materials The materials were identical to those of Experiment 1: We used three sets of 50 general knowledge questions taken from the Nelson and Narens (1980) norms to create the tests.

Procedure The procedure was similar to that in Experiment 1, with an added time constraint. Participants were again given three different tests, with test list structure manipulated within participants. Presentation order was counterbalanced across participants such that participants were divided into three groups of $n = 7$ in each group. Instead of being given 10 min to complete each test, as in Experiment 1, participants were given 10 sec to complete each question, a total of 8 min and 20 sec per test. The experimenter kept time and said "Next" each time 10 sec had passed. Participants were instructed to move on to the following question without working on any previous or upcoming questions. They were told to give as many correct responses as possible and were instructed to either make a guess or not respond to questions for which they were not sure of the answer. As in Experiment 1, participants were asked to estimate how many questions out of the previous 50 they believed they had answered correctly and to write this number on the back of the test.

Results

A repeated-measures ANOVA was used, with test list structure (easy to hard, random, and hard to easy) as the within-subjects variable. There were three dependent measures: performance, postdictions, and bias in postdictions (see Table 2).

Postdictions Mean postdictions, in terms of number of answers estimated correct out of 50, are presented in Table 2. Results showed a similar pattern to results of Experiment 1, though they did not reach significance [$F(2,40) = 2.41$, $MSe = 27.10$, $p = .10$, $\eta_p^2 = .11$]. The order of the three conditions was the same as in Experiment 1; however, the

Table 2 Mean postdictions, performance, and bias by test list structure in Experiment 2

	Postdictions		Performance		Bias (Difference)	
Test List Structure	*M*	*SE*	*M*	*SE*	*M*	*SE*
Easy to hard	20.62	1.84	16.62	1.15	4.00	1.34
Random	18.19	1.98	15.90	1.09	2.29	1.66
Hard to easy	17.19	1.97	16.81	1.06	0.38	1.60

only significant pairwise comparison was between the easy-to-hard and hard-to-easy conditions [$t(20) = 2.16$].

Performance Mean performance, in terms of number of answers correct out of 50, is presented in Table 2. No significant difference was found between conditions [$F(2, 40) = 0.59$, $MSe = 8.08$, $\eta_p^2 = .03$].

Bias in postdictions Bias data 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 [$F(2, 40) = 3.53$, $MSe = 19.51$, $\eta_p^2 = .15$]. Participants were more optimistic about their performance in the easy-to-hard condition than in the hard-to-easy condition [$t(20) = 2.56$]. The difference between the easy-to-hard and the random conditions did not reach significance [$t(20) = 1.30$] but was in the predicted direction: Participants were more optimistic about their performance when easier questions appeared earlier on the test than when the questions were randomly ordered. No significant difference was found between the random condition and the hard-to-easy condition [$t(20) = 1.41$]. Participants were overly optimistic about their performance when the easiest questions were presented at the beginning of the test [$t(20) = 2.99$]. That is, participants' ratings of their performance were significantly higher than their actual performance. Postdictions were not significantly different from performance in the random [$t(20) = 1.38$] or hard-to-easy [$t(20) = 0.24$] conditions.

The absolute difference between postdictions and performance did not significantly differ with test list structure [$F(2, 40) = 0.73$, $MSe = 7.96$, $\eta_p^2 = .04$], and averaged 10.6 % across conditions. Participants were equally (in)accurate in estimating their performance under each of the three test list structures.

Experiment 3

In Experiment 1, we replicated Weinstein and Roediger's (2010) discovery that changing the order in which test questions are arranged can have an impact on evaluations of test performance. In Experiment 2 we showed that this bias persists when attention is equalized across questions. However, it is possible that participants are not forming a global impression of the tests themselves, but rather remembering what or how much they physically wrote on the tests. Experiment 3 was conducted to examine whether participants formed test list structure bias as an impression of the test itself, or as a result of remembering how much they mechanically wrote. In addition to placing control on the amount of time participants spent on each question, as in Experiment 2, participants were given additional instructions to read the questions without answering them. After reading through the questions, participants were asked to predict how well they would perform on the test.

Method

Participants Twenty-three Case Western Reserve University students from introductory psychology classes participated in the study, in partial fulfillment of a course requirement.

Design As in Experiments 1 and 2, Experiment 3 used a within-subjects design, with test list structure as the manipulated variable. We analyzed participants' predictions of performance.

Materials The materials were identical to those of Experiments 1 and 2, with three sets of 50 general knowledge questions taken from the Nelson and Narens (1980) norms.

Procedure The procedure was similar to that in Experiment 2, with additional instructions not to write any of the answers. Participants were told that the experimenters were interested in how people form impressions of tests, not in actual test performance. Participants were again given three different tests, with test list structure manipulated within participants. Presentation order was counterbalanced across participants such that participants were divided into three groups, two groups of $n = 8$ and one of $n = 7$. As in Experiment 2, an experimenter kept time and said "Next" each time 10 sec had passed. Participants were instructed to focus on reading the current question without reading ahead or backtracking. After reading through a test, participants were asked to estimate how many questions out of the previous 50 they believed they would answer correctly if they were to take the test immediately, before looking up any answers, and to write this number on the back of the test. The experiment lasted around 45 min.

Results

A repeated-measures ANOVA was used, with test list structure (easy to hard, random, and hard to easy) as the within-subjects variable. There was one dependent measure: predictions (see Table 3). Since participants were not actually answering any questions, there was no measure of performance to analyze.

Table 3 Mean predictions by test list structure in Experiment 3

	Predictions	
Test List Structure	*M*	*SE*
Easy to hard	26.13	1.92
Random	23.02	1.74
Hard to easy	20.20	1.56

Consequently, there were no measures of bias or absolute error.

Predictions Mean predictions, in terms of number of questions out of 50 estimated to be correctly answerable, are presented in Table 3. Test list structure had a significant effect on predictions [$F(2,44) = 3.56$, $MSe = 56.94$, $p = .037$, $\eta_p^2 = .14$]. Pairwise comparisons revealed that the easy-to-hard condition had significantly higher predictions than the hard-to-easy condition [$t(22) = 2.53$]. There was no significant difference between the easy-to-hard and random [$t(22) = 1.32$] or random and hard-to-easy [$t(22) = 1.45$] conditions. Although the data suggest that participants might be more optimistic when they are not required to actually answer the questions, the differences between mean predictions in Experiment 3 and mean postdictions in Experiments 1 and 2 was not found to be significant [$F(2, 40) = 2.83$, $MSe = 121.80$, $\eta_p^2 = .12$]. It seems possible that participants would be more optimistic when making predictions as compared with postdictions. However, because the participants who took part in Experiment 3 were a different group of students, tested during a different time of the year than the participants in Experiments 1 and 2, we cannot rule out the possibility that they would actually have performed better, thus not having an impact on bias. The important finding is that test list structure impacted predictions following the same pattern as test list structure does in impacting postdictions.

Experiment 4

Experiment 4 was intended to extend this pattern to multiple-choice testing. One application for postdictions from outside the laboratory involves deciding when to cancel scores from a standardized test. Such tests are often mostly or exclusively in a multiple-choice format. However, previous demonstrations have utilized a free-response test format, so Experiment 4 used multiple-choice tests to determine the generality of these findings across test formats.

Method

Participants Twenty-four Case Western Reserve University students from introductory psychology classes participated in the study, in partial fulfillment of a course requirement.

Design Experiment 4 used a within-subjects design, with test list structure as the manipulated variable and with analyses conducted on postdictions, performance, and bias in postdictions.

Materials The materials were based on those of Experiment 1: We used three sets of 50 general knowledge questions taken from the Nelson and Narens (1980) norms to create the tests. The tests were transformed into multiple-choice format, with options chosen so that easy questions remained easy (e.g., selecting *Paris* as the capital of France, with *Dublin*, *Vienna*, and *London* as alternatives) and hard questions remained hard (e.g., selecting *Arthur* as the 21st U.S. president, with *Garfield*, *Harrison*, and *Cleveland* as alternatives).[2]

Procedure The procedure was similar to that in Experiment 2, with participants given 10 sec to complete each multiple-choice question. As in the previous experiments, participants were given three different tests, with test list structure manipulated within participants. Presentation order was counterbalanced across participants such that participants were divided into three groups of $n = 8$ in each group. They were not required to answer every question. After the completion of each test, participants were asked to estimate how many questions out of the previous 50 they believed they had answered correctly, and to write this number on the back of the test.

Results

A repeated-measures ANOVA was used, with test list structure (easy to hard, random, and hard to easy) as the within-subjects variable. There were three dependent measures: performance, postdictions, and bias in postdictions (see Table 4).

Postdictions Mean postdictions, in terms of number of answers estimated correct out of 50, is presented in Table 4. Results were similar to those of the previous experiments, with test list structure having a significant effect on postdictions [$F(2,46) = 8.39$, $MSe = 14.23$, $\eta_p^2 = .27$]. Pairwise comparisons revealed significant differences between the easy-to-hard and random conditions [$t(23) = 2.11$], the easy-to-hard and hard-to-easy conditions [$t(23) = 3.47$], and the random and hard-to-easy conditions [$t(23) = 2.61$]. The magnitude of postdictions in Experiment 4 was slightly higher compared with Experiments 1 and 2, which can be explained by the different testing formats: multiple choice compared with free response. With four possible responses to each question in a 50-question test, chance performance would be around 12.5 questions correct. In a free-response test, there is no such “guaranteed” chance performance, because participants are providing the answers themselves.

Performance Mean performance, in terms of number of answers correct out of 50, is presented in the middle column of Table 4. There was no significant effect of test structure on

[2] All 24 participants scored higher on easy questions ($M = 41.58$) than on hard questions [$M = 17.71$; $t(23) = 13.68$, $p < .01$].

Table 4 Mean postdictions, performance, and bias by test list structure in Experiment 4

	Postdictions		Performance		Bias (Difference)	
Test List Structure	*M*	*SE*	*M*	*SE*	*M*	*SE*
Easy to hard	24.42	1.53	19.67	1.00	4.75	1.12
Random	22.04	1.29	19.42	0.78	2.25	1.48
Hard to easy	19.96	1.18	19.79	0.81	0.54	1.29

number of questions answered correctly [$F(2, 46) = 0.10$, $MSe = 9.11$, $\eta_p^2 = .01$].

Bias in postdictions Bias data are shown in the far right column of Table 4 for each test list structure condition. Participants again tended to be optimistic, with postdictions exceeding performance in all three conditions; however, the difference was significant only in the easy-to-hard [$t(23) = 4.23$] condition and not in the random [$t(23) = 1.52$] or hard-to-easy [$t(23) = 0.42$] conditions. Test list structure had a significant effect on bias [$F(2, 46) = 5.51$, $MSe = 19.51$, $\eta_p^2 = .19$]. There were significant differences between the easy-to-hard and random conditions [$t(23) = 2.15$] and between the easy-to hard and hard-to-easy conditions [$t(23) = 3.24$]. The difference between the random and hard-to-easy conditions did not reach significance [$t(23) = 1.26$]. In addition, test list structure did not have a significant effect on absolute (unsigned) differences between postdictions and performance [$F(2,46) = 0.95$, $MSe = 15.05$].

General discussion

In Experiment 1, we successfully replicated findings from Weinstein and Roediger (2010), in which participants were more optimistic about their performance on tests in which easy questions were presented first than they were about tests in which difficult questions were presented first or in which questions were presented in a random order. In Experiment 2, we found that this bias for the easy-to-hard condition persisted when participants were instructed to spend an equal amount of time on each question. In addition to the equalized attention introduced in Experiment 2, in Experiment 3 participants were given further instructions to read the test questions without responding to them. When asked how they believed they would perform on the tests, participants gave higher predictions for tests in which questions were ordered from easy to hard than for tests in which questions were ordered randomly or from hard to easy. Experiment 4 extended the effect of test list structure on postdictions to multiple-choice testing.

In Experiment 1, the replication of Weinstein and Roediger (2010), we supported the finding that retrospective test evaluations are subject to memory bias. We were able to translate their findings from a computerized testing format in which participants were given one question at a time to a paper-and-pencil format in which participants were able to flip through the entire test. This difference in procedure augments the findings from Weinstein and Roediger by translating their task to a testing situation that is still common in many classrooms, as well as demonstrating the retrospective test list structure bias in a situation in which participants were given all test questions simultaneously. The results from Experiment 1 suggest that participants form a global impression of the test that is disproportionately influenced by the difficulty level of the earlier questions presented. Experiment 2 dispelled the possibility that the test list structure bias found for our testing format was caused by participants spending more time with, or giving more attention to, earlier questions. This was accomplished by imposing equalized attention across questions and still finding a postdiction bias for tests in which easier questions were presented at the beginning. Experiment 3, like Experiment 2, was designed to rule out a competing explanation to the global impression theory. A bias for the easy-to-hard condition persisted when participants did not provide any written answers to test questions, refuting the idea that participants form impressions of the tests according to how much they physically write at the beginning of the tests. Experiment 4 extended the pattern of these results from a free-response format to multiple choice, a format used in many situations, particularly in testing at the college level and in standardized testing. Thus, these experiments support the hypothesis put forth by Weinstein and Roediger (2010, 2012) that a retrospective bias for the test as a whole is responsible for the effect. This global anchoring bias shares similarities with what has been demonstrated in impression formation research, an area usually removed from the realm of cognitive psychology.

The experiments reported here all used a within-subjects design that may have the potential for carryover effects between conditions. However, our data suggest that such carryover effects are not responsible for the biases we observed. Performing a post-hoc analysis on the first trials culled from Experiments 1, 2, and 3 yielded a significant effect of test list structure on postdictions/predictions [$F(2, 74) = 10.15$, $p < .001$]. Mean combined postdictions/predictions for the first trial were 25.50 in the easy-to-hard condition, 20.32 in the random condition, and 15.73 in the hard-to-easy condition, with the overall mean at 20.52. Experiment 4 employed a different dependent variable (multiple-choice testing) on a different scale but showed a similar pattern: When only the first trial is considered, there was a significant effect of test list structure on postdictions [$F(2, 21) = 7.01$, $MSe = 31.64$], with means of 28.50, 21.75, and 18.12 in the easy-to-hard, random, and hard-to-easy conditions.

The importance of first impressions applies not only to people, but also to situations. Metacognitive judgments are typically evaluated in cognitive terms: Is a particular memory accessible or strong? However, metacognitive judgments are in part impressions, and can be thought of in a similar manner to the way social and personality psychologists have studied impressions of people. Just as social psychology has long demonstrated the importance of first impressions in our perception of individuals, first impressions of tests have a disproportional effect on test difficulty judgments. Achieving fair perception of people requires overcoming many of the biases that contaminate person perception; accurate retrospective metacognition may require us to overcome biases involved in test perception.

References

Ahlering, R., & Parker, L. (1989). Need for cognition as a moderator of the primacy effect. *Journal of Research in Personality, 23,* 313–317. doi:10.1016/0092-6566(89)90004-4

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258–290. doi:10.1037/h0055756

Crano, W. D. (1977). Primacy versus recency in retention of information and opinion change. *The Journal of Social Psychology, 101*, 87–96. doi:10.1080/00224545.1977.9923987

Devine, M., & Yaghlian, N. (n.d.). *Center for teaching excellence test construction manual: Construction of objective tests*. Retrieved from http://www.cte.cornell.edu/documents/Test%20Construction%20Manual.pdf

Dunlosky, J., & Thiede, K. W. (2013). Metamemory. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 283–298). New York: Oxford University Press.

Fergus, J. (1997). The professional woman writer. In E. Copeland & J. McMaster (Eds.), *The Cambridge companion to Jane Austen* (pp. 12–31). Cambridge: Cambridge University Press.

Forgas, J. (2011). Can negative affect eliminate the power of first impressions? Affective influences on primacy and recency effects in impression formation. *Journal of Experimental Social Psychology, 47,* 425–429. doi:10.1016/j.jesp.2010.11.005

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

Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods, 45,* 1115–1143. doi:10.3758/s13428-012-0307-9

Webster, D., Richter, L., & Kruglanski, A. (1996). On leaping to conclusions when feeling tired: Mental fatigue effects on impressional primacy. *Journal of Experimental Social Psychology, 32,* 181–195. doi:10.1006/jesp.1996.0009

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

Weinstein, Y., & Roediger, H. L., III. (2012). The effect of question order on evaluations of test performance: How does the bias evolve? *Memory & Cognition, 40,* 727–735. doi:10.3758/s13421-012-0187-3