THE ORDER OF QUESTIONS ON A TEST AFFECTS HOW WELL STUDENTS BELIEVE THEY PERFORMED

BY

GREGORY FRANCO

A thesis submitted to the Victoria University of Wellington in fulfilment of the requirements for the degree of Doctorate of Philosophy

Victoria University of Wellington
2015

Abstract

We know that students are more optimistic about their performance after they take a test that progresses from the easiest to hardest questions than after taking one that progresses in the opposite order[1]. In fact, these "Easy-Hard" students are more optimistic than "Hard-Easy" students even when the two groups perform equally. The literature explains this *question order bias* as a result of students' failing to sufficiently adjust, in the face of new information, their extreme initial impressions about the test. In the first two of six studies, we investigated the possibility that a biased memory for individual questions on the test is an alternative mechanism driving the question order bias. The pattern of results was inconsistent with this mechanism, but fit with the established impression-based mechanism. In the next four studies, we addressed the role that the number of test questions plays in determining the size of the question order bias, discovered that warning students is only a partially effective method for reducing the bias, and established a more precise estimate of the bias' size. Taken together, this work provides evidence that the question order bias is a robust phenomenon, likely driven by insufficient adjustment from extreme initial impressions.

[1] Although the research in this thesis is my own, I conducted it in a lab and supervised a team comprised of research assistants and honours students. I also received advice and direction from my supervisors. Therefore, I often use the word "we" in this thesis to reflect these facts.

Acknowledgements

Thank you to everyone who has ever been a classmate, colleague, teacher, mentor, partner, family member, or friend. Thank you to everyone who has ever given me a chance to try my hardest to succeed—whether that chance be my first, second, third, or fourth. Thank you to those who criticise my work. I know you only want to help me produce the best outcome possible. I haven't always been good at listening to you, so thanks for your persistence. Thank you to Maryanne Garry, who keeps throwing me in the deep end over and over. That experience is sometimes fun and often terrifying, but it is always worth the initial frantic flailing. I've never learned so much so quickly. Thank you to Dr. Matt Crawford for your helpful comments and expertise. Most of all, thank you to the optimists in my life—particularly the most persuasive one. In other words, thanks Lib.

Next, thank you to these people whom I either promised myself or them that I would thank by name: Mom, Dad, Colleen, Janelle, the Rat Pack, the Santa Cruz crew, Dragons & Lasers (PewPewROAR), Cassie Verdon, Jackson Miller, Dr. Jeff Foster, Dr. Eryn Newman, Dr. Brittany Cardwell, Robert Michael, Mevagh Sanson, the remaining members of the Garry Lab past and present, Dr. Matt McCrudden, and Dr. Yana Weinstein. I'm sure you all know what you've done to help me get to this point—or not. Memory is funny like that.

Table of Contents

List of Figures and Tables

Chapter 1: Introducing the Question Order Bias

Of all the emails that land in a teacher's inbox every semester, one of the most common reads something like, "Dear Professor, I just saw my grade on the test and I thought I had done much better than that. Can you check to see that it was scored correctly?" But almost inevitably, the test was scored correctly. So why do so many students leave tests with such distorted views about their performance? There are many explanations for this phenomenon. Sometimes students overestimate because they hold unrealistic expectations based on their performance on previous tests—even when those tests are not related to the current one (Clayson, 2005; see Moore & Healy, 2008 for a review). Sometimes students are simply incompetent—unable to know what they do not know (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Ehrlinger, Johnson, Banner, Dunning, & Kruger 2008; Kennedy, Lawton, & Plumlee, 2002; Kruger & Dunning, 1999). Sometimes the problem stems from the way students naturally form impressions about a test as they take it (Weinstein & Roediger, 2010, 2012). This third scenario is especially likely when the questions on the test were arranged in order by difficulty. Students believe they did better on tests that progressed from easy questions to difficult questions (Easy-Hard tests) compared to tests that progressed in the opposite order (Hard-Easy tests) even when the tests contained the exact same questions (Feldman & Bernstein, 1977, 1978; Jackson & Greene, 2014; Jones, Rock, Shaver, Goethals, & Ward, 1968; Weinstein & Roediger, 2010, 2012).

The evidence that students are accurate in this belief is mixed. Until recently, it was believed that Easy-Hard tests foster a slight—about 3%—advantage in performance over Hard-Easy tests (see Aamodt & McShane, 1992 for a meta-analysis). One explanation for this advantage was that when students begin with a run of easy questions before they answer the hard questions, they have less anxiety and frustration. These reductions can translate to higher performance (Aamodt & McShane, 1992; Munz & Smouse, 1968; for a review, see Vander Schee, 2013). But newer evidence suggests any advantage in

performance produced by question order may even be smaller than previously suggested—perhaps so small as to be trivial (Vander Schee, 2013; Weinstein & Roediger, 2010, 2012).

Regardless of any differences in actual performance on the test, students think they performed better when they took an Easy-Hard test compared to a Hard-Easy test (Feldman & Bernstein, 1977, 1978; Jackson & Greene, 2014; Jones, et al., 1968; Weinstein & Roediger, 2010, 2012). This *question order bias* can be found even when experimenters control for differences in actual performance on the two tests (Jackson & Greene, 2014; Weinstein & Roediger, 2010, 2012). The most recent demonstrations show that the bias occurs across a variety of test formats: paper and pencil, digital, multi-choice, and/or cued recall (Jackson & Greene, 2014; Weinstein & Roediger, 2010).

There is no single explanation for why the question order bias occurs, but researchers have considered several candidate mechanisms. Without immediate feedback, there is no objective way for students to know how well they performed. So, they instead must rely on a different cue as a shortcut—or heuristic—to help them judge their performance. People rely on heuristics to make a wide variety of judgements (see Alter & Oppenheimer, 2009 for a review). For example, when information is easier to read, whether because it stands out in high contrast from the background or is written in clean, tidy font, people inaccurately judge that information to be better learned than information that is difficult to read (Alter, Oppenheimer, Epley, & Eyre, 2007; Rhodes & Castel, 2008). Each candidate mechanism proposes students rely on a different cue to estimate their performance on a test.

Candidate Mechanism 1: Affect-as-Information, "I feel good about the test."

One cue people may use is how good or bad they feel when they make the judgement. For the purposes of this mechanism, *affect* is defined as "the specific quality of 'goodness' or 'badness' experienced as a feeling state (with or without consciousness)" (Slovic, Finucane, Peters, & MacGregor, 2007, p. 1333). It is possible that when students take tests ordered by difficulty, they experience

differing levels of positive and negative affect—particularly at the beginning of the test. Hard-Easy students may begin with more negative affect, and Easy-Hard students may begin with more positive affect. These initial differences may translate to higher or lower estimates of performance at the end of the test.

Generally, people make more positive judgements about targets when they experience more positive affect (for a review, see Clore & Huntsinger, 2007). The *affect-as-information* theory proposes that if people attribute their affective state to the target being judged, their judgement will reflect that fact (Clore, Wyer, Dienes, Gasper, Gohm, & Isbell, 2001). For example, people experiencing high positive affect often judge targets to be more honest, preferred, trustworthy, and more—even when the target was not the true source of people's positive feelings (Slovic et al., 2007).

Another judgement that could be influenced by how good or bad people feel is how well they performed on a test. Peoples' initial affective response to a target can set the mood for the way they view subsequent interactions with it (Zajonc, 1980). Students who take a Hard-Easy test (Hard-Easy students) experience more difficulty and anxiety early on compared to those who take an Easy-Hard test (Easy-Hard students; Munz & Smouse, 1968). Those experiences result in more initial negative affect and lower overall performance (Hinze & Rapp, 2014). Once this "mood" of negativity is set, Hard-Easy students may not fully recover even when they begin to answer easier questions—the proverbial wind has been taken from their sails. As a result, Hard-Easy students end the test with more negative affect than Easy-Hard students and therefore make more pessimistic estimates of their performance.

The empirical evidence for this affect-driven explanation is sparse. If affect plays a substantial role in driving the question order bias, we would expect to see Hard-Easy and Easy-Hard subjects make affective judgements about the test in ways that reflect a respectively negative or positive affect. One study asked subjects to rate their level of enjoyment throughout the test—an affective judgement (Weinstein & Roediger, 2012). These ratings were

symmetrical regardless of question order and Hard-Easy subjects made higher ratings of overall enjoyment by the end of the test than Easy-Hard subjects. These findings suggest Hard-Easy subjects have no problem recovering from their initial experiences of difficulty and even their affective judgements are not biased by their initial affective experiences. It is therefore unlikely that affect drives the question order bias for judgements of performance either.

There is a second feature of the question order bias inconsistent with an affective explanation: as mentioned previously, question order does not seem to change actual performance. If the question order bias is driven by affective differences between students who take opposite orders of a test, we would expect Hard-Easy students to consistently perform better than Easy-Hard students. Recall that according to the affect-as-information hypothesis, Hard-Easy students would have more negative affect throughout the test than their Easy-Hard counterparts. Negative affect signals a threat and people tend to devote more cognitive resources to processing their environment so they can avoid any negative outcomes associated with the threat (see Clore, Schwarz, & Conway, 1994 for a review; Schwarz & Bless, 1991). In a testing situation, one negative outcome would be poor performance. If Hard-Easy students exert more effort on the test than Easy-Hard students, they should also perform better, but researchers do not find that pattern (Aamodt & McShane, 1992; Vander Schee, 2013; Weinstein & Roediger, 2010, 2012). It is therefore unlikely the question order bias is driven by differences in affect created by the order of test questions.

Candidate Mechanism 2: On-Line Impression Formation, "I think I did well."

The second candidate mechanism proposes students rely on their impression about the ease or difficulty of the test to inform how well they *think* they performed overall, but this impression is biased by the difficulty of the first few test questions.

People form impressions about nearly anyone and anything they encounter (Ambady & Skowronski, 2008). They form these impressions by

making inferences based on interactions with or information about a target (Brown & Bassili, 2002; Schaller, 2008; Uleman & Kressel, 2013; Winter & Uleman, 1984). For example, if a stranger buys you fresh-squeezed orange juice when you are sad, you are highly likely to infer that she is a caring or kind person. These inferences can be made somewhat automatically and without the intention or even the awareness of the observer (Uleman, Adil Saribay, & Gonzales, 2008).

Observers form impressions about targets during an interaction using a three-step process. First, a given behavior (say, someone buying you orange juice) activates a relevant trait (say, kindness). Next, an associative link is formed between the trait and the target. Finally, the observers make a dispositional inference about the target based on the trait (the person who bought you orange juice *is* kind; Crawford, Skowronski, Stiff, & Scherer, 2007; Skowronski, Carlton, Mae, & Crawford, 1998). Observers often use their memory for that association, or inference, rather than a memory for the specific behavior from which it was formed, to make subsequent judgments about the target (the kind stranger is attractive)—especially when these links have been frequently re-visited and updated based on subsequent interactions (Carlston & Skowronski, 1986). Judgements based on such impressions are classified as on-line judgement tasks under Hastie and Parks' (1986) framework.

When observers have to incorporate more than one interaction with a target, they are notoriously bad at doing so accurately. Judgements based on these overall impressions tend to be biased towards the qualities suggested by observers' first few interactions with the target(Anderson, 1965; 1973; Anderson & Hubert, 1963; Crano, 1977). For example, one study showed that when subjects were given a list of a stranger's traits that began favourably and ended unfavourably, they rated that stranger more positively than when the list was presented in the opposite order (Anderson & Barrios, 1961). These primacy effects for impression-related judgements can be found about events, strangers, products, legal evidence, and more (Anderson & Norman, 1964; Crano, 1977;

Hastie & Park, 1986; Mantonakis, Rodero, Lesschaeve, & Hschwaastie, 2009; see Hogarth & Einhorn, 1992 for a review). These effects arise because people do not sufficiently adjust their initial impressions about a target to account for subsequent information.

There are two main explanations for why people do not adjust sufficiently (Crano, 1977). The first explanation shares many qualities with the affect-as-information hypothesis. In fact, whether a target causes positive or negative affect is sometimes part of what makes up people's impression of the target itself (Forgas & Bower, 1987). Once people form an initial impression, they interpret any subsequent ambiguous information to fit more closely with the established context (Asch, 1946; Hamilton & Zanna, 1974; Kaplan, 1971; Zanna, & Hamilton, 1977). For example, in one study, subjects interpreted ambiguous descriptors of a stranger more positively when they had just seen a different, positive descriptor compared to when they had seen a negative one (Hamilton & Zanna, 1974).

The second explanation for why people don't sufficiently adjust from their initial impressions is analogous to a decline in or fatigue of attention to new stimuli over time. Once people have formed an initial impression based on preliminary interactions with a target, they give less weight to information from subsequent interactions (Belmore, 1987; Feldman & Bernstein, 1977). This *attentional fatigue* explanation differs from the previous one because it predicts that information encountered later in a sequence is attended to and influences the overall impression less than early information, but later information still retains its original meaning (Crano, 1977). This mechanism suggests manipulations that encourage people to pay more attention to sufficient adjustment can reduce or eliminate the amount their judgements are biased by their initial impressions. Taken individually or together, these two explanations for insufficient adjustment demonstrate that early experiences with a target can guide and limit how people process subsequent experiences.

Thus, when people need to make a judgement based on an overall impression formed from many interactions with a target, their judgements are often biased to reflect the qualities of the first few interactions. We can extend these ideas to account for the way students judge their performance on a test. Students' initial impressions of a test would be based on their first few interactions with it—its first few questions. This initial impression is accurate when the first few questions feel similar to the questions on the test as a whole. But when questions are arranged by difficulty, initial questions are extremely dissimilar to the questions on the test as a whole. Thus, the students' impressions end up biased. Students' initial impressions would instead begin at an extreme point ("This test is extremely easy/difficult") that reflects only the way the test begins. Once students form this extreme initial impression, they might not sufficiently adjust it as they encounter new questions. By the end of the test, students who began with easy questions may hold an overall impression that the test was easier, and overestimate their performance accordingly, compared to those who began with difficult questions (Weinstein & Roediger, 2010, 2012). In other words, the second candidate mechanism proposes Easy-Hard students mistakenly think they did better than Hard-Easy students because they do not sufficiently adjust their extreme initial impressions.

Candidate Mechanism 3: A Memory-Based Strategy, "I remember easy questions."

Candidate Mechanism 2 proposes students use a cue formed on-line during the test (their impression of ease or difficulty) to help them judge their performance. Candidate Mechanism 3 proposes students instead judge their performance based on how well they think they performed on questions they can *remember* from the test.

There is good reason to expect students rely on their memories for individual questions as opposed to an overall impression of ease or difficulty to estimate their test performance. One factor that influences which cue people use

as a heuristic is the cue's subjective validity for the judgement at hand (Shah & Oppenheimer, 2008). An abstracted impression of how easy or difficult the test was may not be the most valid cue for judging test performance. A more valid cue is something more directly related to performance. For example, how well do students think they performed on individual questions from the test?

To answer this question, students could attempt to recall every question on the test and tally the number they believe they answered correctly. Then, they could use that tally to estimate their performance. This attempt would result in the question order bias if students are better able to recall questions from the first half of the test than the second half.

Students could also adopt a similar strategy, but instead of exhaustively searching their memory for every test question, they could stop once they reach an internal criterion of what is sufficient to make their estimate (Simon, 1956). In other words, students may recall a few questions from the test, estimate their performance on those questions, and use that estimate to inform their judgement of overall performance. If students adopt this approach, the questions that are most available for recall will heavily influence how well they believe they performed. This scenario is analogous to the use of an *availability heuristic* (Schwarz et al., 1991; Tversky & Kahneman, 1973).

The availability heuristic is simple: if it is easy to recall examples of something, people mistake that ease as evidence that there are many similar "somethings" in existence. For example, subjects incorrectly guessed that there are more words in the English language that begin with a letter such as "k, r, n, l, and v" than words that feature one of those letters as their third letter because it is easier to recall examples of the former (Tversky & Kahneman, 1973).

For the availability heuristic to explain the question order bias, students would need to find it disproportionally easy to remember examples of early test questions. This possibility might seem unlikely, because we know that people typically recall many more late items than early items from long lists—and tests

are often long lists of questions (Brown, Neath, & Chater, 2007; Murdock, 1962). Still, it remains an empirical question.

Both the exhaustive recall and availability heuristic approaches would be classified as memory-based judgement tasks under Hastie and Park's (1986) framework. Experiments 1a-d and 2 investigate the extent to which such a memory-based approach drives the question order bias.

Chapter 2: Investigating the Memory for Questions Mechanism

Hastie and Park's (1986) on-line vs memory-based judgement task framework predicts that when people make judgements about a target using a memory-based strategy, there should be a close relationship between the specific interactions with the target people can recall and the qualities of the judgements people make. For example, the more positive interactions people can recall about a person, the more positively that person will be rated. By contrast, when people use an on-line strategy, no such relationship is predicted.

In Experiments 1a-d, we investigate the evidence that students use a memory-based strategy to estimate their test performance. To do so, we find out which questions students can recall after taking either an Easy-Hard or a Hard-Easy test. If students use a memory-based strategy, we would expect them to be better able to recall early test questions than later test questions because Easy-Hard students are typically more optimistic about their performance than their Hard-Easy counterparts.

As the primary aim of this experiment was to investigate which questions are easiest for people to recall from a test arranged by difficulty, we did not measure subjects' actual performance nor take a measure of their estimated performance.

Experiments 1a-d

Method

Designs. Across Experiments 1a-d, we used a 2 group (Test question order: Easy-Hard, Hard-Easy) between subjects design.

Subjects. For all experiments in this thesis, we used Amazon's Mechanical Turk (MTurk; https://www.mturk.com/mturk/welcome) to recruit English-speaking subjects[2]. We recruited 76 people to participate in Experiment

[2] Mechanical Turk is an online subject pool. MTurk subjects complete experiments and surveys and are given small amounts of Amazon credit that they can use to purchase things on Amazon.com. These subjects are diverse and the data from studies run online using MTurk often produces similar results to those run in a laboratory or with other online subjects (Buhrmester, Kwang, & Gosling, 2011; Germine et al., 2012; Mason & Suri, 2011).

1a; 43 in Experiment 1b; 54 in Experiment 1c, and 67 in Experiment 1d for $0.50[3].

Procedures. These experiments had two phases. In the first phase, we told subjects:

> For this HIT[4], you will be answering 50 general knowledge questions at your own pace. Your aim is to maximise the number of correct responses you give. Please answer each question from your own memory. Avoid searching for the answers on the internet or other external sources. Please enter an answer for each question. If you do not know the answer, please enter a plausible guess.

Next, subjects took a 50 question trivia test[5]. For all experiments in this paper, we used the same trivia tests as Weinstein and Roediger (2010; see also Nelson & Narens, 1980)[6]. We asked subjects to take tests that contained 50 questions, arranged by difficulty. Questions that many people tended to answer correctly (according to the Nelson & Narens norms) were considered easy, and questions that many people tended to answer incorrectly were considered difficult. In the norms, approximately 97.40% of people answered the easiest questions correctly, and fewer than .004% answered the most difficult questions correctly. Each test was arranged in order from the easiest to the hardest or vice versa.

After subjects answered the last question, the second phase began. In this phase, all subjects performed the same recall task, but with four small variations. In Experiments 1a-d, we told subjects

---

[3] Across experiments 1a-d, one hundred and thirty two people did not complete the entire experiment and we therefore did not include them in our data analyses.

[4] A Human Intelligence Task (HIT) is a term used on MTurk to refer to a voluntary task such as these psychology studies.

[5] For all experiments in this thesis except Experiment 3, subjects took one of two versions of the trivia test. These versions were counterbalanced across subjects. See Appendix A for the full versions of both tests.

[6] There is now an updated version of these norms published (Tauber, Dunlosky, Rawson, Rhodes, & Sitzman, 2013).

> You were asked a total of 50 trivia questions. We need to know which questions you can remember, so now your task is to try to recall as many questions as you can. We are not interested in the exact wording of each question—just type enough so that we know which particular question you are remembering. You DO NOT have to recall the questions in the order they were presented.

In Experiment 1b, for each question they recalled, subjects rated their confidence that they had answered it correctly on the test. They made these ratings using a Likert scale from 1 (*not at all confident*) to 5 (*very confident*). In Experiment 1c, we offered subjects an incentive for recalling a large number of questions—double their compensation. In Experiment 1d, subjects received the same instructions as in Experiments 1b and 1c combined. Subjects then reported all the questions they could recall[7].

Instructional manipulation check and compliance check. For all experiments in this thesis, after the experiment proper concluded, we asked subjects to respond to a series of questions designed to identify those who were not taking the task seriously, as well as those who had not complied with the instructions. Because the proposed mechanisms that underly the question order bias are heavily dependent on people's interactions with a target, we excluded subjects who did not pass these checks to ensure that each subject was genuinely trying to take the test. For all experiments, when these subjects were included in our analyses, the patterns of our results remained similar, but (unsurprisingly) *p*-values and confidence intervals varied.

To identify people who did not take the test seriously, we used an Instructional Manipulation Check (Downs, Holbrook, Sheng, & Cranor, 2010; Oppenheimer, Meyvis, & Davidenko, 2009)—a paragraph about the 2012 Olympic games that ends with the sentence, "Which city will be hosting the

[7] For our analyses on confidence ratings in Experiments 1b and 1d, see Appendix D.

2012 Summer Olympic Games?" followed by 10 cities as options[8]. Embedded in the paragraph was this sentence: “This question is included to make sure you are paying attention to the study. Please disregard the rest of this paragraph and choose the third option.” The third option was "Shanghai." Only those subjects who chose the third option were included in our analyses.

To identify subjects who did not comply with the instructions, we told them:

> We are interested in response times and patterns of responses between individuals who answered the previous questions using their own knowledge and those who looked up the answers using an external source. Regardless of your strategy, you will be paid for this HIT. Please indicate your strategy below.

Choices were "I answered each question from my own memory", "I answered most questions from my own memory, but I consulted an external source for at least one question", "I answered about half of the questions from my own memory", "I answered most questions using an external source, but used my own memory for at least one question", and "I answered each question using an external source." Only subjects who chose "I answered each question from my own memory" were included in our analyses.

Results and Discussion

We excluded 27 (36%) subjects in Experiment 1a, 13 (24%) in Experiment 1b, seven (13%) in Experiment 1c, and 11 (16%) in Experiment 1d for failing the IMC and/or indicating that they consulted an external source for at least one question. For all experiments in this thesis, our exclusion rates are comparable, if not, lower than those found by Oppenheimer et al. (2009). After exclusions, 49 subjects remained in Experiment 1a, 30 remained in Experiment 1b, 47

[8] For studies that were conducted after 2012, the question was modified to “Which country hosted the 2012 Summer Olympic Games?”

remained in Experiment 1c, and 56 remained in Experiment 1d[9]. See Appendix E for details about demographics.

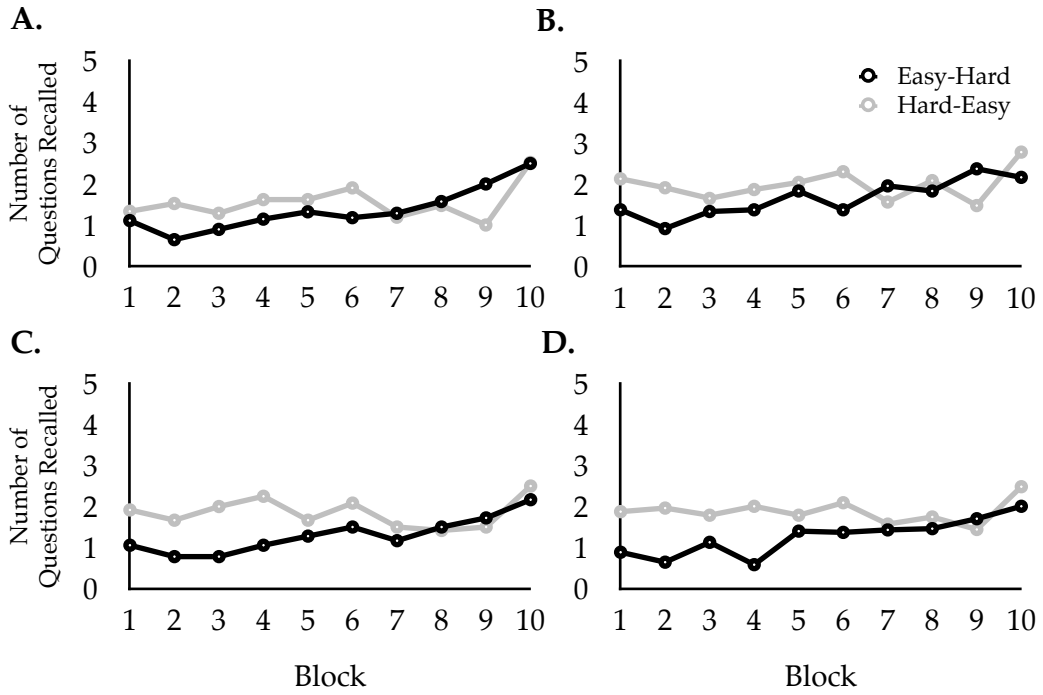
Number of questions recalled. In addition to analysing the serial position of each question subjects recalled, we also measured the total number of questions subjects recalled. Across all experiments, there was a very small trend for Easy-Hard subjects to recall fewer questions than their Hard-Easy counterparts. The individual trends for each study are shown in Table 1. A close inspection of Figures 1a-d show that this overall trend was concentrated around questions that appeared in roughly the first half of the test.

Table 1. Experiments 1a-d, the mean number of questions subjects recalled from their tests.

Experiment	Question Order	Mean (SD)	95% CI_{diff}
1a	Easy-Hard	13.64 (8.02)	[-3.09, 6.77]
	Hard-Easy	15.48 (9.06)	
1b	Easy-Hard	13.00 (9.96)	[-0.74, 11.74]
	Hard-Easy	18.50 (6.57)	
1c	Easy-Hard	16.38 (10.68)	[-2.95, 9.85]
	Hard-Easy	19.83 (11.11)	
1d	Easy-Hard	12.55 (9.80)	[0.64, 11.66]
	Hard-Easy	18.70 (10.55)	

Serial positions. We now turn to the primary issue: did people report more questions from the beginning of the test than from any other part of the test? If so, a memory-based strategy might explain the question order bias. To address this issue, we asked a research assistant, blind to condition, to identify the serial position of each question subjects recalled. Contrary to the memory-based strategy hypothesis, people recalled more questions not from the beginning of the test, but from the end. This pattern was consistent regardless of test order, incentive, and whether subjects rated their confidence that they answered those questions correctly on the test (Figures 1a-d). In null-hypothesis

[9] The pattern of exclusions did not vary depending on test order for any of Experiments 1a-d, all p's > 0.10.



Figures 1a-d. Mean number of questions recalled in each block of 5 questions on the test by serial position.

A. Experiment 1a: Block 1—Easy-Hard (M=1.11, SD=0.96), Hard-Easy (M=1.33, SD=1.32) Block 10—Easy-Hard (M=2.50, SD=1.40), Hard-Easy (M=2.52, SD=1.40)

B. Experiment 1b: Block 1—Easy-Hard (M=1.06, SD=1.06), Hard-Easy (M=1.92, SD=1.00) Block 10— Easy-Hard (M=2.17, SD=1.20), Hard-Easy (M=2.50, SD=0.67)

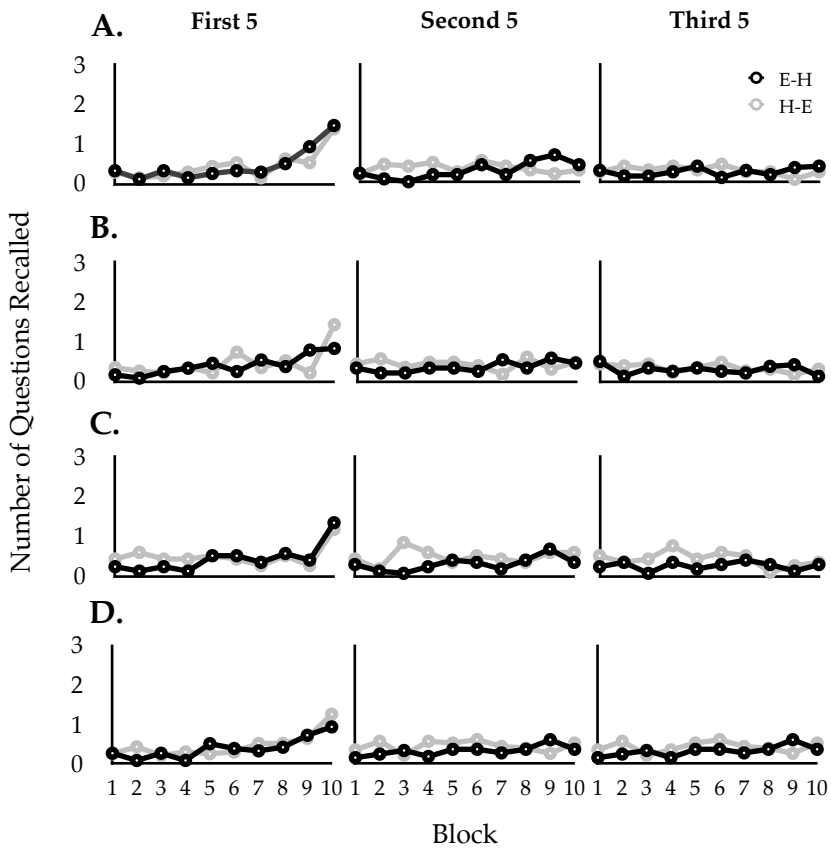
C. Experiment 1c: Block 1—Easy-Hard (M=1.38, SD=1.17), Hard-Easy (M=2.13, SD=1.17) Block 10—Easy-Hard (M=2.17, SD=1.49), Hard-Easy (M=2.78, SD=1.35)

D. Experiment 1d: Block 1—Easy-Hard (M=0.94, SD=1.09), Hard-Easy (M=1.70, SD=1.52) Block 10—Easy-Hard (M=1.88, SD=1.73), Hard-Easy (M=2.48, SD=1.34)

significance-testing (NHST) terms, for each of Experiments 1a-d, we found significant one-way Repeated-Measures ANOVAs on the mean number of questions subjects recalled from each 5-question block of the test,1a: $F(9, 490) = 5.55$, $p < 0.01$, 1b: $F(9, 290) = 2.53$, $p < 0.01$, 1d: $F(9, 550) = 2.04$, $p = 0.04$, 1c: $F(9, 460) = 2.19$, $p = 0.02$, 1d: $F(9, 550) = 2.53$, $p < 0.01$[10].

This pattern of recall by serial position provides an explanation for why Hard-Easy subjects showed a consistent trend to recall more questions than

[10] We also ran a version of Experiment 1a in which subjects estimated their performance after recalling five questions from anywhere on the test. This study is referred to as "Franco & Garry, Availability" in the mini meta-analysis reported later in this thesis. We found the same pattern of recall as in Experiments 1a-d, that is, subjects recalled more questions from the end of the test than from the beginning, regardless of question order.



Figures 2a-d. Mean number of questions recalled in each block of 5 questions on the test by serial position—separated to depict the first five questions reported by each subject, the second five, and the third five.

Easy-Hard subjects. Subjects likely spent more time and effort processing the hard questions than the easy questions. These memory-aiding resources may have helped Hard-Easy subjects more easily recall questions from the beginning of the test (Craik & Lockart, 1972; Johnston & Uhl, 1976). By contrast, the hard questions were already relatively available for recall in the Easy-Hard group because they were seen very recently, so Easy-Hard subjects may not have received the same boost in overall number of questions recalled.

At first glance, the data across Experiments 1a-d do not fit with the idea that students use their memory for individual test questions to inform their global estimates of performance. There is no evidence that subjects could better recall questions from the beginning of the test than those from the end. But the previous analyses only considered what students would recall if their search

was exhaustive. As previously mentioned, students may take a shortcut and adopt an approach similar to the availability heuristic instead. This approach would be biased by the questions that are the most available for recall (Tversky & Kahneman, 1973). We would find evidence for this memory-based strategy if questions from the beginning of the test were more available than those from the end.

In order to investigate this possibility, we constrained our analyses to the first few questions subjects reported—assuming those questions were the ones that sprung to mind most easily. As Figures 2a-d make abundantly clear, subjects tended to report recent items first. In fact, as Table 2 shows, the very last question was reported first more frequently than both the first and second questions combined in Experiments 1a-d. It is therefore unlikely the question order bias is driven by students' memory for the individual questions on the test. Instead, the most likely explanation remains Candidate Mechanism 2: students insufficiently adjust from their initial impressions about the test and rely on those impressions to make their estimates of overall performance.

Recent test questions may have been especially available for several reasons, but two are especially likely in this scenario. First, recency effects are pervasive in free-recall scenarios (Baddeley, Eysenk, & Anderson, 2009). It is easier to temporally discriminate recent items from their antecedents, much like nearer telephone posts are easier to discriminate spatially than those that are further away (Crowder, 1976). Second, as is typical in free recall scenarios with an impression formation component, subjects may have preferentially encoded experiences that were inconsistent with their impressions, but did not assign those experiences extra weight when adjusting their impression (for a review, see Hastie, 1980).

Our data show that people are good at recalling recent test questions when we ask them to recall as many questions as they can—in fact, those questions are often the first people report. Perhaps, then, asking people to recall the last few questions from the test before estimating their performance would

increase the salience of these questions. When cognitive feelings such as ease of retrieval are highly salient, people are more likely to use them when making ambiguous judgements such as estimating performance on a test (Kühnen, 2010; see Greifeneder, Bless, & Pham, 2010 for a review). If we increase people's use of questions that are easily recalled (recent questions) we should reduce people's use of their impressions to estimate their performance. This reduction could produce the opposite of the typical question-order bias. That is, Easy-Hard test takers would be more pessimistic about their performance than Hard-Easy test takers. We address this possibility in Experiment 2.

Table 2. Experiments 1a-d proportion of subjects who reported the initial, second and final test questions first

		Serial Position		
Experiment	**Test Order**	**Initial Question**	**Second Question**	**Final Question**
1a	Easy-Hard	0.11	0.00	0.46
	Hard-Easy	0.00	0.00	0.48
1b	Easy-Hard	0.11	0.00	0.22
	Hard-Easy	0.08	0.00	0.25
1c	Easy-Hard	0.00	0.00	0.29
	Hard-Easy	0.04	0.00	0.35
1d	Easy-Hard	0.03	0.00	0.24
	Hard-Easy	0.04	0.00	0.33

Experiment 2

Method

Design. We used a 2 (Test question order: Easy-Hard, Hard-Easy) X 2 (Recall: Yes, No) between subjects design.

Subjects. Based on the results of a previous study (reported in Chapter 5 as Franco & Garry: MTurk Replication, p. 46) and our expected exclusion rates, we aimed to collect 560 subjects—140 subjects for each between subjects cell. Due to the way our survey platforms (MTurk and Qualtrics) interact, it is possible to collect more subjects than requested. Because of this fact, this and all subsequent experiments have *N*s that deviate slightly from our intended sample sizes. For this study, we recruited 564 people on MTurk to participate for $0.50[11].

Materials and procedure. The procedure for Experiment 2 matched those of Experiment 1a with three exceptions. First, directly after the test, we instructed half of subjects to recall and write down only the last five questions they saw and the other half simply skipped the recall phase. Second, all subjects then estimated the number of questions they answered correctly on the entire test. Third, after estimating performance, subjects answered the question, "Did you notice anything in particular about the way that the questions were arranged on the trivia test you just took?" If anyone answered "yes" they elaborated on what they noticed.

Results and Discussion

We excluded 119 (21%) subjects were excluded for failing the IMC and/or reporting that they consulted an external source for at least one question on the test[12]. After exclusions, 445 subjects remained. See Appendix E for details about demographic information.

---

[11] Ninety nine people did not complete the entire experiment and we therefore did not include them in our data analyses.

[12] The pattern of exclusions did not vary depending on test order X^2 (1, N = 119) = 0.09, p = 0.99.

Scoring criteria. Questions on the test were scored according to a formula. If the first three letters of a subject's answer matched the first three letters of the correct answer (as indicated by the Nelson & Narens, 1980 norms), the question was scored as correct. For example, for the question "What is the name of the comic strip character who eats spinach to increase his strength?" "Popeye" was the correct answer. Answers that began with "Pop" were scored as correct. In a few cases, there were common misspellings or alternative spellings in the first three letters of an answer. For these questions, we included the misspellings and alternative spellings in our formula. For example, for the question "What is the name of the Chinese religion founded by Lao Tse?" "Taoism" and "Daoism" are both correct answers. Therefore, answers that began with "Tao" or "Dao" were scored as correct.

Number of correct answers. Before turning to our primary question, we first calculated subjects' actual performance. We replicated Weinstein and Roediger's (2010, 2012) finding that subjects performed similarly, regardless of whether they took the Easy-Hard (M = 56.16%, SD = 15.58) or Hard-Easy (M = 58.22%, SD = 15.96) test, M_{diff} = 2.06%, 95% CI [-0.86, 5.00]. In addition, subjects who recalled five questions before they estimated their performance (M = 59.20%, SD = 15.14) answered a mean of 4% more questions correctly than subjects who did not (M = 55.20%, SD = 16.16), 95% CI [1.08, 6.92]. This second result was surprising because the recall task only came after the test was completed and therefore could not have affected people's performance. This difference in mean performance held no influence over how biased people's estimates about their performance were because we corrected people's estimates to account for individual performance.

Question order bias. For these and many subsequent analyses in this thesis, we must consider how best to frame the effects of question order on people's estimates of performance. On the one hand, we could say that people who take Easy-Hard tests are optimistic because their mean estimates are higher than their actual mean performance (Weinstein & Roediger, 2010; 2012).

But the problem with using actual performance as a reference point is that performance often depends on arbitrary criteria set by the person who marks the test (see Weinstein & Roediger 2010).

On the other hand, we could say that people who take Easy-Hard tests are optimistic compared to some "status quo" benchmark such as a randomly-ordered test. But such an approach comes with its own problems. First, there is no evidence that randomly-ordered tests are the norm among real-life educators. Second, there is no guarantee that a single randomly-ordered test begins with questions that are representative of the test's entire range of difficulty. In fact, it is plausible that such a test could end up being arranged identically to an Easy-Hard or a Hard-Easy test. To limit these problems and to be consistent with previous research (Weinstein & Roediger, 2012), we interpreted opposite test orders relative to each other, comparing people's estimates after an Easy-Hard test with others' estimates after a Hard-Easy test.

Also consistent with Weinstein & Roediger's (2010, 2012) approach, we corrected these estimates for subjects' actual performance. We did so by subtracting the number of questions each subject actually answered correctly from the number they estimated they answered correctly. This correction resulted in a difference score that we refer to as a *bias score*. Subjects who overestimated their performance received a positive bias score, and subjects who underestimated received a negative bias score. We performed this correction to account for any differences in actual test performance between groups[13].

Did calling attention to the last few questions on the test change the way people estimated their performance? If so, we should have led Hard-Easy subjects to overestimate and Easy-Hard subjects to underestimate. But that pattern is not what we found. The question order bias was equally as large

[13] An alternative method of correcting for individual performance is to measure subjects' performance estimates and include their actual performance in the model as a covariate with Test Order and Recall. We conducted these analyses for Experiments 2, 3, and 4 and found the same patterns of results described in the main text.

regardless of whether subjects recalled test questions before estimating their overall performance (see Figure 3). In fact, recalling questions before estimating performance simply made everyone more pessimistic regardless of test question order, M_{diff} = 3.86%, 95% CI [1.40, 6.34]. In NHST terms, there were main effects of Recall and Test question order, $F(1, 443) = 9.46$, $p < 0.01$ and $F(1, 443) = 78.37$, $p < 0.01$, respectively, but no significant interaction $F(3, 441) = 1.67$, $p = 0.20$. Taken together, these findings show that drawing subjects' attention to the last few questions on the test before they estimated their performance only trivially affected the size of the question order bias.

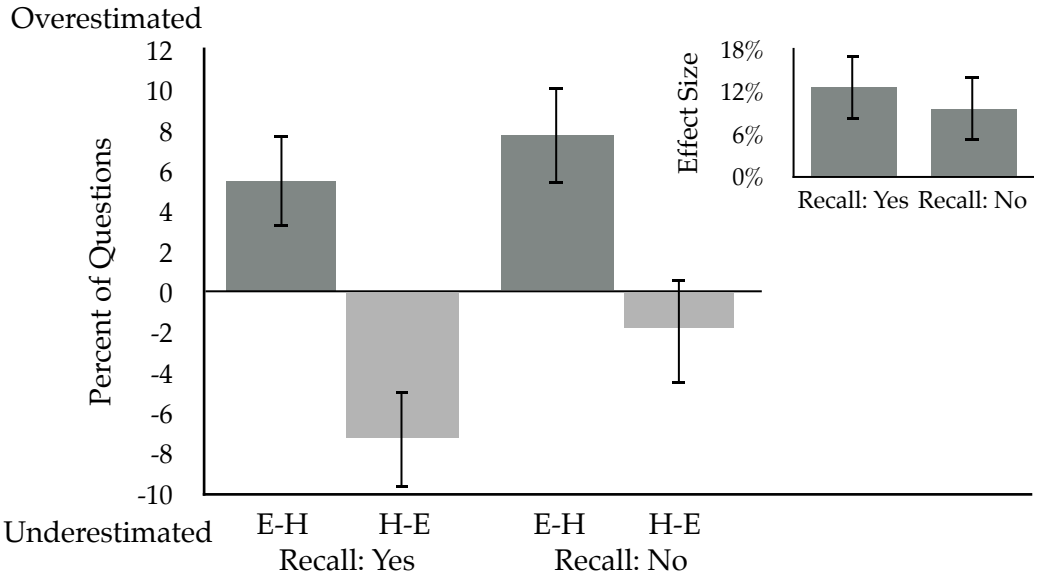


Figure 3. The mean difference, in percentage, between the number of questions people estimated they answered correctly and the actual number of questions they answered correctly. Inset axes: The mean size of the question order bias, between people who recalled five questions before estimating and people who did not. All error bars represent 95% CIs for each cell mean.

Subjects who noticed the arrangement of their test questions. A critic might wonder if the question order bias depends on whether or not students are aware that the test is arranged by difficulty. Perhaps if people were aware of how the test was arranged at the time of estimation, they would be able to account for this potential source of bias and make an accurate estimate of their performance. We did not forewarn any of our subjects about the arrangement of

the tests in this study, but 93 subjects (21%) accurately reported noticing that their test was arranged by difficulty. Forty-five subjects had taken a test that began with easy questions and 48 had taken a test that began with difficult questions. When subjects noticed that the test was arranged by difficulty, they were similarly biased compared to those who did not report noticing the test was arranged by difficulty, $F(1, 443) = 1.84$, $p = 0.18$. Reporting having noticed the question order did not interact with whether or not subjects were asked to recall five questions before estimating their performance, $F < 1$. In addition, the three-way interaction between recall, question order, and whether subjects reported noticing the question order was nonsignificant, $F < 1$.

But subjects who accurately reported noticing that their test was arranged by difficulty made more pessimistic performance estimates than subjects who did not report noticing the way their test was arranged, M_{diff} = 4.10%, 95% CI [1.06, 7.16]. In NHST terms, subjects who reported noticing the way their test was arranged made significantly more pessimistic performance estimates (M = -2.04%, SD = 12.84) than subjects who did not report noticing the way their test was arranged (M = 2.20%, SD = 15.76), $F(1, 443) = 7.02$, $p = 0.01$. Perhaps these subjects held a naïve theory that when a test is arranged by difficulty, regardless of whether the difficulty ascends or descends throughout the test, that test becomes more difficult overall.

Accuracy of recall. The task of correctly recalling each of the last five questions on the test proved to be difficult. Easy-Hard subjects recalled a mean of 53% (SD = 21%) of the last five questions from their test and Hard-Easy subjects recalled 56% (SD = 20%), M_{diff} = 3% , 95% CI [-9%, 2%]. In NHST terms, these two means did not differ significantly $t(220) = 1.97$, $p = 0.23$. This difficulty could possibly explain why subjects in the Recall condition made more pessimistic estimates than subjects in the No Recall condition. Perhaps subjects misattributed their experience of retrieval difficulty as informative to their estimate of performance, as people tend to do for judgments of frequency and confidence (see Alter & Oppenheimer, 2009).

One alternative explanation for our results is that not enough subjects could correctly constrain their recall to only the questions from the end of their tests. If this scenario is true, our manipulation might not have made the end of the test sufficiently salient to reduce subjects' reliance on their initial impressions. We tested this possibility by determining whether subjects who were better at constraining their recall to questions from the end of their test had less-extreme bias scores than those who were poor at constraining their recall. We ran an ANCOVA with bias scores as the dependent variable, the order of test questions as a between-subjects factor, and the percentage of questions that subjects correctly recalled from the last block of the test as a covariate. We found no evidence to support this alternative explanation. Those who were better and those who were worse at constraining their recall to questions from the end of the test showed no systematic differences in the magnitude of their bias scores regardless of test order, $F(3, 218) = 2.17$, $p = 0.14$. In addition, the percentage of questions subjects correctly recalled from the end of their tests had no main effect on bias scores either, $F(1, 220) < 1$.

Spontaneous discounting of availability. Some of the results from Experiments 1a-d and 2 open the door to another memory-based explanation for the question order bias. Consider the fact that Experiments 1a-d showed there was a clear trend for subjects to recall the most recent test questions first. In other words, these questions were likely to be the most available for recall. Students might know those questions are only easy to recall because of their recency, not their frequency of occurrence. This knowledge could cause students to spontaneously discount the availability of recent test questions and over-correct their estimates of performance (Oppenheimer, 2004). In general, when people can attribute the availability of exemplars to a source unrelated to the judgement they are making, they don't use that availability to inform the judgement (Schwarz et al., 1991). If students realise certain questions are only highly available due to their recency, they might discount the information those questions would provide for an availability heuristic. Depending on whether

the questions being discounted are easy or difficult, students would then make correspondingly pessimistic or optimistic estimates of performance on the test as a whole.

But one result from Experiment 2 does not fit with this spontaneous discounting explanation. If students over-correct their estimates when they discount the availability of recent test questions, we should have seen more subjects over-correct when the source of that availability is made even more salient. As a result, the question order bias should have been larger for subjects whom we instructed to recall the last 5 questions on the test before they estimated their overall performance. But we did not find that pattern. Therefore, the question order bias is unlikely to be caused by students spontaneously discounting the availability of recent test questions.

Summary of Experiments 1-2

In the first two experiments, we considered a memory-based explanation for the question order bias: that people estimate their performance by recalling individual questions from the test, but the set of questions recalled contains a disproportionate number of early questions. In Experiment 1, we found no evidence that early questions are especially available for recall—in fact, later questions were remembered best and first. In Experiment 2, we showed that even drawing subjects' attention to the last few questions before they estimated their performance did not change the size of the question order bias. Taken together, these results suggest the question order bias is not driven by students' memory for individual questions on the test.

The best explanation for the question order bias remains the idea, first proposed by Weinstein and Roediger (2010, 2012), that students form an initial impression about their performance—one that reflects their experiences as the test begins, and one from which they do not sufficiently adjust to account for later questions. Our results from Experiment 2 suggest that it may be difficult to encourage students to ignore these unhelpful first impressions. In fact, sticky, stubborn first impressions have a long and venerable history in social

psychology (Hastie & Park, 1986; McConnell, Sherman, & Hamilton, 1997). But do students always rely on their impressions when they estimate their performance on a recent test? That approach makes sense as a way for students to quickly aggregate and use a large amount of relevant information from a large number of questions. But do students still use this approach when there are only a few questions to consider?

Chapter 3: Test Length as a Potential Boundary Condition

The length of a test may be a factor that predicts whether students will display the question order bias. On the one hand, primacy effects in impression formation are found for lists of target-relevant information that are as short as five items long (Anderson, 1973). On the other hand, there is little uncertainty about students' overall performance when there are only a few questions to incorporate and evaluate. It might be sufficiently easy and more accurate for students to thoroughly recall each test question in turn and evaluate whether they think they answered the question correctly or not, tallying as they go. That is, students may instead adopt a memory-based approach for short tests. If so, the arrangement of test questions shouldn't matter—assuming it was easy for students to accurately remember all questions on the test. Everyone should make similar estimates of performance.

Until this point, the question order bias has only been examined for somewhat lengthy tests ranging from 24 to 100 questions long(Feldman & Bernstein, 1977; Weinstein & Roediger, 2012). In Experiment 3, we measure the size of the question order bias for tests that range from 3 to 50 questions.

Experiment 3

Method

Subjects. Based on our pilot studies and past exclusion rates, we aimed to recruit 420 subjects[14]. Before exclusions, we recruited 431 subjects on MTurk to participate in exchange for $0.50[15].

---

[14] We ran a different pilot study for each of Experiments 3, 4, and 5. These pilot studies collected roughly 20 participants per cell. We then identified a key comparison between cells that we wanted to make. Based on the margin of error of the confidence interval required to reliably detect that effect with 99% confidence and the sample standard deviation from that study, we estimated the number of subjects we would need in those cells to reach that level of precision (Cumming, 2012).

[15] Thirty two people did not complete the entire experiment and we therefore did not include them in our data analyses.

Design. We used a 2 (Test item order: Easy-Hard, Hard-Easy) X 4 (Test Length: 50 Questions, 25 Questions, 10 Questions, 3 Questions) between subjects design.

Materials and procedure. To begin, we gave subjects the same instructions as in Experiment 2, but altered to them refer to the appropriate number of questions. Next, subjects began their test. Tests were either arranged from the easiest to the most difficult question or vice versa (according to the Nelson & Narens, 1980 norms). The 50 question test was same as Test A, used in Experiment 2. Each other test contained the same range (97% to 0.40%) and comparable mean levels of accuracy. The 50 question test had a mean normed accuracy of 48%, the 25 question test had 46%, the 10 question had 43%, and the 3 question test had 49% (see Appendix B for the questions and answers for each remaining test). After their test, subjects estimated the number of questions they had answered correctly. Finally, we asked them to "Please tell us a little bit about how you made your estimation of how many questions you answered correctly on the test." Subjects responded by filling in a blank text box.

Results and Discussion

Before turning to our primary question, we excluded 112 subjects failing the IMC and/or indicating that they had looked up the answer to at least one of the questions on the test[16]. These exclusions left 319 subjects for analysis. See Appendix E for details about demographics.

Did subjects show the question order bias even after short tests? To answer this question, we compared the difference in subjects' bias scores between those who took an Easy-Hard test and those who took a Hard-Easy test. These results appear in Figure 4. Consistent with the idea that people use a recall-based strategy rather than an impression-based strategy for estimates on shorter tests, the question order bias (difference between Easy-Hard and Hard-Easy subjects' bias scores) decreased as the number of test questions decreased.

---

[16] The pattern of exclusions did not vary depending on test order X^2 (1, N = 112) = 7.19, p = 0.21.

The question order bias was trivial, and even slightly reversed for the 3 question test. In NHST terms, people's performance-corrected estimates after an Easy-Hard test were significantly more optimistic than after a Hard-Easy test for each test length (p_{50} = 0.01; p_{25} = 0.01; p_{10} = 0.05) except for three questions long, $t(311) = 0.32$, $p = 0.75$.

Our qualitative data were also consistent with the idea that people use a different strategy to estimate their performance on short tests than on long tests. We asked subjects to explain, in their own words, what strategy they used to estimate their performance. We were particularly interested in subjects who mentioned they used a strategy that might be seen as reliant on an overall impression or as reliant on a more recall-based strategy such as retrospectively tallying questions or using the availability heuristic. A volunteer, blind to condition, coded each response into one of five categories: Impression/Guess, Retrospective Tally, On-Line Tally, Irrelevant/Vague and Other (see Appendix C for examples of responses that were coded into each category).

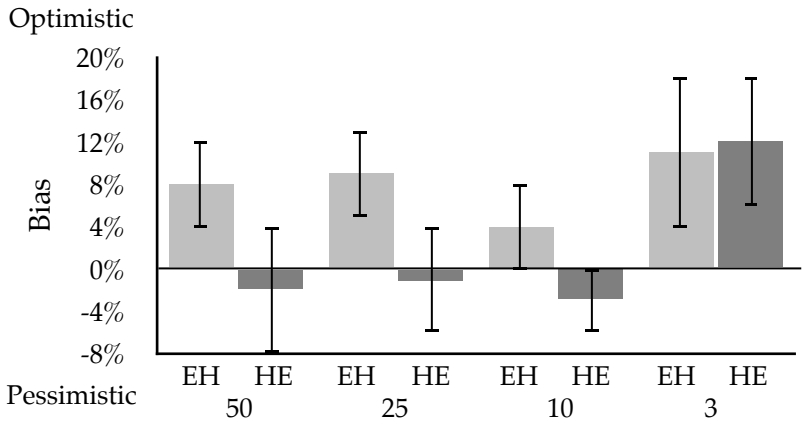


Figure 4. The mean difference between the number of questions people estimated they answered correctly and the actual number of questions they answered correctly, standardised as a percentage of the total number of questions on the test. All error bars represent 95% CIs for each cell mean.

As can be seen in Table 3, subjects were clearly more likely to report using a strategy that relied on a guess or an overall impression about their performance when the test had more questions. Conversely, subjects reported using a retrospective, memory-based approach much more often for estimations

about the short tests, especially for the 3-question test. In NHST terms, we found that people's self-reported strategies were not equally distributed across the tests of each different length, X^2 (12, N = 319) = 67.01, $p < 0.01$[17].

We also compared the size of the question order bias between subjects who reported using a "Retrospective" strategy an those who reported relying on an "Impression/Guess." Subjects who reported using a "Retrospective" strategy showed a very small question order bias (M_{diff} = 3.10%), 95% CI [-2.60%, 8.77%]. By contrast, subjects who reported relying on an "Impression/Guess" showed a substantial question order bias (M_{diff} = 16.65%), 95% CI [6.78%, 26.52%]. In NHST terms, there was a significant Test order by self-reported strategy interaction, F(1, 141) = 5.50, p = .02.

Table 3. The percentage of people who used each self-reported strategy

Strategy Type	**3 Questions (Expected)**	**10 Questions (Expected)**	**25 Questions (Expected)**	**50 Questions (Expected)**
Impression/ Guess	7.32 (22.87)	24.44 (25.11)	37.97 (22.04)	45.59 (18.97)
Retrospective Tally	78.05 (39.59)	50.00 (43.45)	31.65 (38.14)	29.41 (32.83)
On-line Tally	0 (2.31)	5.56 (2.54)	3.80 (2.23)	3.33 (1.47)
Irrelevant/Vague	12.2 (15.68)	16.67 (17.21)	26.58 (15.11)	22.06 (13.00)
Other	2.44 (1.54)	3.33 (1.69)	0 (1.49)	1.47 (1.28)

It is also important to note that subjects tended to overestimate their performance on the 3-question test—regardless of the order of questions (Figure 4). This pattern could be a byproduct of the strategy subjects used to make their estimate. If subjects thoroughly recalled all three questions from the test and retrospectively evaluated their performance on each, it would be expected that the average subject believes they answered one (easy) question correctly, and

[17] Some of the cells in our table had no subjects. Because of this missing data, the Chi-squared test we ran may be invalid. To account for this problem we ran a second analysis in which we only compared the patterns of responses between people who used the "Retrospective Tally" and the "Impression/Guess" strategies. Again, we found that people's self-reported strategies were not equally distributed across the tests of each different length, X^2 (3, N = 243) = 49.39, $p <$ 0.01.

one (difficult) question incorrectly. The only unknown would be the question of medium difficulty. Subjects would be more likely to judge that they answered the medium question correctly because generally, people err on the side of optimism for those types of judgements (Metcalfe, 1998; see Metcalfe & Shimamura, 1994 for a review). Such an error could produce the overall pattern of optimism we found for subjects who took 3-question test.

Taken together, the data from this experiment suggest students do not show the question order bias for short tests. When there are fewer questions on the test, there is less uncertainty about students' performance on each question once the test is completed. When there is less uncertainty about performance, students are less likely to use a composite impression about the test as a whole to estimate their performance. Instead the results of Experiment 3 suggest students use a more memory-based strategy such as recalling the questions from their test and retrospectively tallying their performance. Such a strategy would be relatively easy to perform for short tests. When students recall and retrospectively tally their performance on all test questions, the question order bias is not formed. Instead, students seem to overestimate their performance on the "medium" question, and accurately tally their performance on the easy and difficult ones—regardless of the order they were presented on the test. Therefore, educators do not need to worry about the question order bias when they administer very short tests.

Chapter 4: Warnings and the Question Order Bias

Of course, it is impractical to expect educators to solely rely on short tests to assess their students' knowledge. We therefore investigated another, more practical way to accomplish the goal of reducing the question order bias—simply warning students that their test is arranged by difficulty.

In order to make use of such a warning to debias their judgements of performance, students must a) be made aware of the potential for bias, b) know the direction of the bias, c) have an estimate of the magnitude of the bias, and d) have the ability to effectively apply this information (Wilson & Brekke, 1994). Because many impressions are formed both unintentionally and outside of awareness, the likelihood of spontaneously recognising the potential for bias is perhaps quite minimal. A warning can make students aware of the fact that self-evaluation of their performance may be biased by the structure of the test itself. Such a warning should prompt attempts to prevent or correct for the potential bias.

There are, however, several ways in which these attempts may go awry. When people adjust their judgements to account for a potential bias, they do so based on their own naïve theories about its magnitude and direction (Petty & Wegener, 1993). But these naïve theories may or may not be accurate. If the warning is not specific enough about the magnitude or direction of the question order bias, students could overcompensate, or worse, make estimates that are even more biased (Lombardi, Higgins, & Bargh, 1987; Strack, Schwarz, Bless, Kubler, & Wanke, 1993). As a result, students who are warned before an Easy-Hard test may end up even more optimistic than they were without the warning or they may overcorrect to the point of being pessimistic about their performance. The same logic, in reverse, can be applied to students who are warned before a Hard-Easy test.

Even if a warning were specific about both direction and magnitude, students may not be able to apply that information effectively. Impression formation involves both automatic and controlled components (McCarthy &

Skowronski, 2011). Even when the more controlled components of impression formation are disrupted, the automatic components still result in an impression —albeit a much weaker one—being formed (Crawford, et al., 2007; Wells, Skowronski, Crawford, Scherer, & Carlston, 2011). Thus, if students cannot exert good control over impression formation or subsequent adjustment, warnings could have little-to-no effect over the size of the question order bias. We already know warnings are ineffective for helping people adjust from irrelevant starting points to make more accurate numerical estimates (Epley & Gilovich, 2005, 2006; Wilson, Houston, Etling, & Brekke, 1996). There are also many other examples of people's failures to discount irrelevant information when making judgments about a target even after having seen a warning (see Wilson & Brekke, 1994 for a review).

Taken together, it is unclear whether warning students about the question order bias will cause them to make less-biased estimates of their test performance. To investigate this issue, we measured the extent to which a warning changes the magnitude or direction of this question order bias in Experiment 4. Then, we provided some evidence for the mechanism behind this change in Experiment 5.

Experiment 4

Method

Subjects. Based on a pilot study, we aimed to collect 500 subjects (125 participants for each between subjects cell). A total of 509 subjects finished the experiment, 55 from Victoria University of Wellington participated online; 128 participated in-person for course credit, and 326 from MTurk received $0.50[18] [19].

Design. We used a 2 (Test order: Easy-Hard, Hard-Easy) X 2 (Instructions: Warning, No Warning) between subjects design.

---

[18] One hundred and eight online subjects from Victoria University of Wellington and 41 MTurk subjects did not complete the entire experiment and were therefore not included in our analyses.

[19] Subjects from each source did not appear evenly in each experimental condition, but the source of each subject did not significantly interact with either factor (all p's > 0.34).

Materials and procedure. First, we gave half the subjects the same initial instructions as subjects who took a 50 question test in Experiment 3. We told the other half of the subjects the same information, but also included a warning that stated:

> The 50 questions on this test are arranged in a very specific way. The test begins with easy [difficult] questions, and then gradually the questions become more difficult [easier]. Therefore, the questions change over the course of the test from easy [difficult] questions to "medium" questions and then to difficult [easy] questions. Research shows that when questions are arranged in this way, the arrangement influences how people think they will do on the whole test. So as you answer the first few questions on this test, you may find yourself forming an impression of what the whole test will be like. Don't do that. Remember, the test starts with easy [difficult] questions, then progresses to "medium" questions, and then to difficult [easy] questions. That means your first impression may not give you good information about how well you will answer all the questions on the test.

All subsequent procedures were identical to the "No Recall" group from Experiment 3.

Results and Discussion

We excluded 137 (27%) subjects for failing the IMC and/or indicating that they had looked up the answer to at least one of the questions on the test[20]. These exclusions left 372 subjects for analysis. See Appendix E for details about demographics.

Performance. We first verified that neither test order nor the presence of a warning affected subjects' actual performance on the trivia test. Table 4 shows that subjects answered nearly half of the questions correctly for all combinations of test orders and warnings. In NHST terms, there were no

---

[20] The pattern of exclusions did not vary depending on test order X^2 (1, N = 137) = 0.56, p = 0.46.

significant main effects or interactions between Test order and Warning on the number of questions people answered correctly on their tests, all Fs < 1.

Table 4. Subjects' mean performance on the trivia test represented by the percentage of correct answers. The numbers in brackets represent 95% confidence intervals of each cell mean.

Test Order	No Warning	Warning
Easy-Hard	47.7% [44.76%, 50.64%]	49.60% [45.66%, 53.52%]
Hard-Easy	49.22% [45.3%, 53.16%]	49.08% [44.62%, 53.56%]

Question order bias. We now turn to our primary question: how effective were warnings in reducing the question order bias? To answer this question, we compared the difference between the bias scores from Easy-Hard subjects and Hard-Easy subjects based on whether or not the subjects had received a warning. The results appear in Figure 5. As the figure shows, although the question order bias was reduced, the warning did not operate symmetrically between the question orders. That is, subjects who took a Hard-Easy test were much less pessimistic when they received a warning than when they received no warning (M_{diff} = 4.36%), 95% CI [0.50%, 8.20%]. By contrast, subjects who took an Easy-Hard test were only slightly less optimistic (M_{diff} = 2.48%), 95% CI [-1.16%, 6.12%]. In NHST terms, there was a significant Test order by Warning interaction, $F(3, 368) = 6.42$, $p = .01$.

What might have caused this lop-sided influence of warnings? One possibility is a self-serving feature of the fundamental attribution error (Ross, 1977). People are more likely to attribute positive experiences to their own, internal factors, but they are more likely to explain negative experiences by seeking out third-party, external factors. (Pyszczynski, Greenberg, and LaPrelle, 1985, see Mezulis, Abramson, Hyde, & Hankin, 2004). In other words, Easy-Hard subjects may have resisted abandoning the idea that they had actually performed well, and therefore hardly heeded the warning. By contrast, Hard-Easy subjects may have readily abandoned the idea that they actually

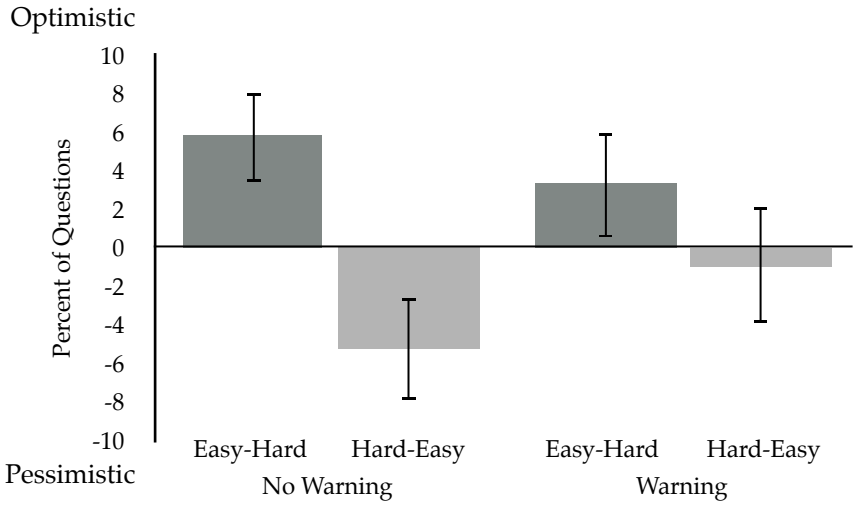


Figure 5. The mean difference between the number of questions people estimated they answered correctly and the actual number of questions they answered correctly, displayed as a percentage of the total number of test questions. All error bars represent 95% CIs for each cell mean.

performed poorly in favour of the idea that the arrangement of the questions made them think they performed worse than they really did.

But it is still unclear exactly how and when subjects may have used the information in the warning. There were a few points during which they could have reacted. For example, subjects may have responded immediately and discounted the feelings of ease or difficulty provided by the first few questions, thus setting a more balanced "central" starting point for their impressions (see Schwarz, 2011 for a review). Alternatively, subjects could have responded to the warning throughout the course of the entire test. After seeing a warning, they may have concentrated on sufficiently adjusting from their extreme initial experience of ease or difficulty (Epley & Gilovich, 2006). Warned subjects may have adopted a less conservative criterion for how much to adjust their impression as they took the test, which resulted in larger adjustments and therefore less extreme estimates of performance (LeBoeuf & Shafir, 2009). Subjects also could have made use of the warning at the end of the test, when they were constructing their estimates of performance. Once they formed their estimate, they may have retrospectively adjusted it up or down depending on their understanding of what the warning suggested (Epley & Gilovich, 2005;

Wilson & Brekke, 1994). Subjects could have made use of the warning at any one or a mix of these time points in an attempt to form an accurate estimate of their performance.

In Experiment 5, we investigated the time point at which subjects heeded the warning about the question order bias. We asked subjects to predict their overall performance after having encountered only a few questions. If subjects who are warned make similar predictions regardless of whether they had just answered several easy questions or several hard questions, it would suggest a warning helped them discount their initially extreme experience of ease or difficulty and set a less-biased impression even after just the first few questions. If subjects predict markedly different performance regardless of whether they are warned, it would suggest the warning's effects took place some time after subjects formed their initial impressions—such as on-line throughout the rest of the test, or after the test was complete.

Experiment 5

Method

Subjects. Based on a pilot study, we aimed to collect 200 subjects (50 for each between subjects cell). Two hundred and two people completed the entire experiment. Twenty two subjects from Victoria University of Wellington participated for course credit and 180 subjects from MTurk participated in exchange for $0.20[21].

Design. We used a 2 (Test question order: Easy-Hard, Hard-Easy) X 2 (Instructions: Warning, No Warning) between subjects design.

Procedure. We used the same procedure as in Experiment 4, but instead of subjects having taken the entire 50 question test before estimating their performance, they predicted their overall performance after just the first five questions. We compared these predictions to investigate the extent to which

[21] Three people from MTurk and 10 people from Victoria University of Wellington did not complete the entire experiment and we therefore did not include them in our data analyses.

warnings about the question order bias can change the way people form their initial impressions about the test they are taking.

Results and Discussion

Before turning to our primary question, we excluded 40 (20%) subjects for failing the IMC and/or indicating that they had looked up the answer to at least one of the questions on the test[22]. These exclusions left 162 subjects for analysis[23]. See Appendix E for details about demographics.

To what extent did warning subjects about the question order bias help subjects to form less extreme initial impressions about their test? As Figure 6 shows, the answer depended on the order of questions. That is, the warning helped Hard-Easy subjects more than Easy-Hard subjects. Hard-Easy subjects predicted markedly better performance after five questions when they saw a warning than when they did not, (M_{diff} = 14.02%), 95% CI [4.90%, 23.18%], and Easy-Hard subjects predicted only somewhat lower performance (M_{diff} = 7.06%),

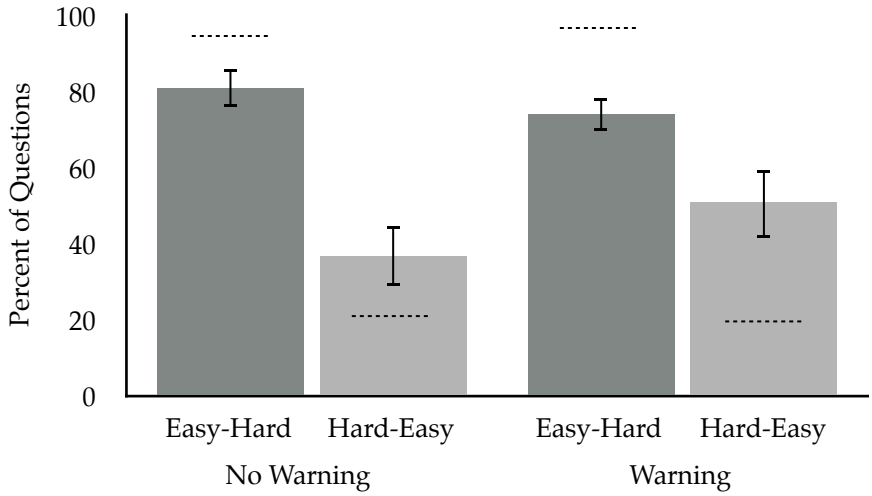


Figure 6. The mean number of questions people predicted they would answer correctly on the test after having answered the first five questions, expressed as a percentage of the overall number of questions. Dotted lines represent actual mean performance for subjects in each cell. All error bars represent 95% CIs for each cell mean.

[22] Subjects were more likely to be excluded for failing the IMC in the Easy-Hard conditions, but were more likely to be excluded for cheating in the Hard-Easy condition, X^2 (1, N = 40) = 10.25, $p < 0.01$.

[23] After exclusions, between 2 and 4 subjects from Victoria University of Wellington appeared in each condition.

95% CI [-1.54%, 15.64%]. In NHST terms, there was a Warning x Test order interaction, $F(3, 158) = 11.03$, $p < 0.01$. This asymmetry can also be explained as an effect of self-serving biases in the same way that people's global estimates of performance can be explained.

On the whole, the results of Experiment 5 are consistent with the idea that a warning reduces the question order bias by helping students form a centralised initial impression of the test that requires less adjustment than it would if there was no warning. Despite this evidence, these results do not rule out the possibility that students who are warned also correct for the question order bias at later time points such as throughout the remainder of the test or once the test is completed.

Summary of Experiments 4 and 5

In Experiments 4 and 5, we evaluated a practical way to mitigate the question order bias even on a lengthy test—forewarning. Warnings can only effectively help students change their judgements of performance if they have control over them in the first place and know how they should change them. Taken together, the results from Experiments 4 and 5 suggest people do have some control over how they form and adjust their impressions about the test and that this control translates to less extreme judgements of performance. With the help of a warning, students can form a less extreme starting point from which to adjust their impressions of a test as it progresses. As a result, when they see a warning that alerts them to the fact their tests are arranged in a potentially biasing way, those who take an Easy-Hard test can make estimates of performance that are more similar to those who take a Hard-Easy test.

The warning reduced the question order bias, and it was particularly effective for subjects who took a Hard-Easy test. But it did not result in all subjects estimating equal performance on average. Why not? Consider that the warning contained no information about the magnitude of the bias. Without this information, subjects were left to guess how much to adjust, and in doing so, may have drawn on their own naïve theories and motivational biases (Petty

& Wegener, 1993). Research on self-serving attributions suggests that after the first few questions on the test, Hard-Easy students should be motivated to attribute their initial experience of difficulty to something external to their self, (i.e. the structure of the test) whereas Easy-Hard students will be motivated to attribute their initial experience of ease to something internal (i.e. their own abilities; Pyszczynski, et al., 1984, see Campbell & Sedikides, 1999 and Mezulis et al., 2004). The warning provides a convenient external scapegoat for Hard-Easy students, but Easy-Hard students need no such scapegoat. Thus, Hard-Easy students should be able to discount their initial experiences of extreme difficulty and more easily adopt centralised, "big picture" initial impressions of the test overall. By contrast, even though Easy-Hard students may recognise that there is the potential for bias based on the structure of the test, their initial impressions may still be fairly biased towards optimism because they should be unmotivated to attribute their initial experience of ease to the structure of the test itself.

Also consider that there are both automatic and controlled components involved in the impression formation process (McCarthy & Skowronski, 2011). It is possible that subjects heeded the warning to the best of their ability, but that automatic processes prevented them from fully centralising their impressions. Thus, even if we used a thorough, explicit warning that contained information about magnitude, it is possible we would find students still display the question order bias. This cognitive process and the motivational process discussed above are not mutually exclusive and could both contribute to the effectiveness of the warning used in Experiments 4 and 5.

Across the previous five studies, we replicated and explored the mechanism behind the finding that Easy-Hard students are more optimistic about their test performance than Hard-Easy students. Overall, this question order bias seems robust as researchers have detected it across a variety of experimental designs: between and within subjects; presentation media: paper and pencil, digital; and test formats: multi-choice, and cued recall (Jackson &

Greene, 2014; Weinstein & Roediger, 2010, 2012). We also provided some evidence that Easy-Hard students and Hard-Easy students do not differ much in how well they actually perform. These findings contribute to the overall debate about whether ordering questions on a test affects students' performance (Aamodt & McShane, 1992; Jackson & Greene, 2014; Vander Schee, 2013; Weinstein & Roediger, 2010, 2012). But given the recent focus in psychology on departure from NHST and its new focus on replication and estimation (see the November 2012 *Perspectives on Psychological Science*, the February 2012 *Observer*, www.psychfiledrawer.org; Cumming, 2012; Michael, Newman, Vuorre, Cumming, & Garry, 2013), we will establish a more precise estimate of the size of the question order bias and the effect of question order on performance than can be inferred from any individual experiment. To accomplish these goals, we performed a mini meta-analysis of several published and unpublished studies that compared subjects' actual performance on and subjective estimates of performance after taking either an Easy-Hard or a Hard-Easy test.

Chapter 5: Meta-Analysis of the Question Order Bias

Here we report a small-scale meta-analysis that synthesises the studies conducted over the course of my thesis to investigate an important question: How much does the order of questions on a test affect people's actual and estimated performance on a test? This type of small-scale meta-analysis is useful for obtaining a precise estimate of the size of an effect over in a single line of research (Cumming, 2012). It is not intended to be exhaustive, and we did not search for studies to include outside of those that were conducted over the course of my thesis.

Method

Subjects. Across the 15 studies that make up this meta-analysis, 2256 subjects were analysed.

Design. Each experiment in this meta-analysis compared two groups (Test order: Easy-Hard, Hard-Easy) between subjects.

Materials and procedure. For all experiments in this analysis, subjects took tests comprised of cued-recall trivia questions that were normed by Nelson and Narens (1980). These questions ranged from 97.40% correct to .004% correct. This percentage functioned as a proxy for the experience of ease or difficulty. Questions that many subjects tended to answer correctly were considered easy and questions that many subjects tended to answer incorrectly were considered difficult. Across studies, the number of questions on the trivia tests ranged from three to 50 (see Table 5). Subjects took a test with questions that were arranged from the easiest to the most difficult or from the most difficult to the easiest. Directly after their test, subjects estimated the number of questions they answered correctly.

Table 5. Information about the subjects and test for each study in the mini meta-analysis.

Study Name	Source of Subjects	Medium	N	Notes
Franco & Garry: MTurk Replication (unpublished)	MTurk	Online	95	50 Trivia Questions
Franco & Garry: MTurk Replication 2 (unpublished)	MTurk	Online	70	50 Trivia Questions
Franco & Garry: MTurk Replication 3 (unpublished)	MTurk and Victoria University of Wellington Undergraduates	Online	191	50 Trivia Questions
Franco & Garry: Test Length (unpublished)	MTurk	Online	68	50 Trivia Questions
Franco & Garry: Test Length (unpublished)	MTurk	Online	79	25 Trivia Questions
Franco & Garry: Test Length (unpublished)	MTurk	Online	90	10 Trivia Questions
Franco & Garry: Test Length Replication (unpublished)	MTurk	Online	151	10 Trivia Questions
Franco & Garry: Test Length Replication 2 (unpublished)	MTurk	Online	108	10 Trivia Questions
Franco & Garry: Test Length (unpublished)	MTurk	Online	82	3 Trivia Questions
Franco & Garry, Recall Last 5 (unpublished)	MTurk	Online	223	50 Trivia Questions; No Recall Condition
Franco & Garry, Recall Last 5 (unpublished)	MTurk	Online	222	50 Trivia Questions; Recall Condition
Franco & Garry: Exhaustive Recall (unpublished)	MTurk	Online	245	50 Trivia Questions; No Recall Condition
Franco & Garry: Exhaustive Recall (unpublished)	MTurk	Online	260	50 Trivia Questions; Recall Condition
Franco, Crawford, & Garry: Warnings, exp 1 (ms in prep)	MTurk and Victoria University of Wellington Undergraduates	Online and In-person	194	50 Trivia Questions; No Warning Condition
Franco, Crawford, & Garry: Warnings, exp 1 (ms in prep)	MTurk and Victoria University of Wellington Undergraduates	Online and In-person	178	50 Trivia Questions; Warning Condition
TOTAL			2256	

For all experiments that were conducted online, after the experiment proper concluded, subjects to responded to a series of questions designed to identify those who who did not take the task seriously, as well as those who had not complied with the instructions. To identify people who did not take the task seriously, we used an IMC. To identify subjects who did not comply with the instructions, we used the same “catch-out” question as we described in Experiment 1. Only subjects who chose "I answered each question from my own memory" were included in our analyses.

Results and Discussion

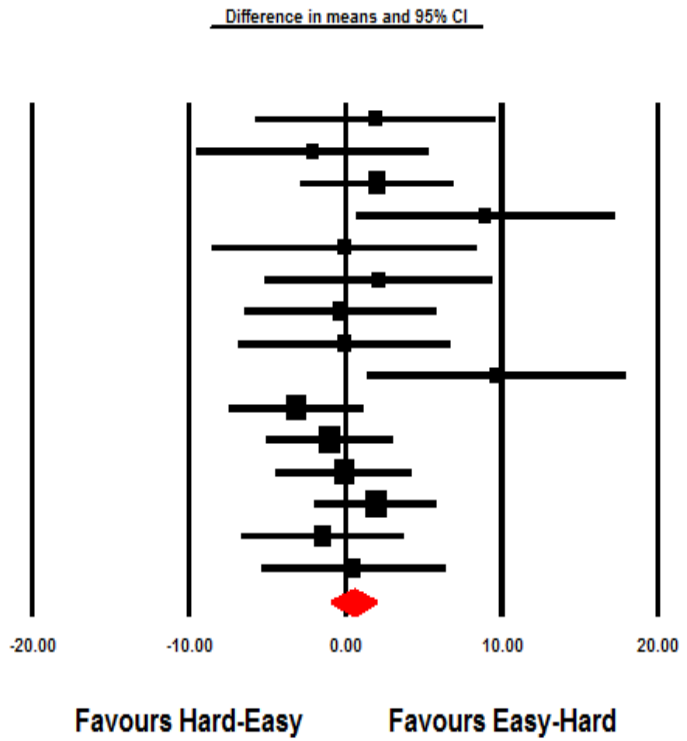
Scoring criteria. For all studies, we marked questions according to the same formula as described in Experiment 2. Consistent with Experiments 2-4, we operationalised the question order bias as a measure of the difference between subjects' bias scores after an Easy-Hard test and those after a Hard-Easy test.

Number correct. We compared the number of questions answered correctly by subjects who took an Easy-Hard test and subjects who took a Hard-Easy test. Figure 7 contains the size of the difference between groups as expressed by a percentage of questions on the test, the standard error, variance, 95% confidence intervals of the effect size, a *z*-score for a comparison between groups, and a *p*-value for each individual comparison in each study and for the meta-analysed random effect.

Overall, subjects who took an Easy-Hard test performed only slightly (although plausibly not at all) better than subjects who took a Hard-Easy test, M_{diff} = 0.48%, 95% CI [-0.99, 1.95]. This low estimate is consistent with recent claims that the order of questions on a test does not greatly influence people's performance (Jackson & Greene, 2014; Vander Schee, 2013; Weinstein & Roediger, 2010, 2012). We found no significant evidence of heterogeneity in our model, $Q(14) = 14.66$, $p = 0.40$.

Question order bias. Figure 8 contains the size of the question order bias (a difference in means, expressed as a percentage of questions on the test), the standard error, variance, 95% confidence intervals of the effect size, a *z*-score for a comparison between groups, and a *p*-value for each individual comparison in each study and for the meta-analysed random effect. Overall, subjects who took an Easy-Hard test made more optimistic estimates of performance than people who took a Hard-Easy test, M_{diff} = 9.10%, 95% CI [6.92, 11.29]. This difference amounts to nearly an entire letter grade in most university courses. Put another way, a student who expected a B might have actually earned a C. This estimate might be slightly conservative because we included studies that used

Study name	Statistics for each study						
	Difference in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Franco & Garry: Mturk Replication (unpublished)	1.900	3.900	15.209	-5.744	9.544	0.487	0.626
Franco & Garry: Mturk Replication 2 (unpublished)	-2.110	3.788	14.346	-9.534	5.314	-0.557	0.577
Franco & Garry: Mturk Replication 3 (unpublished)	2.020	2.475	6.125	-2.830	6.870	0.816	0.414
Franco & Garry: Test Length [50 questions] (unpublished)	8.920	4.196	17.603	0.697	17.143	2.126	0.034
Franco & Garry: Test Length [25 questions] (unpublished)	-0.080	4.298	18.476	-8.505	8.345	-0.019	0.985
Franco & Garry: Test Length [10 questions] (unpublished)	2.100	3.700	13.686	-5.151	9.351	0.568	0.570
Franco & Garry: Test Length Replication [10 questions] (unpublished)	-0.300	3.120	9.735	-6.415	5.815	-0.096	0.923
Franco & Garry: Test Length Replication 2 [10 questions] (unpublished)	-0.100	3.438	11.820	-6.838	6.638	-0.029	0.977
Franco & Garry: Test Length [3 questions] (unpublished)	9.660	4.216	17.775	1.397	17.923	2.291	0.022
Franco & Garry: Recall Last 5 [No recall] (unpublished)	-3.180	2.164	4.683	-7.421	1.061	-1.469	0.142
Franco & Garry, Recall Last 5 [Recall] (unpublished)	-1.040	2.038	4.153	-5.034	2.954	-0.510	0.610
Franco & Garry: Availability [No recall] (unpublished)	-0.100	2.205	4.864	-4.423	4.223	-0.045	0.964
Franco & Garry: Availability [Recall] (unpublished)	1.940	1.975	3.901	-1.931	5.811	0.982	0.326
Franco, Crawford, & Garry: Warnings [No Warning] (unpublished)	-1.480	2.642	6.979	-6.658	3.698	-0.560	0.575
Franco, Crawford, & Garry: Warnings [Warning] (unpublished)	0.520	2.997	8.983	-5.354	6.394	0.173	0.862
	0.479	0.748	0.560	-0.987	1.945	0.640	0.522



Meta Analysis

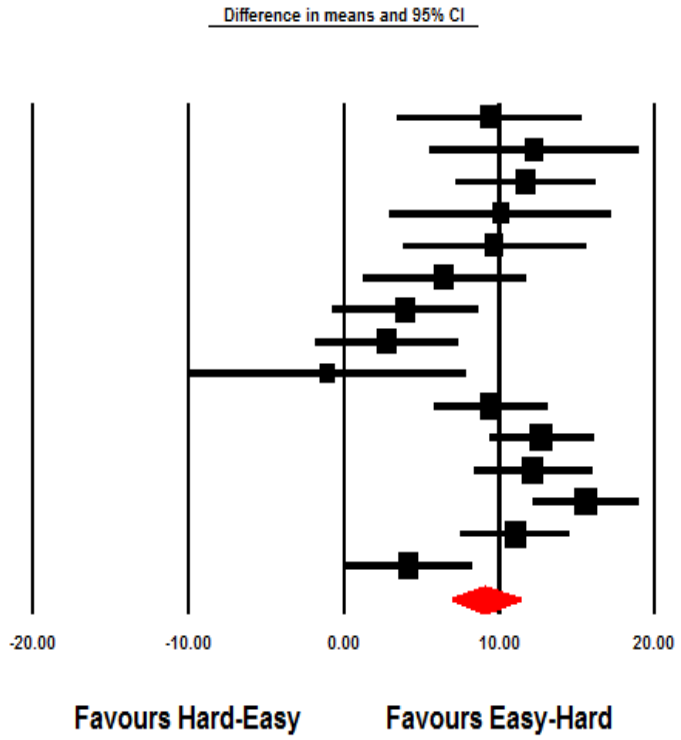
Figure 7. The overall meta-analysis for performance. The forest plot depicts the mean difference between the percent of questions people answered correctly on an Easy-Hard test and on a Hard-Easy test. A result that "favours Hard-Easy" means people who took a Hard-Easy test answered more questions correctly than people who took an "Easy-Hard" test. A result that "favours Easy-Hard" means the opposite.

manipulations meant to mitigate the question order bias—regardless of whether those manipulations were effective. For example, the bias is likely to be considerably smaller for studies that used very short tests or forewarned subjects before they began. Such studies were likely responsible for the high proportion of variance due to heterogeneity we found in our model, $Q(14) = 47.23$, $p < 0.01$, $I^2 = 70.36$.

The constrained scope of our meta-analysis limits our ability to compare our estimate of the difference in actual performance between Easy-Hard and Hard-Easy students with estimates from previous meta-analyses. Our study estimates the difference in performance to be smaller than previously estimated (Aamodt & McShane, 1992). It is possible that the studies that show a consistent advantage in performance for Easy-Hard students contain systematic differences that make the effect larger than the effect found in our studies. A comprehensive meta-analysis of both samples and a wider search for studies from the file drawer is a fruitful avenue for future research.

One related methodological limitation of this mini meta-analysis is the fact that all studies used very similar materials. The question order bias may vary in size based on the content of the test. For example, it is possible that the bias takes hold only when the questions provide a wide range in experience from extreme ease to extreme difficulty. Each test in this meta-analysis provided such a range because they were comprised of questions that were normed to do so. Perhaps tests that provide a more restricted range of difficulty might not produce extreme enough initial impressions between Easy-Hard students and Hard-Easy students. If the gap in initial impressions is not large to begin with, students would not need to adjust much to meet in the middle and provide relatively similar estimates of performance. Future research on this topic would have implications for the generalisability of the question order bias from highly normed and structured tests to non-academic testing settings such as eyewitness interviews.

Study name	Statistics for each study						
	Difference in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Franco & Garry: Mturk Replication (unpublished)	9.420	3.018	9.106	3.506	15.334	3.122	0.002
Franco & Garry: Mturk Replication 2 (unpublished)	12.300	3.419	11.687	5.599	19.001	3.598	0.000
Franco & Garry: Mturk Replication 3 (unpublished)	11.740	2.291	5.246	7.251	16.229	5.125	0.000
Franco & Garry: Test Length [50 questions] (unpublished)	10.120	3.618	13.092	3.028	17.212	2.797	0.005
Franco & Garry: Test Length [25 questions] (unpublished)	9.720	2.988	8.929	3.863	15.577	3.253	0.001
Franco & Garry: Test Length [10 questions] (unpublished)	6.500	2.642	6.982	1.321	11.679	2.460	0.014
Franco & Garry: Test Length Replication [10 questions] (unpublished)	4.000	2.380	5.662	-0.664	8.664	1.681	0.093
Franco & Garry: Test Length Replication 2 [10 questions] (unpublished)	2.800	2.329	5.427	-1.766	7.366	1.202	0.229
Franco & Garry: Test Length [3 questions] (unpublished)	-1.000	4.512	20.358	-9.843	7.843	-0.222	0.825
Franco & Garry: Last 5 [No recall] (unpublished)	9.500	1.860	3.458	5.855	13.145	5.109	0.000
Franco & Garry, Last 5 [Recall] (unpublished)	12.740	1.693	2.865	9.422	16.058	7.526	0.000
Franco & Garry: Availability [No recall] (unpublished)	12.220	1.925	3.704	8.448	15.992	6.350	0.000
Franco & Garry: Availability [Recall] (unpublished)	15.600	1.723	2.970	12.222	18.978	9.052	0.000
Franco, Crawford, & Garry: Warnings [No Warning] (unpublished)	11.040	1.767	3.121	7.577	14.503	6.249	0.000
Franco, Crawford. & Garry: Warnings [Warning] (unpublished)	4.200	2.079	4.324	0.124	8.276	2.020	0.043
	9.101	1.115	1.243	6.916	11.286	8.164	0.000



Meta Analysis

Figure 8. The overall meta-analysis for the question-order bias. The forest plot depicts the mean percent difference between the estimates of performance made by people who took an Easy-Hard test and people who took a Hard-Easy test. A result that "favours Hard-Easy" means people who took a Hard-Easy test were more optimistic than people who took an "Easy-Hard" test. A result that "favours Easy-Hard" means the opposite. *P*-Values that read "0.000" mean $p < 0.001$.

Chapter 6: General Discussion

Summary

In six studies, we investigated how the question order bias is formed, its magnitude, and some of its limits. Experiments 1a-d and 2 eliminated memory for performance on individual questions as a plausible explanation for this bias. These results are consistent with studies that show there is often little relationship between the specific exemplars people can recall and the global judgements people make about the source of those exemplars (see Hastie & Park, 1986 for a review). Instead, a combination of extreme initial impressions and insufficient adjustment from those extreme impressions remains the most plausible explanation for the question order bias (Jackson & Greene, 2014; Weinstein & Roediger, 2010, 2012).

Experiment 3 explored the length of the test as a boundary condition for the question order bias. The results of Experiment 3 suggest that students adopt a different strategy to estimate their performance after a long test than they do after a short test. After a short test, subjects were more likely to report recalling each question and retrospectively tallying their performance than they were to report using a gut-feeling or a guess. This apparent change in strategy resulted in subjects avoiding the question order bias after very short tests. The question order bias is therefore a larger concern for educators who use lengthy tests than those who rely only on very short quizzes.

In Experiment 4, we demonstrated that specific warnings about the influence question order has on people's estimates of their performance can help students make less-biased estimates, but they are much more effective for students who take a Hard-Easy test. These results suggest that students have at least some control over the information they use to estimate their performance on a test.

In Experiment 5, we provided evidence that students exert this control from the very beginning of their tests. Subjects who were warned made less

extreme predictions about their overall performance after having seen just a few questions on the test—particularly subjects who took a Hard-Easy test. These results suggest students heed the warnings by forming a less extreme initial impression of the test's overall level of difficulty. As a result, there is less distance between groups' initial impressions from which to adjust when accounting for later test questions.

Finally, our meta-analysis established a more precise estimate for how large the question order bias is and how much arranging questions on a test in order by difficulty affects students' actual performance. We found that the question order bias is large and robust—resulting in a mean difference in estimates for Easy-Hard subjects and Hard-Easy subjects of almost an entire letter-grade—but the order of questions on the test had little effect on subjects' actual performance.

The Practicality of Warnings

Because warnings can at least partially mitigate the question order bias, it might be tempting to conclude the bias is little cause for concern among educators. This conclusion would be premature. When we look closer at which groups were best at heeding our warnings, only people who took a Hard-Easy test make substantially different predictions with and without a warning. Because of this fact, the utility of a warning in a real-life classroom context is somewhat limited.

Educators often distribute multiple versions of a test to discourage cheating (Vander Schee, 2013). Suppose one version ascended in difficulty and one descended. In this scenario, educators might consider warning their students about these arrangements to reduce any relative difference in judgements of performance between their students and make the situation "fairer." Alternatively, some educators may already have a single favourite way of arranging their test questions. If their favourite arrangement is Easy-Hard, Experiment 4 demonstrates a warning would help very little to avoid their students leaving with overly optimistic beliefs about their performance.

Perceived Cohesiveness

Warnings might not be an entirely practical way to handle the question order bias in a classroom after all. What might be a better way to protect students from undue optimism or pessimism? One obvious answer is immediate feedback after the test. This approach is becoming increasingly practical because of multiple choice tests and scoring algorithms.

But what about when immediate feedback is impossible or impractical? The social perception literature suggests another promising route to unwinding the impressions that drive the question order bias: disrupting the perceived cohesiveness of the target being judged (Crawford, Sherman, & Hamilton, 2002; McConnell et al., 1997). For example, people given a set of 30 descriptors about three members of the same family would be more likely to form an impression of those people as a group than if the same descriptors were applied to three randomly selected people from anywhere in the world. In the latter scenario, people may have no previously-formed impression on which to base a judgement about the group as a whole. Instead, people would need to retrospectively construct an impression of the group based on memories about its individual members before making an overall judgement about the group (McConnell et al., 1997).

Following this logic, if students were asked to estimate their performance on a 30 question test divided into 3 blocks of 10 (written by teachers A, B, and C), we might not see the question order bias. Students who take a divided test are less likely to have formed an impression about the test as a whole. Instead, they might have several sub-impressions—one for each of blocks A, B, and C. If they are asked to estimate their performance on the test overall, they might need to retrospectively construct an impression about the ease of the test overall by recalling and incorporating their impressions for each individual block. One block was easy, one was difficult, and one was of medium difficulty. As a result, the students who take such a divided test may show no difference in estimates of performance regardless of the order of

questions. It is easy to recall and tally three impressions—much like it is easy to recall and tally three questions (see Experiment 3). If so, the implications of arranging test questions by difficulty could vary for team-taught courses compared to solo-taught courses—especially when questions are "batched" by teacher.

Drop-out Rate

The majority of studies in this thesis were conducted entirely online. One aspect of online data collection that makes it different from in-person data collection is experiments conducted online are not monitored by an experimenter. As a result, online subjects are more likely to leave the experiment before finishing, or "drop out." Subjects may also drop out for a number of additional reasons, not necessarily related to the fact the experiment is unmonitored. For example, some may drop out because of internet connectivity issues, environmental distractions, or emergencies. But subjects in our online experiments consistently failed to complete the entire test when it was arranged from Hard-Easy more often than when it was arranged from Easy-Hard.

Recall that when people take a test that begins with difficult questions, their anxiety levels raise and this increase in anxiety may negatively affect performance (Aamodt & McShane, 1992; Munz & Smouse, 1968; for a review, see Vander Schee, 2013). Suppose this increase in anxiety also made our online subjects more likely to give up on the test. If so, this explanation could account for the higher drop-out rates in the Hard-Easy condition for our online experiments.

The subjects in the Hard-Easy condition who did complete the test may have been more resilient to increases in anxiety than their counterparts who dropped out. As a result, the Hard-Easy group's mean anxiety level might have been artificially deflated due to its high drop-out rate. On one hand, this artificial deflation may at least partially explain why we only found a trivial and unreliable advantage in performance for Easy-Hard group. On the other

hand, previous in-person investigations of the question order bias also found no reliable differences in performance between test orders (Jackson & Greene, 2014; Weinstein & Roediger, 2010; 2012).

In real-world educational contexts, this pattern of drop outs may have even more dire implications. Unlike the subjects in our studies, students who don't complete their tests in school still receive marks for their performance. If students are more likely to give up when taking a Hard-Easy test in the classroom, studies conducted in labs (or online) may underestimate the effects of question order on performance. Imagine that one half of a class takes an Easy-Hard test and the other takes a Hard-Easy test. As established in Chapter 5, those students who complete the test should perform similarly regardless of test order. Unfortunately, students who take a Hard-Easy test may be more likely to give up on the test. As a result, those who took a Hard-Easy test would have a lower mean score than those who took an Easy-Hard test. Future research on students in actual educational environments should take this information into account when evaluating the effects of test order on performance.

Test Length

Another fruitful avenue for future research would be to establish more precise boundary conditions for the question order bias. For example, Experiment 3 demonstrated that people don't display the question order bias for tests that are 3 questions long, but that they do show a small and unreliable bias for tests that are 10 questions long. Future research could establish a more precise estimate of how many questions an educator could expect to include on his or her test before the question order bias becomes a concern.

One approach to answering this question is to issue eight separate tests between 3 and 10 questions long to subjects that are arranged either from Easy-Hard or from Hard-Easy. This approach would require a very large number of subjects to reliably detect the question order bias (roughly 50 per cell), so experimenters should adopt a within subjects design to conserve power. The

mini meta-analysis within this thesis demonstrates the question order bias remains of similar size regardless of whether the tests are given within subjects or between groups, so such a design should be appropriate for establishing this boundary condition.

If this research is conducted, it could help address one important limitation of Experiment 3. Namely, it is impossible to determine whether our subjects did not display the question order bias on the 3-question test because they chose a different strategy to make their estimates (i.e. a retrospective strategy) or because they simply did not encounter enough material to form a biased first impression, so they could not use the same strategy as subjects who took longer tests. Studies of impression formation in other domains typically use at least five descriptors to demonstrate primacy effects (Andersen, 1973). If subjects don't display the question order bias for tests that are five questions long, this result could be interpreted as evidence they still choose a retrospective strategy even though an impression-based strategy is theoretically available.

Predicted Future Performance

Each experiment in this thesis measured the effect of question order either on students' retrospective estimates of their performance, or prospective estimates of their performance on a test they were taking. Future research could investigate the degree to which students show the question order bias when they predict performance on a future test. It is possible that when students predict future performance, they ignore their initial impressions. Instead they may rely on how their performance changed over time and how well they did at the end because these aspects of a test seem more relevant to the future than performance at the beginning (Zauberman, Diel, & Aiely, 2006).

The evidence for this hypothesis is mixed. In one study, students who finished a test with easy questions predicted better performance in the future (Jones et al., 1968). When students took (what they believed to be) an intelligence test in either ascending or descending order by difficulty, they

predicted better future performance when they finished with many correct answers than when they finished with many incorrect answers. But another study found the opposite—a pattern more like the question order bias (Feldmen & Bernstein, 1977). When students took a geometry test during which students had to identify a previously-seen target shape hidden within a new, more complex design, they predicted better future performance when they began with many correct answers than when they finished with many incorrect answers.

One possible explanation for these conflicting patterns of results has implications for how the question order bias unfolds for predictions about different styles of test. Subjects in the Jones et al. (1968) study who took a test in ascending order by difficulty could have predicted better future performance because they believed they discovered the "trick" to answering the logical "intelligence" questions. The test used in the Feldman and Bernstein (1977) study may not have afforded the same belief. Future research could investigate this possibility by keeping the test itself constant, but manipulating how likely Hard-Easy subjects are to attribute their increase in performance to a gain in ability throughout the test. For example, experimenters could use normative feedback throughout the test to give subjects the impression that their performance is improving (or decreasing) due to the difficulty of the test or due to the possibility that the subject is getting better or worse at the task. Subjects who believe their performance is changing due to changes in their own abilities (finding or losing the "trick") should estimate future performance based on the difficulty of the most recent questions. Those who believe their performance is only changing due to changes in the difficulty of the questions themselves should show the opposite pattern—the traditional question order bias—when estimating their performance.

Actual Future Performance

Regardless of the direction of the question order bias when people predict their performance in the future, the bias has practical implications for

student's decisions about what and how long to study for an upcoming test. Undergraduates commonly take a test at the end of the semester followed by a final exam just a few weeks later. The question order bias might influence how students decide to study during this interval. On the one hand, those who leave their test believing they had performed (or will perform) relatively well might not study as hard nor as long as students who leave the test believing they performed (or will perform) poorly. On the other hand, their perceived high level of performance may give them confidence and subsequently cause them to study harder for the next test. Both of these scenarios are especially likely if students believe they are in control of their own level of learning, attribute their performance to internal factors, and are intrinsically motivated to perform well (Bandura, 1997; Dweck, 1991; Weiner, 1985, 2000).

Future research should investigate whether the arrangement of questions on one test could influence students' performance on a future test. Such an experiment could be run with very similar methods to either of the Weinstein and Roediger (2010, 2012) studies. Subjects would simply take a test arranged from Easy-Hard or from Hard-Easy in session 1. Then, subjects will be excused for a period of time (days or weeks) during which they will need to study for their second test. Finally, subjects will come back and take the second test.

Informational vs Affective Judgements

The studies comprising this thesis suggest that students form initial impressions about a test based on the first few questions and fail to sufficiently adjust those impressions to account for the remainder of the questions. Then, students estimate their overall performance on the test using this biased impression. This approach is similar to the way people make judgements about other people, products, and legal evidence (Anderson & Norman, 1964; Crano, 1977; Hastie & Park, 1986; Mantonakis et al., 2009; see Hogarth & Einhorn, 1992 for a review). But people do not adopt this approach for all judgements about the events they experience. In fact, people commonly base their judgements

about certain experiences on their peak intensity and the way the experiences end (Fredrickson & Kahneman, 1993; Hastie & Park, 1986; Jensen, Martin, & Cheung, 2005; Redelmeier & Kahneman, 1996; Redelmeier, Katz, & Kahneman, 2003; Stone, Schwartz, Broderick, & Schiffman, 2005). For example, some subjects were asked to make judgements about a colonoscopy they had just received. Half of the subjects underwent a standard procedure, and some underwent a longer procedure that ended less painfully than the standard. Subjects who underwent the longer procedure retrospectively rated the experience as being less painful and less unpleasant, despite experiencing an objectively greater amount of real-time pain during the procedure (Redelmeier, et al., 2003). Recent research demonstrates that these effects extend to judgements about how people remember and evaluate future study and test scenarios based on whether a recent scenario's end was mild or severe in difficulty (Finn, 2010; Finn & Miele, in press; Hoogerheide & Paas, 2012).

Why do people make judgements about medical procedures based on this *peak-end rule*, but make judgements about tests and people based off of their first impression? One theory predicts that people will adopt one strategy when making an *affective* judgement, but another strategy when making an *informational* judgement (see Zauberman et al., 2006). When people rate affective traits of an experience or object such as attractiveness, satisfaction, and comfort, they generally rely on the peak intensity of those traits, the intensity of those traits at the end of their encounter, and the overall trend. By contrast, when people rate evaluative traits such as intelligence, performance, and friendliness, they generally rely on their initial impressions of the target. But Weinstein and Roediger (2012) found no asymmetry in subjects' ratings of how much they were enjoying the test while they took it, regardless of whether the questions were arranged Easy-Hard or Hard-Easy. That is, subjects made a seemingly affective judgement—their level of enjoyment—but were not biased by their peak level of enjoyment or how the test ended.

How might we reconcile these conflicting results? One possible avenue comes from the exact wording of the judgement subjects made as they took the test. Experimenters asked subjects to rate "How much are you enjoying the test?" at each of 10 intervals throughout the test. This question may have elicited moment-to-moment judgements about enjoyment analogous to the *real-time evaluations* gathered in typical peak-end rule studies. (Redelmeier & Kahneman, 1996; Redelmeier et al., 2003). These moment-to-moment affective judgements do not typically differ between groups that end with high intensity and those that end with low intensity. If future studies instead ask subjects to make a global affective judgement such as "How enjoyable was this test?", they may find a dissociation between subjects who take an Easy-Hard and those who take a Hard-Easy test. Both test orders should contain the same peak intensity of enjoyment, but Easy-Hard tests should end with lower levels of enjoyment than Hard-Easy tests. If the literature on informational versus affective judgements is correct, subjects who take an Easy-Hard test should rate the test as less enjoyable overall than those who take a Hard-Easy test (Zauberman et al., 2006).

The Question Order Bias in Other Domains

The question order bias is likely not limited to educational tests. Recent research shows that the order in which eyewitnesses answer questions about an event an bias their judgements about how well they performed on the set of questions and their confidence about their performance (Michael & Garry, 2015). These results have implications in a court room. Eyewitnesses who seem more confident in their memories are more credible to juries than those who less confident (Penrod & Cutler, 1997).

Conclusion

When educators arrange their tests in order by difficulty, they may do so with benevolent intentions. Some studies show educators typically choose to arrange their tests in order from the easiest question to the most difficult question in order to decrease anxiety and help weaker students retain their

confidence through the easy questions before they encounter the difficult ones (Munz & Smouse, 1968; for a review, see Vander Schee, 2013). But this arrangement does not typically improve students' actual performance—what is worse, it fosters a positive illusion about their actual performance that will soon be destroyed when the grades are announced. It is not a leap to worry that this illusion, or its later destruction, may encourage students to change their decisions about what they study and how long they study. Therefore, it is important that we understand what causes the bias, what its limitations are, and how to reduce or eliminate it.

References

Aamodt, M. G., & McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores. *Public Personnel Management*, *21*, 151-160.

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219-235. doi: 10.1177/1088868309341564

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*, 569-576. doi: 10.1037/0096-3445.136.4.569

Ambady, N. E., & Skowronski, J. J. (2008). *First Impressions.* Guilford Publications.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology, 70*, 394-400. doi: 10.1037/h0022280

Anderson, N. H. (1973). Serial position curves in impression formation. *Journal of Experimental Psychology, 97*, 8-12. doi: 10.1037/h0033774

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

Anderson, N. H., & Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior,* 2, 379-391. doi: 10.1016/S0022-5371(63)80039-0

Anderson, N. H., & Norman, A. (1964). Order effects in impression formation in four classes of stimuli. *The Journal of Abnormal and Social Psychology*, *69*, 467-471. doi: 10.1037/h0047472

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

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. East Sussex: Psychology Press.

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*, Freeman, New York.

Bassili, J. N. (Ed.). (1989). *On-line Cognition in Person Perception.* Psychology Press.

Belmore, S. M. (1987). Determinants of attention during impression formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 480-489. doi: 10.1037/0278-7393.13.3.480

Brown, R. D., & Bassili, J. N. (2002). Spontaneous trait associations and the case of the superstitious banana. *Journal of Experimental Social Psychology*, *38*, 87-92. doi: 10.1006/jesp.2001.1486

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*, 539-576. doi: 10.1037/0033-295X.114.3.539

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3-5. doi:10.1177/1745691610393980

Campbell, W. K., & Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology*, *3*, 23-43. doi: 10.1037/1089-2680.3.1.23

Carlston, D. E., & Skowronski, J. J. (1986). Trait memory and behavior memory: The effects of alternative pathways on impression judgment response times. *Journal of Personality and Social Psychology, 50*, 5-13. doi: http://dx.doi.org/10.1037/0022-3514.50.1.5

Clayson, D. E. (2005). Performance overconfidence: Metacognitive effects or misplaced student expectations? *Journal of Marketing Education*, *27*, 122-129. doi: 10.1177/0273475304273525

Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences, 9*, 393-399. doi: 10.1016/j.tics.2007.08.005

Clore, G. L., Schwarz, N., & Conway, M. (1994). Affective causes and consequences of social information processing. *Handbook of Social Cognition, 1*, 323-417.

Clore, G. L., Wyer R. S., Dienes, B., Gasper, K., Gohm, C. L., & Isbell, L. (2001). Affective feelings as feedback: Some cognitive consequences. In L. L. Martin & G. L. Clore (Eds.), *Theories of mood and cognition: A user's handbook* (pp. 27–62). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684. doi: 10.1016/S0022-5371(72)80001-X

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

Crawford, M. T., Sherman, S. J., & Hamilton, D. L. (2002). Perceived entiativity, stereotype formation, and the interchangeability of group members . *Journal of Personality and Social Psychology, 83,* 1076–1094. doi: 10.1037/0022-3514.83.5.1076

Crawford, M. T., Skowronski, J. J., Stiff, C., & Scherer, C. R. (2007). Interfering with inferential, but not associative, processes underlying spontaneous trait inference. *Personality and Social Psychology Bulletin, 33,* 677-690. doi: 10.1177/0146167206298567

Crowder, R. G. (1976). *Principles of Learning and Memory.* Hillsdale, NJ: Laurence Erlbaum Associates.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010, April). Are your participants gaming the system?: Screening mechanical turk workers. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 2399-2402. ACM.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*, 83-87.

Dweck, C. S. (1991). Self-theories and goals: Their role in motivation, personality, and development. In *Nebraska Symposium on Motivation*, 1990, University of Nebraska Press, Lincoln, pp. 199–235.

Epley, N., Gilovich, T., (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. J*ournal of Behavioral Decision Making, 18*, doi: 199–212. 10.1002/bdm.495

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*, 311-318, doi: 311-318. 10.1111/j.1467-9280.2006.01704.x

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*, 98-121. doi: 10.1016/j.obhdp.2007.05.002

Feldman, R. S., & Bernstein, A. G. (1977). Degree and sequence of success as determinants of self-attribution of ability. *The Journal of Social Psychology*, *102*, 223-231. doi: 10.1080/00224545.1977.9713268

Feldman, R. S., & Bernstein, A. G. (1978). Primacy effects in self−attribution of ability. *Journal of Personality*, *46*, 732-742. doi: 10.1111/j. 1467-6494.1978.tb00194.x

Finn, B. (2010). Ending on a high note: Adding a better end to effortful study. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36,* 1548-1553.

Finn, B. & Miele, D.B. (in press). Hitting a high note on math tests: Experiences of success influence test preferences. *Journal of Experimental Psychology: Learning, Memory and Cognition.*

Forgas, J. P., & Bower, G. H. (1987). Mood effects on person-perception judgments. *Journal of Personality and Social Psychology, 53*, 53-60. doi: 10.1037/0022-3514.53.1.53

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65*, 45–55. doi: 10.1037/0022-3514.65.1.45

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*. *19*, 847-857 doi:10.3758/s13423-012-0296-9

Greifeneder, R., Bless, H., & Pham, M. T. (2010). When do people rely on affective and cognitive feelings in judgment? A review. *Personality and Social Psychology Review, 15*, 107-141. doi: 10.1177/1088868310367640.

Hastie, R. (1980). Memory for behavioral information that confirms or contradicts a personality impression. In R. Hastie, T. M. Ostrom, R. S. Wyer, Jr., D. L. Hamilton, & D. E. Carlston (Eds.), *Person memory: The cognitive basis of social perception* (pp. 155-178). Hillsdale, NJ: Erlbaum.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological review, 93*, 258-268. doi: 10.1037/0033-295X.93.3.258

Hamilton, D. L., & Zanna, M. P. (1974). Context effects in impression formation: Changes in connotative meaning. *Journal of Personality and Social Psychology, 29*, 649-654. doi: 10.1037/h0036633

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28*, 597-606. doi: 10.1002/acp.3032

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief adjustment model. *Cognitive Psychology*, *24*, 1–55. doi: 10.1016/0010-0285(92)90002-J

Hoogerheide, V., & Paas, F. (2012). Remembered utility of unpleasant and pleasant learning experiences: Is all well that ends well? *Applied Cognitive Psychology, 26,* 887-894.

Jackson, A., & Greene, R. L., (2014). Impression formation of tests: Retrospective judgments of performance are higher when easier questions come first. *Memory and Cognition, 42,* 1325-1332. doi: 10.3758/s13421-014-0439-5

Jensen, M. P., Martin, S. A., & Cheung, R. (2005). The meaning of pain relief in a clinical trial. *Journal of Pain*, *6*, 400 – 406. doi: 10.1016/j.jpain.2005.01.360

Johnston, W. A., & Uhl, C. N. (1976). The contributions of encoding effort and variability to the spacing effect on free recall. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 153-160. doi: 10.1037/0278-7393.2.2.153

Jones, E. E., Rock, L., Shaver, K. G., Goethals, G. R., & Ward, L. M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology*, *10*, 317-340. doi: 317. 10.1037/h0026818

Kaplan, M. F. (1971) Context effects in impression formation: The weighted average versus the meaning-change formulation. *Journal of Personality and Social Psychology*, *19*, 92-99. doi: 10.1037/h0031098

Kennedy, E. J., Lawton, L., & Plumlee, E. L. (2002). Blissful ignorance: The problem of unrecognized incompetence and academic performance. *Journal of Marketing Education*, *24*, 243-252. doi: 10.1177/0273475302238047

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121-1134. doi: 10.1037/0022-3514.77.6.1121

Kühnen, U. (2010). Manipulation-checks as manipulation: Another look at the ease-of-retrieval heuristic. *Personality and Social Psychology Bulletin, 36,* 47-58. doi: 10.1177/0146167209346746

LeBoeuf, R.A., Shafir, E. (2009). Anchoring on the "Here" and "Now" in time and distance judgments. *Journal of Experimental Psychology*, *35*, 81–93. doi: 10.1037/a0013665

Lombardi, W. J., Higgins, E. T., & Bargh, J. A. (1987). The role of consciousness in priming effects on categorization assimilation versus contrast as a function of awareness of the priming task. *Personality and Social Psychology Bulletin*, *13*, 411-429. doi: 10.1177/0146167287133009

Mantonakis, A., Rodero, P., Lesschaeve, I., & Hastie, R. (2009). Order in choice: Effects of serial position on preferences. *Psychological Science*, *20*, 1309-1312. doi: 10.1111/j.1467-9280.2009.02453.x

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*, 1-12. doi:10.3758/s13428-011-0124-6

McCarthy, R. J., & Skowronski, J. J. (2011). The interplay of controlled and automatic processing in the expression of spontaneously inferred traits: A PDP analysis. *Journal of Personality and Social Psychology*, *100*, 229-240. doi: 10.1037/a0021991

McConnell, A. R., Sherman, S. J., & Hamilton, D. L. (1997). Target entitativity: implications for information processing about individual and group targets. *Journal of Personality and Social Psychology*, *72*, 750-762. doi: 10.1037/0022-3514.72.4.750

Metcalfe, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics?. *Personality and Social Psychology Review, 2,* 100-110. doi: 10.1207/s15327957pspr0202_3

Metcalfe, J. E., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing.* The MIT Press.

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, *130*, 711-747. doi: 10.1037/0033-2909.130.5.711

Michael, R. B., Newman, E. J., Vuorre, M., Cumming, G., & Garry, M. (2013). On the (non)persuasive power of a brain image. *Psychonomic Bulletin & Review, 20*, 720–725. doi: 10.3758/s13423-013-0391-6

Michael, R.B., & Garry, M. (2015) *Ordered questions bias eyewitnesses and jurors.* Manuscript submitted for publication.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115, 502-517. doi: 10.1037/0033-295X.115.2.502

Munz, D. C., & Smouse, A. D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. *Journal of Educational Psychology, 59*, 370–374. doi: 10.1037/h0026224

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*, 482-488. doi: doi: 10.1037/h0045106

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

Oppenheimer, D. M. (2004). Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science, 15,* 100-105.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867-872. doi: 10.1016/j.jesp.2009.03.009

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law, 1*, 817-845. doi:10.1037/1076-8971.1.4.817

Petty, R. E., & Wegener, D. T. (1993). Flexible correction processes in social judgment: Correcting for context-induced contrast. *Journal of Experimental Social Psychology*, *29*, 137-165. doi: 10.1006/jesp.1993.1007

Pyszczynski, T., Greenberg, J., & LaPrelle, J. (1985). Social comparison after success and failure: Biased search for information consistent with a self-serving conclusion. *Journal of Experimental Social Psychology*, *21*, 195-211. doi: 10.1016/0022-1031(85)90015-0

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

Redelmeier, D. A., Katz, J., & Kahneman, D. (2003). Memories of colonoscopy: a randomized trial. *Pain, 104*, 187-194. doi: 10.1016/S0304-3959(03)00003-4

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology, 10*, 173-220.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615-625. doi: 10.1037/a0013684

Schaller, M. (2008). Evolutionary bases of first impressions. *First Impressions*, Guilford Publications, 15-34.

Schwarz, N. (2011). Feelings-as-information theory. *Handbook of Theories of Social Psychology*, *1*, 289-308.

Schwarz, N., & Bless, H. (1991). Happy and mindless, but sad and smart? The impact of affective states on analytic reasoning. *Emotion and Social Judgments*, 55-71.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*, 195-202. doi: 10.1037/0022-3514.61.2.195

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129-138. doi: 10.1037/h0042769

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychological Bulletin, 134*, 207-222. doi: 10.1037/0033-2909.134.2.207

Skowronski, J. J., Carlston, D. E., Mae, L., & Crawford, M. T. (1998). Spontaneous trait transference: Communicators take on the qualities they describe in others. *Journal of Personality and Social Psychology*, *74*, 837-848. doi: 10.1037/0022-3514.74.4.837

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research, 177*, 1333-1352. doi: 10.1016/j.ejor.2005.04.006

Stone, A. A., Schwartz, J. E., Broderick, J. E., & Shiffman, S. S. (2005). Variability of momentary pain predicts recall of weekly pain: A consequence of the peak (or salience) memory heuristic. *Personality and Social Psychology Bulletin, 31*, 1340–1346. doi: 10.1177/0146167205275615

Strack, F., Schwarz, N., Bless, H., Kübler, A., & Wänke, M. (1993). Awareness of the influence as a determinant of assimilation versus contrast. *European Journal of Social Psychology*, *23*, 53-62. doi: 10.1002/ejsp.2420230105

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. *Behaviour Research Methods*, *45,* 1115-1143. doi:10.3758/s13428-012-0307-9

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

Uleman, J. S., Adil Saribay, S., & Gonzales, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*, 329-360. doi: 10.1146/annurev.psych.59.103006.093707

Uleman, J. S., & Kressel, L. M. (2013). A brief history of theory and research on impression formation. In D. E. Carlston (Ed.). *Oxford handbook of social cognition* (pp. 53-73). New York, NY: Oxford University Press.

Vander Schee, B. A. (2013). Test Item Order, Level of Difficulty, and Student Performance in Marketing Education. *Journal of Education for Business*, *88*, 36-42. doi: 10.1080/08832323.2011.633581

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*, 548-573. doi: 10.1037/0033-295X.92.4.548

Weiner, B. (2000). Intrapersonal and interpersonal theories of motivation from an attribution perspective. *Educational Psychology Review, 12*, 1-14, doi: 10.1023/A:1009017532121

Weinstein, Y., & Roediger, H. L. (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 III, H. L. (2012). The effect of question order on evaluations of test performance: how does the bias evolve?. *Memory & Cognition, 40*, 727-735 doi: 727-735.10.3758/s13421-012-0187-3

Wells, B. M., Skowronski, J. J., Crawford, M. T., Scherer, C. R., & Carlston, D. E. (2011). Inference making and linking both require thinking: Spontaneous trait inference and spontaneous trait transference both rely on working memory capacity. *Journal of Experimental Social Psychology*, *47*, 1116-1126. doi: 10.1016/j.jesp.2011.05.013

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*, 117-142 doi: 117.10.1037/0033-2909.116.1.117

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, *125*, 387-402. doi: 10.1037/0096-3445.125.4.387

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology*, *47*, 237-252. doi: 10.1037/0022-3514.47.2.237

Zajonc, R.B., 1980. Feeling and thinking: Preferences need no inferences. *American Psychologist 35*, 151–175. doi: 10.1037/0003-066X.35.2.151

Zanna, M. P., & Hamilton, D. L. (1977). Further evidence for meaning change in impression formation. *Journal of Experimental Social Psychology, 13*, 224-238. doi: 10.1016/0022-1031(77)90045-2

Zauberman, G., Diehl, K., & Ariely, D. (2006). Hedonic versus informational evaluations: Task dependent preferences for sequences of outcomes. *Journal of Behavioral Decision Making*, *19*, 191-211. doi: 10.1002/bdm.516

Appendix A

Test A: Question (Arranged from Easiest to Hardest)	Answer
What is the name of the comic strip character who eats spinach to increase his strength?	popeye
What is the name of the long sleep some animals go through during the entire winter?	hibernation
What is the capital of France?	paris
Which sport uses the terms "gutter" and "alley"?	bowling
What is the name of the remains of plants and animals that are found in stone?	fossils
What is the name of Dorothy's dog in "The Wizard of Oz"?	toto
What is the last name of the man who rode horseback in 1775 to warn that the British were coming?	revere
What is the last name of the singer who recorded "Heartbreak Hotel" and "All Shook Up"?	presley
What kind of metal is associated with a 50th wedding anniversary?	gold
What is the name of the bird that cannot fly and is the largest bird on Earth?	ostrich
What is the name of the thick layer of fat on a whale?	blubber
What is the only liquid metal at room temperature?	mercury
For which country is the Yen the monetary unit?	japan
What is the last name of the first person to set foot on the moon?	armstrong
What is the word that means a nautical mile per hour?	knot
What is the largest planet in the solar system?	jupiter
In which game are men crowned?	checkers
What is the name of the liquid portion of whole blood?	plasma
What is the name of the legendary one-eyed giant in Greek mythology?	cyclops
What is the name of deer meat?	venison
What is the longest river in South America?	amazon
What is the name of the chapel whose ceiling was painted by Michaelangelo?	sistine
What animal runs the fastest?	cheetah
What was the last name of the man who was the radio broadcaster for the "War of the Worlds"?	welles
What is the last name of the author who wrote "The Old Man and the Sea"?	hemmingway
What is the name of the extinct reptiles known as "terrible lizards"?	dinosaurs
What is the last name of the scientist who discovered radium?	curie
Of which country is Buenos Aires the capital?	argentina
In which city is the U.S. Naval Academy located?	annapolis
What is the last name of the man who invented the phonograph?	edison
What is the last name of the first signer of the "Declaration of Independence?"	hancock
Of which country is Nairobi the capital?	kenya
What is the name of the Roman emperor who fiddled while Rome burned?	nero
What Italian city was destroyed when Mount Vesuvius erupted in 79 A.D.?	pompeii
What is the last name of the man who wrote "Canterbury Tales?"	chaucer
What is the last name of the author who wrote "Brave New World?"	huxley
What is the last name of the man who first studied genetic inheritance in plants?	mendel
What is the only word the raven says in Edgar Allen Poe's poem "The Raven?"	nevermore

For which country is the Rupee the monetary unit?	india
Which sport uses the terms "stones" and "brooms"?	curling
What is the name of the North Star?	polaris
What was the name of the Apollo lunar module that landed the first man on the moon?	eagle
In what profession was Emmett Kelly?	clown
In which city is Michelangelo's statue of David located?	florence
What is the last name of the artist who painted "Guernica?"	picasso
Over which river is the George Washington Bridge?	hudson
What is the name of the brightest star in the sky excluding the sun?	sirius
The general named Hannibal was from what city?	carthage
What is the last name of the man who is regarded as the national poet of Scotland?	burns
What is the last name of the union general who defeated the confederate army at the civil war battle of Gettysburg?	meade

Test B: Question (Arranged from Easiest to Hardest)	**Answer**
What is the name of the horse-like animal with black and white stripes?	zebra
What was the name of Tarzan's girlfriend?	jane
What is the name of the molten rock that runs down the side of a volcano during an eruption?	lava
Which sport is associated with Wimbledon?	tennis
What is the name of the rubber object that is hit back and forth by hockey players?	puck
What is the name of an inability to sleep?	insomnia
What is the term for hitting a volleyball down hard onto the opponent's court?	spike
What is the name for a medical doctor who specializes in diseases of the skin?	dermatologist
What is the last name of the author who wrote "Romeo and Juliet"?	shakespeare
What is the name of the process by which plants make their food?	photosynthesis
In what park is "Old Faithful" located?	yellowstone
What is the name for a cyclone that occurs over land?	tornado
What is the name of the large hairy spider that live near bananas?	tarantula
What is the name of the navigation instrument used at sea to plot position relative to the magnetic north pole?	compass
Which breed of cat has blue eyes?	siamese
What is the last name of the man who proposed the theory of relativity?	einstein
What is the name for the astronomical bodies that enter the Earth's atmosphere?	meteors
Which game uses a rubber ball and little metal pieces?	jacks
What is the name of the short pleated skirt worn by men in Scotland?	kilt
What is the name of the ocean that is located between Africa and Australia?	indian
What is the name of the automobile instrument that measures mileage?	odometer
In which sport is the Stanley Cup awarded?	hockey
Which games uses a doubling cube?	backgammon
What is the last name of the man who assassinated President John F. Kennedy?	oswald
What is the name of the organ that produces insulin?	pancreas
What is the last name of the woman who began the profession of nursing?	nightingale
What is the name of the first artificial satellite put in orbit by Russia in 1957?	sputnik
What is the name of the three-leaf clover that is the emblem of Ireland?	shamrock
What is the last name of Batman's secret identity in the Batman comics?	wayne

What is the capital of New York?	albany
In which game are the standard pieces of Staunton design?	chess
What is the last name of the author of the book "1984?"	orwell
In what European city is the Parthenon located?	athens
What is the last name of the man who invented the telegraph?	morse
What was Frank Lloyd Wright's profession?	architect
Of which country is Budapest the capital?	hungary
In which city is Heathrow Airport located?	london
What is the name of the river on which Bonn is located?	rhine
In which city does the cotton bowl take place?	dallas
What is the capital of Denmark?	copenhagen
In what ancient city were the "Hanging Gardens" located?	babylon
What is the last name of the man who invented dynamite?	nobel
What is the capital of Delaware?	dover
What is the name of the ship on which Charles Darwin made his scientific voyage?	beagle
What is the capital of Chile?	santiago
What is the name of the mountain range that separates Asia from Europe?	ural
What is the name of the Chinese religion founded by Lao Tse?	taoism
What is the name of the instrument used to measure windspeed?	anemometer
What is the name of the villainous people who lived underground in H.G. Wells' book "The Time Machine?"	morlocks
What is the last name of the man who supposedly killed Jesse James?	ford

Appendix B

25 Question Test: Question	Answer
What is the name of the comic strip character who eats spinach to increase his strength?	popeye
What is the capital of France?	paris
What is the name of the remains of plants and animals that are found in stone?	fossils
What is the last name of the man who rode horseback in 1775 to warn that the British were coming?	revere
What kind of metal is associated with a 50th wedding anniversary?	gold
What is the name of the thick layer of fat on a whale?	blubber
For which country is the Yen the monetary unit?	japan
What is the word that means a nautical mile per hour?	knot
In which game are men crowned?	checkers
What is the name of the legendary one-eyed giant in Greek mythology?	cyclops
What is the longest river in South America?	amazon
What animal runs the fastest?	cheetah
What is the name of the extinct reptiles known as "terrible lizards"?	dinosaurs
Of which country is Buenos Aires the capital?	argentina
What is the last name of the man who invented the phonograph?	edison
Of which country is Nairobi the capital?	kenya
What Italian city was destroyed when Mount Vesuvius erupted in 79 A.D.?	pompeii
What is the last name of the author who wrote "Brave New World?"	huxley
What is the only word the raven says in Edgar Allen Poe's poem "The Raven?"	nevermore
Which sport uses the terms "stones" and "brooms"?	curling
What was the name of the Apollo lunar module that landed the first man on the moon?	eagle
In which city is Michelangelo's statue of David located?	florence
Over which river is the George Washington Bridge?	hudson
The general named Hannibal was from what city?	carthage
What is the last name of the union general who defeated the confederate army at the civil war battle of Gettysburg?	meade

10 Question Test: Question	Answer
What is the name of the comic strip character who eats spinach to increase his strength?	popeye
What is the name of the remains of plants and animals that are found in stone?	fossils
What is the name of the bird that cannot fly and is the largest bird on Earth?	ostrich
What is the word that means a nautical mile per hour?	knot
What is the name of deer meat?	venison
What is the last name of the man who invented the phonograph?	edison
What is the last name of the man who wrote "Canterbury Tales?"	chaucer
Which sport uses the terms "stones" and "brooms"?	curling
What is the last name of the artist who painted "Guernica?"	picasso
What is the last name of the union general who defeated the confederate army at the civil war battle of Gettysburg?	meade

3 Question Test: Question	Answer
What is the name of the comic strip character who eats spinach to increase his strength?	popeye
What is the last name of the author who wrote "The Old Man and the Sea"?	hemmingway
What is the last name of the union general who defeated the confederate army at the civil war battle of Gettysburg?	meade

Appendix C

Table C1. Examples of each self-reported strategy from Experiment 3.

Strategy Type	Example
Impression/Guess	”I feel like I may have gotten more right than wrong, but probably not by a large margin.”
Retrospective Tally	“I could not remember the Union general, I think the Confederate general was either Robert E. Lee. I think it was possibly Stonewall Jackson. I also had no clue who wrote that book, so I just guessed using popular authors at this time. I am confident of my last answer.”
On-line Tally	"Tried to keep track a long the way.”
Irrelevant/Vague	“I like history and trivia.”
Other	"I knew I'd gotten all the answers right.”

Appendix D

Supplementary Analyses

Experiment 1a-d Confidence Ratings

In Experiments 1b and 1d, after each question subjects recalled, we asked how confident they were that they had answered that question correctly. Perhaps subjects who take an Easy-Hard test are traditionally more optimistic about their performance because they are confident about the questions they can recall; by contrast, perhaps Hard-Easy people do the opposite, and are less confident about the questions they can recall—regardless of when they appeared on the test. But we found no evidence for this proposition. In fact, we found evidence that points in the opposite direction. In Experiment 1b, Easy-Hard subjects (M = 3.35, SD = 1.80) were less confident that they had answered the questions they recalled correctly than Hard-Easy subjects, (M = 3.97, SD = 1.54; M_{diff} = 0.62 95% CI [-0.69, 1.94]. In NHST terms, there was a nonsignificant trend for Easy-Hard subjects to be less confident than Hard-Easy subjects $t(27) = 0.97$, $p = 0.34$. In Experiment 1d, Easy-Hard subjects (M = 3.85, SD = 1.52) were again less confident than Hard-Easy subjects (M = 4.42, SD = 1.25; M_{diff} = 0.57 95% CI [-0.22, 1.37]. In NHST terms, there was again a nonsignificant trend for Easy-Hard subjects to be less confident than Hard-Easy subjects $t(48) = 1.43$, $p = 0.16$.

Experiment 2 Time Spent on the Test

The order of questions on the test hardly affected how long subjects took to take the test, M_{diff} = 3.59 seconds, 95% CI [-51.48, 58.66]. In NHST terms, the difference between the average time subjects spent on each question was nonsignificant $t(443) < 1$.

Appendix E

Table E1. Information about the subjects in each experiment.

Experiment	Condition	Males	Females	Mean Age (SD)
1a	Easy-Hard	11	14	33.89 (13.06)
	Hard-Easy	10	14	31.15 (8.88)
1b	Easy-Hard	9	9	31.63 (8.27)
	Hard-Easy	6	6	32.26 (10.50)
1c	Easy-Hard	15	9	28.72 (11.21)
	Hard-Easy	14	9	37.83 (11.88)
1d	Easy-Hard	17	15	32.33 (11.62)
	Hard-Easy	14	10	34.83 (12.18)
2	Easy-Hard, No Recall	48	42	30.50 (10.05)
	Easy-Hard, Recall	50	47	30.30 (10.43)
	Hard-Easy, No Recall	40	41	32.70 (12.86)
	Hard-Easy, Recall	49	41	32.00 (10.47)
3	Easy-Hard, 3	19	24	30.07 (10.45)
	Hard-Easy, 3	20	19	30.44 (11.22)
	Easy-Hard, 10	27	23	30.68 (10.08)
	Hard-Easy, 10	18	22	31.50 (9.75)
	Easy-Hard, 25	18	24	29.88 (10.82)
	Hard-Easy, 25	18	19	31.68 (13.95)
	Easy-Hard, 50	17	18	31.63 (9.96)
	Hard-Easy, 50	16	17	31.94 (10.49)
4	Easy-Hard, Warning	42	46	29.91 (12.63)
	Easy-Hard, No Warning	43	66	26.62 (10.30)
	Hard-Easy, Warning	38	52	27.56 (12.05)
	Hard-Easy, No Warning	31	54	30.01 (13.45)
5	Easy-Hard, Warning	18	26	34.59 (12.87)
	Easy-Hard, No Warning	23	19	30.60 (10.74)
	Hard-Easy, Warning	18	19	30.22 (10.68)
	Hard-Easy, No Warning	18	21	31.54 (12.53)