Ordered Questions Bias Eyewitnesses and Jurors

By Robert B. Michael

A thesis submitted to Victoria University of Wellington
In fulfilment of the requirements for the degree of
Doctor of Philosophy

Victoria University of Wellington
2016

Abstract

Eyewitnesses play an important role in the justice system. But suggestive questioning can distort eyewitness memory and confidence, and those distorted beliefs influence jurors (Loftus, 2005; Penrod & Cutler, 1995). Recent research, however, hints that suggestion is not necessary. Simply changing the order of a set of trivia questions altered people's beliefs about their accuracy on those questions (Weinstein & Roediger, 2010, 2012). I wondered to what degree eyewitnesses' beliefs—and in turn the jurors who evaluate them—would be affected by this simple change to the order in which they answer questions[1]. Across a number of experiments in Part 1 of my thesis[2], I show that the order of questions matters: Eyewitnesses reported higher accuracy and were more confident about their memory when questions seemed initially easy than when they seemed initially difficult. In addition, jurors' beliefs about eyewitnesses closely matched those of the eyewitnesses themselves. But why does the order of questions matter? How does this simple rearrangement produce these alarming effects? Across a number of experiments in Part 2 of my thesis, I explore the extent to which the data are consistent with an explanation where eyewitnesses rapidly form an impression of their performance that is resistant to change. Taken together, these findings have implications for eyewitness metacognition and for eyewitness questioning procedures.

---

[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 that fact. You will also see that I use the word "we" in a different context when referring to what is known (or not known) in the wider scientific community.

[2] Portions of this thesis were adapted from a published manuscript and another in preparation. More specifically, the six experiments in Part 1 were adapted from:

Michael, R. B. & Garry, M. (2015). Ordered questions bias eyewitnesses and jurors. *Psychonomic Bulletin & Review*, 1-8. doi:10.3758/s13423-015-0933-1

and the five experiments in Part 2 were adapted from:

Michael, R. B. & Garry, M. (manuscript in preparation). How do ordered questions bias eyewitnesses?

Acknowledgments

My sincere thanks to Victoria University of Wellington for all of the funding that has supported my research.

Thank you to the many undergraduate and honours students for your help with creating materials and collecting data, and for your excellent contributions to our discussions.

To my fellow Garry labsters, past and present: Seema, Stef, Kim, Deryn, Mel, Eryn, Jeff, Brittany, Gregg, Mevagh, Robbie, and Cassandra. What a crew! I consider every one of you an inspiration. Your advice, mentoring, and camaraderie have been vital. I hope our friendships continue long into the future.

My thanks to a number of academic collaborators and mentors, without whose help I could not have accomplished a fraction of this work. In particular, thank you to Gina Grimshaw and Matt Crawford for many helpful discussions, to Dan Bernstein and Steve Lindsay for their incredible support, and to Irving Kirsch, Eryn Newman, Matti Vuorre, Geoff Cumming, Gabriel Braniff, and last but not least, Beth Loftus, for their amazing work as co-authors and collaborators.

I owe an unimaginably enormous debt of thanks to my supervisor Maryanne Garry, who has continuously supported my progression from a petrified undergraduate to a capable and passionate academic. This experience has fundamentally changed my life for the better, and I am where I am today in no small part because of Maryanne's guidance and care.

Finally, to family and friends: Thank you for all the advice, support, chewed ears, shoulders, reality checks, and occasional indulgences. You have all helped me on this journey in some way, large or small, and for that I am eternally grateful.

I dedicate this thesis to the memory of my mother.

Contents

List of figures and tables

Part 1

Chapter 1

On your lunch break you decide to deposit that cheque you have been meaning to take care of. While waiting in line at the bank for the next available teller, a commotion breaks out. A man shouts, "Everybody get down!" You turn to see that he is brandishing a gun. You drop to the floor. The man dumps an empty duffel bag on the counter and signals the nearest bank teller, who begins stuffing the bag with money. Once filled, the man snatches the bag back, backs out of the bank, and gets into a car which speeds away.

Your memory for this event is now an important element in a criminal investigation. Aware of that fact, you do your best to play back the event in your mind as you wait for the police to arrive: the colour of the man's shirt; the defining features of his face; the gun; the duffel bag bulging with cash; the getaway car. A few minutes pass before the police arrive. They begin interviewing the eyewitnesses—including you.

How might that line of questioning proceed? Would the officer—following commonly recommended practice (Clarke, Milne, & Bull, 2011; Paulo, Albuquerque, & Bull, 2013)—begin by asking you some basic questions that feel relatively easy to answer, establishing rapport before moving on to the hard-hitting questions: *What's your name? Age? Address? Where do you work? Where were you standing? What were you doing just before the robbery began?* Alternatively, would the officer—eager to find the robber—perhaps instead launch straight into the hard-hitting questions that feel relatively difficult to answer, before returning to the easier questions: *What was the man wearing? What type of gun did he have? Can you describe his face? What about the duffel bag—what did it look like? What colour, make and model was the getaway car?*

Now ask yourself the following question: How would the arrangement of the officer's questions influence what you believe about your memory for the robbery? More specifically: How would the order of difficulty of those questions—a seemingly trivial feature—change how you think about your memory as an eyewitness? And—supposing for a moment that the arrangement of questions does, in fact, sway your beliefs—then to what extent would your changed beliefs influence

what other people, like jurors, think and believe about your memory as an eyewitness? These are the primary questions I address in Part 1.

EYEWITNESS MEMORY

Memory plays a pivotal role in the criminal justice system. When people witness criminal activity—or even the circumstances surrounding that activity—they become a valuable resource. Why? Because when eyewitnesses retrieve the information they have previously encoded and stored in memory, they can share those retrieved details and help other people reconstruct the events of a crime. In doing so, eyewitnesses assist triers-of-fact—like jurors and judges—in building a coherent story of how a criminal event unfolded (Pennington & Hastie, 1986, 1988, 1992). But more importantly, eyewitnesses can provide information that holds probative value, indicating the guilt or innocence of suspects.

Against a backdrop of other available evidence—like DNA, fingerprints, shoeprints, and bullet analysis, for example—eyewitness memory might contribute only one small piece to the crime puzzle. But when that other evidence is scarce, or worse—non-existent, eyewitness memory becomes critical and can be highly influential. Eyewitnesses, that is, are persuasive—particularly when they express themselves confidently (Cutler, Penrod, & Dexter, 1990; Douglass, Neuschatz, Imrich, & Wilkinson, 2010).

That persuasive power of the eyewitness would be entirely appropriate if human memory was flawless. In fact, it would be extremely convenient for the justice system—not to mention a remarkable cognitive feat—if memory functioned like a video camera. More specifically, if incoming information was recorded as a reliable representation of reality and could be "played back" in full fidelity, without error, then memory's role in the criminal justice system would change. Memory would no longer be merely probative—it would be imperative. We could say, with utmost confidence, that a suspect is guilty because an eyewitness has a memory of that suspect committing the crime.

Unfortunately, memory does not work that way—although many people believe it does (Simons & Chabris, 2011, 2012). Decades of scientific research shows that memory does not work like a video camera. Instead, as pioneering memory researcher Elizabeth Loftus states in her TED talk, memory works more like your

own wikipedia page: You might be the author of an original memory, but then you and others can go in and make changes (Loftus, 2013). Unlike wikipedia, however, we do not keep a reliable record of the changes that have been made, nor who made them; incoming information isn't conveniently tagged with its source (Johnson, Hashtroudi, & Lindsay, 1993). Without that record, it becomes virtually impossible—without additional corroboration—to know the accuracy of what's remembered.

This ambiguity of memory accuracy poses a problem, because there are serious consequences when a memory system works more like a wikipedia page than a video camera. Consider the work of The Innocence Project, a global organisation dedicated to exonerating wrongfully convicted people through DNA testing. This organisation has now helped exonerate 337 people who were wrongfully convicted (Innocence Project, 2016). These people spent, on average, 14 years of their lives in prison for crimes they did not commit. And these 337 people are only the cases we know about; they represent a fraction of those who have been wrongfully convicted. It is impossible to know and exceedingly difficult to even estimate how many innocent people are in prison this very moment (Risinger, 2007). But how do these travesties of justice happen? There are a number of causes, including: false, often coerced confessions; unvalidated or improper forensic science; police and prosecutorial misconduct; lying informants; and inadequate defence attorneys (Innocence Project, 2016). But the greatest contributing factor to the wrongful conviction of an innocent person is also, somewhat ironically, the least dubious. Eyewitness misidentification—a memory error—plays a key role in more than 70% of these convictions that were overturned through DNA testing (Innocence Project, 2016). At the heart of this societal problem lies the fragility of memory.

DISTORTED MEMORIES AND BELIEFS

Psychological scientists have long known that memory is easily distorted. Take the striking results from changing a single word in a question: In one study, witnesses reported cars in an accident travelled faster when a question suggested the cars *smashed* into rather than *hit* each other (Loftus & Palmer, 1974). In another study, witnesses were more likely to report seeing a non-existent broken headlight when a question suggested its presence using the word *the*, rather than the more ambiguous

a (Loftus & Zanni, 1975). More than four decades of research now shows that questions can transmit misleading suggestions that distort memory (see Loftus, 2005, for a review).

But questions can distort more than the details of memories; they can exert equally interesting influences on metacognition and metamemory—that is, the thoughts, beliefs, and strategies people have about their own thinking and memory. For instance, eyewitnesses incorrectly answer misleading questions both quickly and confidently (Loftus, Donders, Hoffman, & Schooler, 1989), and people generally provide more information—but monitor less for accuracy—when forced to answer questions compared with when they decide themselves what to report (Koriat & Goldsmith, 1996). These studies show that questions can change not only the content of eyewitnesses' memories—but also how eyewitnesses think and what they believe about their memory.

It is alarming that suggestive questions can so easily distort eyewitnesses' memories and beliefs. But more alarming still are the potential consequences of these distortions. Consider how other people, like jurors, evaluate the accuracy or veridical status of information that an eyewitness reports. There is no tool available to a juror, or to anyone for that matter, that can distinguish true from false memories with absolute certainty. Jurors must instead rely on cues that signal accuracy, like the eyewitness's behaviour.

One of these behavioural cues that is highly influential is the confidence with which eyewitnesses express themselves, illustrated by the following mock-jury study (Cutler et al., 1990). In this study, two sets of subjects—eligible, experienced jurors and undergraduates—watched a videotaped trial involving eyewitness evidence. The researchers manipulated a collection of 10 different factors—drawn from psychological theory—that relate to the quality of a witness's memory, including the disguise of the perpetrator, the length of time between the witnessed event and the witness's identification, and the witness's confidence. Subjects made a dichotomous verdict (guilty or not guilty) and estimated the probability that the witness's identification was correct. Both measures were virtually identical across nine of the ten manipulations. That result alone is alarming. It shows that jurors' evaluations of an eyewitness's memory are unmoved by several factors that can be

informative about the quality of that memory. But most important was the finding that jurors believed the witness's identification was 7% more likely to be correct when the witness testified that she was 100% confident, compared to 80% confident. A 7% difference might not seem like a large difference. But it is dangerous to underestimate the influence of a factor that, at first glance, seems trivial. Across a large number of cases, for example, effects traditionally considered small accumulate, leading to meaningful outcomes (Abelson, 1985; Rosenthal, 1990). Additionally, in a legal context where people's freedom is at stake, it is vital that we minimise the influence of extraneous factors as much as possible. In summary, this finding shows that jurors are sensitive to eyewitness confidence, using confidence as a cue to the accuracy of an eyewitness's memory.

On the one hand, it might seem like good intuitive sense to rely on confidence as a proxy for truth. Moreover, some research supports this idea. We know, for example, that although the relationship between confidence and accuracy varies, it grows stronger as the conditions for memory become optimal (Bothwell, Deffenbacher, & Brigham, 1987; Brewer, Weber, Wootton, & Lindsay, 2012; Brewer & Wells, 2006; Deffenbacher, 1980). But the problem is that this relationship can easily be undermined, as illustrated in the following study (Wells & Bradfield, 1998). Subjects watched a security camera video of an armed robbery, and then attempted to identify the gunman in a photo lineup. Because the culprit was not in the lineup, all identifications were false. The experimenter then gave some subjects confirmatory feedback, saying "Good, you identified the suspect." Subjects then answered a number of questions assessing their memory for, and beliefs about, the witnessed event. Worryingly, this simple confirmatory feedback made eyewitnesses more certain about their lineup choice. Moreover, the feedback made eyewitnesses reappraise their memory, reporting that they had a better view of the culprit, could make out more details of his face, and that they paid more attention. These eyewitnesses were also more willing to testify in court. In summary, this study shows that a single phrase can distort the relationship between confidence and accuracy: Witnesses can be confident they are right, when they are completely wrong. That finding also fits with more recent research on the confidence-accuracy

relationship, wherein illusions of familiarity turn confidence into a signal of inaccuracy rather than accuracy (DeSoto & Roediger, 2014).

The implications of a hijacked confidence-accuracy relationship for the criminal justice system are clear. Most jurors will not know when the relationship between confidence and accuracy has been thwarted. It is unlikely, for example, that the court will hear how the officer congratulated the eyewitness on their lineup choice. The following study demonstrates the consequences of this lack of awareness (Douglass et al., 2010). In this study, mock jurors watched a videotaped eyewitness interview and then evaluated that eyewitness on a number of dimensions. Unbeknownst to the jurors, some of them were viewing an eyewitness who had been given confirmatory feedback about an earlier, false lineup identification. The results clearly showed that artificially inflated eyewitness confidence is persuasive to jurors, because jurors rated those witnesses who were given confirmatory feedback as more accurate in and confident about their identification, as having paid more attention, as having had a better view, and as having a better memory for faces. These findings are important because they show that confident witnesses persuade jurors, even when that confidence is an unreliable indicator of accuracy.

Taken together, this research on eyewitnesses and jurors shows that it is vital we understand the factors that influence memory and confidence. One common element underpins these manipulations that affect people's memories and beliefs. That common element is a degree of suggestion or deception. But what if suggestion or deception is unnecessary? What if there is another important property—a property of the set of questions given to eyewitnesses—that can influence what people believe about what they remember?

THE INFLUENCE OF ORDERED DIFFICULTY

Recently, researchers in cognitive psychology discovered a new important property of tests in an educational context: The order in which questions are arranged (Jackson & Greene, 2014; Weinstein & Roediger, 2010, 2012). Across a series of experiments, subjects answered a set of 100 trivia questions, arranged either from the easiest to the most difficult question, or from the most difficult to the easiest question. Subjects then estimated how many of the trivia questions they had answered correctly. Despite virtually identical objective performance, the subjects

who first answered the easy questions believed they answered more questions correctly than the subjects who first answered the difficult questions. A number of candidate mechanisms could explain the influence of ordered questions; I examine these potential explanations empirically in Part 2.

Of course, the influence of order more generally dates back a long way. Some of the earliest research on memory, for example, famously established the importance of the temporal features of information: The classic *serial position effect* shows the relative advantage in recalling information encountered early and late in a series, compared to recalling information encountered in the middle (Ebbinghaus, 1913). Order—and in particular information encountered early rather than late—is also important in shaping the impressions we form of other people (Anderson & Barrios, 1961; Asch, 1946), the attributions we make about people's intellectual ability (Jones, Rock, Shaver, Goethals, & Ward, 1968), and even the way jurors build a story about how events could have happened (Pennington & Hastie, 1986, 1988, 1992).

But what is novel and surprising about these recent trivia studies is that they highlight the importance of not simply order in general, but more specifically the order of *difficulty* of a set of questions, in changing people's beliefs. That property of order seems trivial at face value—after all, everyone answers the same overall set of questions—and yet its influence is not. It is also a property that has not been systematically examined until now, and could lead to applications in fields other than education—like eyewitness memory.

We were intrigued by these trivia study findings and wondered to what extent the order of questions would influence eyewitnesses' beliefs about the accuracy and quality of their memory. Of course, it is not obvious that the order of questions should influence eyewitnesses at all. Whereas trivia questions can be drawn from a virtually infinite pool, questions put to eyewitnesses are typically from a more limited set, addressing a specific and recent event. This relative constraint should provide fewer opportunities for uncertainty, reducing eyewitnesses' reliance on heuristic processing—cognitive shortcuts that can result in biased judgments from seemingly innocuous manipulations (Shah & Oppenheimer, 2008; Tversky & Kahneman, 1974). It would be surprising and worrying if a simple

change to the order of questions put to eyewitnesses could change how they appraise their memories.

OVERVIEW OF EXPERIMENTS

To what extent does the order of questions put to eyewitnesses change what they believe about their memory? To answer that question, we showed people a video of a simulated crime, and then after a short delay, tested their memory for that crime. Importantly, we arranged the questions on the memory test in one of two ways: from the easiest through to the most difficult, or from the most difficult to the easiest. Immediately afterward, we asked people: [1] to estimate how many questions they thought they had answered correctly, and [2] how confident they felt about the accuracy of their memory for the events in the video.

In Experiment 1, the order of questions influenced people's beliefs about their memory: Subjects who first answered easy questions believed they answered more questions correctly than subjects who first answered difficult questions—even though both groups actually answered about the same number correctly. Subjects who first answered easy questions were also more confident about the accuracy of their memory.

In Experiment 2, we showed that the influence of the order of questions was not limited to a particular type of test: The same pattern of results appeared when we changed the test from a 2-alternative forced choice recognition test to a cued-recall test. That finding suggests that it is reasonable to expect that question order could have an influence in the field, where questions put to eyewitnesses are typically in a cued-recall format.

In Experiment 3, we showed that the influence of the order of questions was not limited to a particular set of materials, nor to a fixed question set size: The same pattern of results appeared when we used a different simulated crime video with a different pool of questions, and 20 rather than 30 questions total. Those findings suggest that the influence of question order is robust and generalizable.

We then wondered: To what extent does the influence of question order extend beyond the eyewitnesses themselves? To answer that question, we asked people to take on the role of jurors, evaluating an eyewitness from a previous study. Our jurors read an eyewitness report consisting of the questions and answers from

the memory test in Experiments 1-3, but unbeknownst to our jurors, we secretly created two versions of this report and gave half our jurors one version, and the other half the other version. In one version, the eyewitness began with high confidence in their answers but became steadily less confident—making the test appear as though it was first easy then became difficult. In the other version, the eyewitness began with low confidence in their answers but became steadily more confident—making the test appear as though it was difficult at first and then became easy. Immediately after reading the eyewitness's report, our jurors answered two questions: [1] how many questions they thought the eyewitness answered correctly, and [2] how confident they were about the accuracy of the eyewitness's memory.

In Experiment 4, we showed that the influence of question arrangement extends beyond eyewitnesses themselves: When eyewitnesses expressed high confidence in their early answers through to low confidence in their late answers, jurors believed that those eyewitnesses answered more questions correctly and were more confident in the accuracy of those eyewitnesses' memories, compared to when eyewitnesses expressed low confidence in their early answers through to high confidence in their late answers.

In Experiment 5, we controlled for a confounding variable and ruled out an alternative explanation for these findings. Specifically, we showed that the results are driven by the eyewitness's confidence, and not by the content of the particular questions associated with that confidence.

In Experiment 6—as in Experiment 3—we showed that the influence of order was not limited to a particular set of materials, nor to a fixed question set size: The same pattern of results appeared when we used the pool of 20 questions from Experiment 3 that are about a different simulated crime event.

Chapter 2

Experiment 1

Method

Subjects. Through pilot work, we determined a sample size of 100 (50 per between subjects cell). We ultimately recruited a total of 102 subjects from Amazon's Mechanical Turk (www.mturk.com)[3]. Age and gender information for this and all other experiments is presented in Appendix A.

Design. We used a simple two groups design with Question Order (low-to high confidence, high-to-low confidence) manipulated between subjects.

Procedure. We used Qualtrics survey software (Qualtrics, Provo, UT) to present instructions and materials in subjects' web browsers. The experiment had four phases. First, we told subjects the study was examining visual and verbal learning styles. We used this minor deception because revealing the true purpose of the experiment—that it was investigating the influence of the order of questions on eyewitness memory—would likely influence the results. We also gave subjects a set of general experimental instructions to follow. Next, subjects answered demographic questions: age, country of origin, country of residence, gender, and level of education. Because none of these measures were reliable covariates in any experiment, I will not discuss them further. However, the interested reader can see Appendix A for age and gender information. Subjects then watched one of two similar videos of a tradesman who stole items from the unoccupied house in which he was working (Takarangi, Parker, & Garry, 2006). We counterbalanced video versions across subjects and conditions.

The second phase began when the video ended. To encourage a small degree of memory decay—similar to what is likely to take place in settings outside of a

[3] Mechanical Turk and Qualtrics—the survey software that we use as our experimental platform—interact in such a way that it is possible to inadvertently collect more data points than requested. Subjects view a brief description of the experiment on Mechanical Turk that includes a link to the survey. If subjects choose to participate, they should click an "Accept" button on Mechanical Turk that reserves the subject a place in the allotted pool. Some subjects do not click this button. Instead, they go directly to the survey and complete it. Occasionally, when returning to Mechanical Turk to click "Accept ", these subjects find that all allocated spaces have already been filled. These additional subjects were included in all analyses.

controlled laboratory experiment—subjects solved Sudoku number puzzles for 10 minutes.

In the third phase, subjects took a surprise memory test consisting of 30 two-alternative forced choice (2AFC) questions about the video.

To construct the order of test items, we used data from an earlier, separate group of 107 subjects. These subjects followed the procedure described so far before answering the 30 questions in random, computer-generated orders. Twenty of the 30 questions came from the same published set of materials as the video (Takarangi et al., 2006). The confidence people report in their answers to these 20 questions, however, skews high, suggesting that none of the questions feels particularly difficult (M = 4.10 on a 5 point scale, range = 2.29 – 4.97; Foster, unpublished data). We therefore generated an additional 10 questions we thought subjects would find difficult. For each of the 30 questions, subjects used a scale from 1 ("Not at all confident") to 5 ("Very confident") to report their confidence they had selected the correct answer. Then, using these confidence ratings, we ordered the 30 questions from the lowest mean confidence (M = 1.73, SD = 1.06) to highest mean confidence (M = 4.79, SD = 0.63) to produce the low-to-high confidence test. We reversed this order to create the high-to-low confidence version[4]. See Appendix B for the complete list of questions, answers, and associated confidence ratings.

Subjects in the current experiment were randomly assigned one of these two test versions. For each question, subjects selected one of the two possible answers they thought was correct, and then used a scale from 1 ("Not at all confident") to 5 ("Very confident") to report their confidence they had selected the correct answer.

The fourth phase followed the memory test. Subjects answered two randomly ordered questions: [1] "The memory test you just took consisted of 30 questions. How many of those questions do you think you answered correctly?" Subjects responded with a number between 0 and 30; [2] "Suppose that you were asked to testify as an eyewitness. How confident would you be in your memory of the events

[4] To consider the possibility that the subjective experience of confidence tracked subjective difficulty, we asked a group of 141 people to rate the difficulty of each randomly ordered question on a scale from 1 ("Very easy") to 5 ("Very difficult"). When we compared items in this group with items in the group of 107 subjects, we found that reported difficulty closely matched reported confidence (r = -.82, 95% CI [-.66, -.91], p < .001, treating items as cases), suggesting that confidence was a good proxy for subjective difficulty.

you saw in the video?" Subjects responded on a scale from 1 ("Not at all confident") to 5 ("Very confident"). Finally, subjects answered a number of questions assessing compliance with the set of general instructions given in Phase 1. The full list of these questions appears in Appendix F. To encourage honest responding, we made subjects aware they would receive compensation irrespective of their answers to these compliance questions.

Results and Discussion

In all experiments, the overall pattern of results was consistent when including or excluding subjects who failed to comply with any of our general instructions. That finding suggests that the general instructions were not a necessary condition to elicit any effects of interest. We therefore included all subjects in the reported analyses. There were no other exclusion criteria.

We first performed a manipulation check by examining mean confidence ratings for individual test questions. These data appear in the top panel of Figure 1 and show that our manipulation worked: "low-to-high" subjects were increasingly confident and "high-to-low" subjects were decreasingly confident. The middle panel of Figure 1 displays mean accuracy for individual test questions and shows a similar pattern—although less cleanly, as a consequence of binary 2AFC scoring. The bottom panel of Figure 1 displays the mean confidence-accuracy relationships for individual test questions and suggests that the order of questions did not affect subjects' insight into their own accuracy. We also found that the order of questions had little effect on overall test performance, $M_{\text{diff}} = 0.65$ (2.17%), 95% confidence interval (CI) [-0.46, 1.76]; $t(100) = 1.15$, $p = .252$.

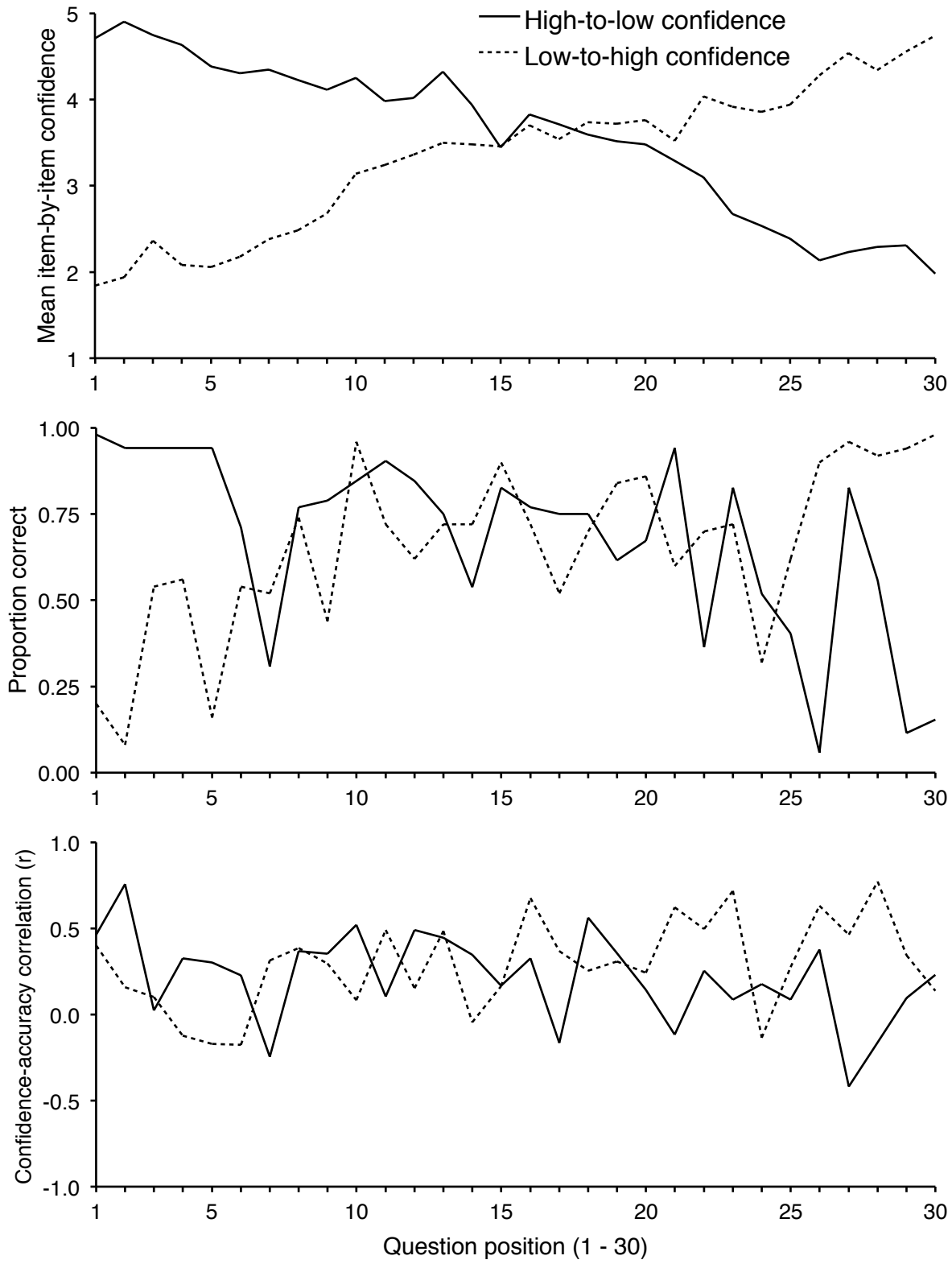

Figure 1. *Top panel:* Mean confidence of a correct answer for each test question, ordered by position on test. *Middle panel:* Proportion of subjects who answered each test question correctly, ordered by position on test. *Bottom panel:* Pearson correlations between confidence and accuracy ratings for each test question, ordered by position on test. Note that the test versions are symmetric, i.e., question 1 in one condition is the same as question 30 in the other condition. Data are from Experiment 1.

We now address our primary questions: To what extent did the order of questions [1] bias subjects' retrospective estimates of their test performance, and [2] affect their confidence in the accuracy of their memory? To answer [1], we subtracted subjects' test scores from their retrospective estimates to produce bias scores. Positive bias scores represent subjects who thought they performed better on the test than they truly did, and negative bias scores represent subjects who thought they performed worse on the test than they truly did. We present mean test scores and mean retrospective estimates of test scores in the top panel of Figure 2. We present mean bias scores in the middle panel of Figure 2. These data show that low-to-high confidence subjects were more pessimistic than high-to-low confidence subjects, M_{diff} = 2.32 (7.73%), 95% CI [0.33, 4.32]; $t(100) = 2.31$, $p = .023$. To answer [2], we examined subjects' mean post-test reports of memory confidence. These data appear in the bottom panel of Figure 2 and show that low-to-high confidence subjects were less confident about the accuracy of their memory: M_{diff} = 0.46 (11.50%), 95% CI [0.08, 0.84]; $t(100) = 2.41$, $p = .018$ (for all experiments in Part 1, we report cell means and SDs in Tables 1 and 2).

In a forensic setting, it is unlikely that most questions put to eyewitnesses will rely simply on recognition—as is the case when questions appear as a forced choice between two alternatives. Instead, most questions will rely on what the eyewitness can recall. Recall tasks are more difficult than recognition tasks and provide less constraint, because the correct answer now comes from a potentially much larger pool of possible answers. On the one hand, that relatively greater freedom to "roam" memory in a recall task could lead to greater variability in the experience of difficulty across questions, diminishing the influence of question order. On the other hand, research using the trivia question paradigm relies on recall and successfully finds an influence of question order (Jackson & Greene, 2014; Weinstein & Roediger, 2010, 2012). Nonetheless, it would be useful to similarly demonstrate such a finding in an eyewitness context. Therefore, to determine the extent to which these effects would generalize to the more real-world situation of open-ended questions, we conducted Experiment 2.

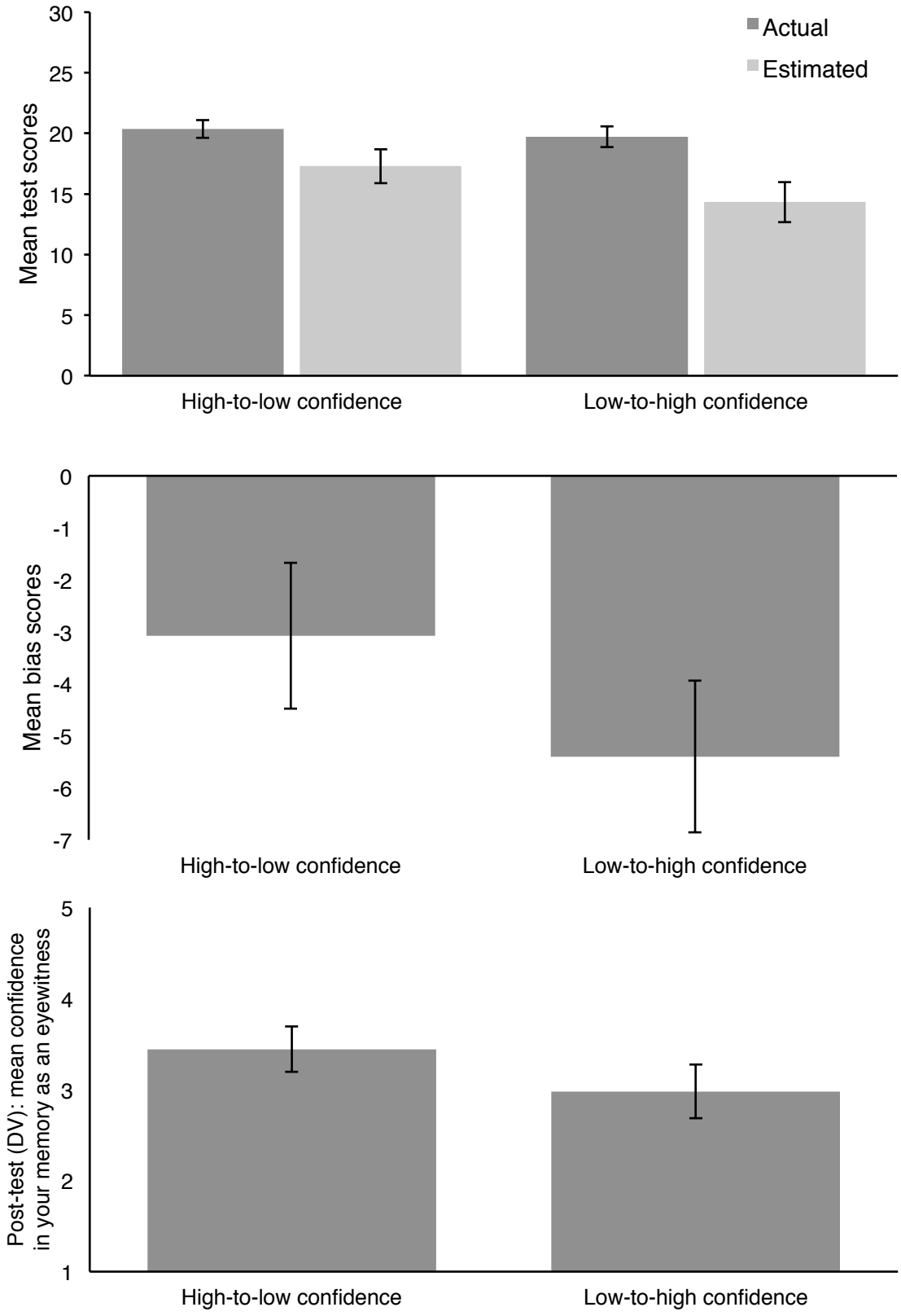

Figure 2. *Top panel:* Mean actual and estimated test scores by condition. *Middle panel:* Mean bias (estimated test score - actual test score) by condition. Positive bias scores represent subjects who thought they performed better than they truly did; negative bias scores represent the opposite. *Bottom panel:* Mean post-test memory confidence by condition. Error bars represent 95% confidence intervals of cell means. Data are from Experiment 1.

Table 1. Experiments 1, 2, and 3 mean scores for Bias and Confidence by condition.

	Bias				Confidence			
	Condition				Condition			
	High-to-low	Low-to-high	M_{Diff}	95% CI	High-to-low	Low-to-high	M_{Diff}	95% CI
Expt 1 (N = 102)	-3.08 (5.03)	-5.40 (5.13)	2.32	0.33, 4.32	3.44 (0.89)	2.98 (1.04)	0.46	0.08, 0.84
Expt 2 (N = 220)	1.64 (5.49)	-2.02 (4.43)	3.65	2.33, 4.98	2.69 (1.08)	2.32 (0.94)	0.37	0.10, 0.64
Expt 3 (N = 205)	-3.25 (4.00)	-5.13 (3.43)	1.88	0.86, 2.90	2.99 (1.01)	2.71 (0.97)	0.28	0.01, 0.55
Meta-analysis			10.33%	7.36, 13.30			0.36	0.19, 0.52

Standard deviations in parentheses. Note: The meta-analysed difference is a percentage, because the number of test questions was 30 in Experiment 1 and 2, but 20 in Experiment 3.

Table 2. Experiments 4, 5, and 6 mean scores for Estimate and Confidence by condition.

	Estimate				Confidence			
	Condition				Condition			
	High-to-low	Low-to-high	M_{Diff}	95% CI	High-to-low	Low-to-high	M_{Diff}	95% CI
Expt 4 (N = 261)	17.47 (4.81)	14.23 (4.29)	3.23	2.12, 4.34	3.15 (0.84)	2.67 (0.84)	0.47	0.27, 0.68
Expt 5 (N = 305)	17.79 (4.52)	13.62 (5.13)	4.18	3.09, 5.27	3.06 (0.85)	2.34 (0.92)	0.72	0.52, 0.92
Expt 6 (N = 316)	12.33 (3.30)	10.46 (3.38)	1.88	1.14, 2.62	3.06 (0.91)	2.64 (0.89)	0.42	0.22, 0.62
Meta-analysis			11.38%	8.77, 14.00			0.54	0.36, 0.72

Standard deviations in parentheses. Note: The meta-analysed difference is a percentage, because the number of test questions was 30 in Experiments 4 and 5, but 20 in Experiment 6.

Experiment 2

Method

Subjects. To boost precision, we recruited a larger sample of 220 Mechanical Turk workers.

Design and Procedure. Experiment 2 followed the design and procedure of Experiment 1, except we converted each 2AFC question into a cued-recall question.

Results and Discussion

We scored subjects' responses to the questions by a computerised keyword search. For example, if a subject's response to the question, "How many toothbrushes were in the bathroom?" included either "six" or "6", it was marked correct. To ensure scoring was not unfairly conservative, the keyword search ignored letter case and whitespace in subjects' responses. In addition, a blind rater hand-scored a random 20% of responses; keyword and hand scores were highly correlated, $r = 0.96$, $p < .001$. For the complete list of questions and keywords, see Appendix C.

This new format replicated the earlier results: the order of questions had little effect on overall test performance, $M_{diff} = 0.61$ (2.03%), 95% confidence interval (CI) [-0.35, 1.57]; $t(218) = 1.25$, $p = .213$, yet low-to-high confidence subjects were more pessimistic, $M_{diff} = 3.65$ (12.17%), 95% CI [2.33, 4.98]; $t(218) = 5.43$, $p < .001$; and were less confident about the accuracy of their memory, $M_{diff} = 0.37$ (9.25%), 95% CI [0.10, 0.64]; $t(218) = 2.73$, $p = .007$. We next ran Experiment 3 to ensure these effects were not tied to specific materials.

Experiment 3

Method

Subjects. We recruited a new sample of 205 Mechanical Turk workers.

Design and Procedure. The design and procedure was the same as Experiment 1, except subjects viewed a different video and answered a different set of twenty 2AFC questions (French, Garry, & Mori, 2011).

To construct the order of test items, we again used data from an earlier, separate group of 106 subjects who answered the 20 questions in random, computer-generated orders. For each question, subjects used a scale from 1 ("Very easy") to 5 ("Very difficult") to rate the difficulty of the question. We used subjective difficulty

because one criticism of the prior experiments was that confidence—despite being a good proxy for subjective difficulty—is not necessarily the same thing. Using these difficulty ratings, we ordered the questions from those rated easiest (M = 1.61, SD = 1.07) to those rated most difficult (M = 4.67, SD = 0.74) to produce the high-to-low confidence test. We reversed this order to create the low-to-high confidence version. Note that this naming convention is merely used for consistency. Subjects were randomly assigned one of these two test versions. See Appendix D for the complete list of questions, answers, and difficulty ratings.

Results and Discussion

As before, we found that low-to-high confidence subjects were more pessimistic, M_{diff} = 1.88 (9.40%), 95% CI [0.86, 2.90]; $t(203) = 3.62$, $p < .001$; they were also less confident about the accuracy of their memory, M_{diff} = 0.28 (7.00%), 95% CI [0.01, 0.55]; $t(203) = 2.05$, $p = .042$. These data show that the influence of the order of questions generalizes to different materials.

In line with the recommendations of Cumming (2013), we obtained more precise estimates of these effect sizes by meta-analysing the results of Experiments 1, 2, and 3, using ESCI software to run two random effects model meta-analyses. These analyses estimate that "low-to-high" eyewitnesses would be 10.33% more pessimistic about their performance than "high-to-low" eyewitnesses, M_{diff} = 10.33%, 95% CI [7.36, 13.30], $z = 6.82$, $p < .001$. These "low-to-high" eyewitnesses would also be 0.36 points, or 9.00%, less confident about what they remember, M_{diff} = 0.36, 95% CI [0.19, 0.52], $z = 4.10$, $p < .001$.

The results of Experiments 1, 2, and 3 show that the order of questions shapes what eyewitnesses believe. Specifically, when people answered questions that initially seemed difficult and then became easy, they were more pessimistic and less confident about their memory compared with others who answered questions that initially seemed easy and then became difficult.

In changing how eyewitnesses appraise their memories, one possible consequence is that jurors will appraise the eyewitness's credibility in the same direction (Douglass et al., 2010). Such a result would have disturbing implications for the justice system. Because jurors tend to rely on eyewitness confidence as a signal of accuracy (Penrod & Cutler, 1995), we asked subjects in Experiments 4, 5,

and 6 to take on the role of a juror, evaluating an eyewitness whose confidence systematically changed over the course of questioning.

Experiment 4

Method

Subjects. We aimed to collect data from 200 people but ultimately recruited 261 Mechanical Turk workers.

Design. We used a two groups design with Question Order (low-to-high confidence, high-to-low confidence) manipulated between subjects.

Procedure. We asked people to take on the role of a juror and answer questions about an eyewitness who had been in a previous study. We told these "jurors" that in the previous study, the eyewitness had taken a memory test after watching the video of Eric the Electrician. The juror's task was not to watch the video but to carefully read the eyewitness's memory test and then answer some questions.

To mirror the real-world scenario where a group of jurors evaluate one eyewitness, all jurors within a condition actually read a single eyewitness's test that we secretly created. In the high-to-low confidence condition, we manufactured the test so that the eyewitness's answers were initially confident but became less confident over the test. In the low-to-high confidence condition, this pattern reversed. We used data from Experiment 1 to help create these two versions of the eyewitness's test. First, we randomly selected, for each of the 30 test questions, which answer the eyewitness had ostensibly chosen. Next, we calculated the mean confidence ratings eyewitnesses had given to each of the 30 questions in Experiment 1, rounding each mean to an integer. We used these integers to select positions on the Likert scales the eyewitness had ostensibly used to report their confidence that each answer was correct. The result of this procedure was a completed eyewitness test with 30 questions, 30 randomly selected answers, and 30 confidence ratings that descended from high to low. We then flipped this entire test to produce the other version.

Subject jurors randomly received either the low-to-high confidence or high-to-low confidence eyewitness test, formatted exactly like the test in Experiment 1. Immediately after reading the eyewitness's test, subjects answered two randomly

ordered questions: [1] "The memory test about Eric the Electrician consisted of 30 questions. How many of those questions do you think the eyewitness answered correctly?" Subjects responded with a number between 0 and 30; [2] "How confident are you about the accuracy of the eyewitness's memory?" Subjects responded on a scale from 1 ("Not at all confident") to 5 ("Very confident").

Results and Discussion

Jurors believed that an initially confident eyewitness was more accurate, estimating that these eyewitnesses answered more questions correctly, M_{diff} = 3.23 (10.77%), 95% CI [2.12, 4.34]; $t(259) = 5.73$, $p < .001$. Jurors also reported more confidence in these eyewitnesses' memories, M_{diff} = 0.47 (11.75%), 95% CI [0.27, 0.68]; $t(259) = 4.54$, $p < .001$.

Note, however, that the pattern of the eyewitness's confidence ratings always covaried with the pattern of questions put to that eyewitness. That is, each of the 30 test questions always appeared with the same confidence rating. This confound leaves open the possibility that jurors were influenced not by the eyewitness's confidence, but by the content of the questions. We ran Experiment 5 to address this counter-explanation.

Experiment 5

Method

Subjects. We aimed to boost precision by increasing observations to 150 per between subjects cell, ultimately recruiting 305 Mechanical Turk workers.

Design and Procedure. The design and procedure was the same as in Experiment 4, except that we randomly assigned, for each subject, which question would appear with each confidence rating. This modification decoupled question content from confidence ratings, while maintaining the ascending or descending pattern of eyewitness confidence.

Results and Discussion

We found again that subjects believed high-to-low confidence eyewitnesses answered more questions correctly, M_{diff} = 4.18 (13.93%), 95% CI [3.09, 5.27]; $t(303) = 7.54$, $p < .001$, and were more confident about the accuracy of these eyewitnesses' memories, M_{diff} = 0.72 (18.00%), 95% CI [0.52, 0.92]; $t(303) = 7.10$, $p < .001$.

Finally, we ran Experiment 6 to demonstrate that these effects were not tied to specific materials.

Experiment 6

Method

Subjects. We aimed to collect 150 observations per between subjects cell, and ultimately recruited 316 Mechanical Turk workers.

Design and Procedure. The design and procedure was the same as in Experiment 5, but used the set of test questions from Experiment 3. We created the two versions of the eyewitness's report using data from Experiment 3. We calculated mean confidence ratings for each of the 20 questions, rounding each mean to an integer so it could be represented on the Likert scale of confidence the eyewitness had ostensibly used. We also randomly selected, for each test question, which answer the eyewitness had ostensibly chosen. We decoupled these questions from their associated confidence ratings by randomly assigning questions to each confidence rating. Subjects randomly received either the low-to-high confidence or high-to-low confidence eyewitness test.

Results and Discussion

We found again that jurors believed high-to-low confidence eyewitnesses answered more questions correctly, M_{diff} = 1.88 (9.40%), 95% CI [1.14, 2.62]; $t(314)$ = 4.99, $p < .001$, and jurors were also more confident about the accuracy of these eyewitnesses' memories, M_{diff} = 0.42 (10.50%), 95% CI [0.22, 0.62]; $t(314)$ = 4.14, $p < .001$.

The findings from Experiments 4, 5, and 6 fit with those of Experiments 1, 2, and 3, in which eyewitnesses thought they answered more questions correctly and reported higher confidence in their memory if their initial experience was one of high confidence. We meta-analysed the results of Experiments 4, 5, and 6 (Cumming, 2013) and estimated that jurors believe "high-to-low" eyewitnesses answer 11.38% more questions correctly, M_{diff} = 11.38%, 95% CI [8.77, 14.00], z = 8.53, $p < .001$. Moreover, jurors are 0.54 points—or 13.50%—more confident about the accuracy of a "high-to-low" eyewitness's memory, M_{diff} = 0.54, 95% CI [0.36, 0.72], z = 5.75, $p < .001$.

Chapter 3

Across six experiments, we found that the order in which eyewitnesses answered questions mattered in two key ways. First, the order changed how eyewitnesses appraised themselves. When questions produced an initial experience of high confidence rather than low confidence, eyewitnesses believed that they were more accurate and were more confident about their memory. Second, the order changed how jurors appraised eyewitnesses. Jurors believed eyewitnesses who initially displayed high confidence were more accurate, and jurors were more confident about those eyewitnesses' memories. This collection of results paints a worrying picture of the malleability of beliefs about memory accuracy.

It is surprising that questions produce different beliefs in witnesses when all that changes is the order those questions are asked. Ultimately, everyone answers the same questions, so it seems reasonable to expect no differences in beliefs. But the influence of order shows that beliefs about memory are shaped not only by the content or phrasing of questions, but also by factors that—on the face of it—are trivial.

In fact, our seemingly trivial manipulation produced effects similar in size to more blatant manipulations affecting eyewitness credibility. An eyewitness who claims to be absolutely certain, for example, is rated more credible than an eyewitness who does not (Tenney, MacCoun, Spellman, & Hastie, 2007), and prosecution eyewitnesses who elaborate their testimony with extra details are more credible, and get more guilty verdicts, than eyewitnesses who do not (Bell & Loftus, 1988, 1989). It is worrying that our subtle manipulation produces effects similar in magnitude to these relatively heavy-handed approaches.

How can we explain our effects? One possibility is that people's attention wanes over the test, resulting in beliefs influenced most by early experience (Crano, 1977). If this *attention decrement* hypothesis is true, then we might expect that the same question would be answered with higher accuracy when it appears early rather than late. To investigate this possibility, we ran a random effects model meta-analysis comprising all three datasets from Experiments 1, 2, and 3. This meta-analysis compared accuracy between groups for the subjectively easiest and most difficult test questions, because each appears first for one group and last for the

other. We found no support for this attention-based explanation: Accuracy is not notably different when a question appears first rather than last, M_{diff} = -0.01, 95% CI [-0.04, 0.02], z = -0.40, p = .686.

An alternative explanation is that the effects are driven by early experience and insufficient adjustments: The subjective ease or difficulty of early questions sets an anchor, and to save effort, people adjust from this anchor only until reaching a plausible impression (Epley & Gilovich, 2006). This explanation is consistent with our data and that from recent research in which subjects held biased impressions of performance throughout a trivia test, and not merely at the end (Weinstein & Roediger, 2010, 2012). Relatedly, Experiments 4-6 suggest that jurors used early information to create a story about the eyewitness's credibility and were slow to revise that story in the face of new information. This explanation fits with the Story Model of juror decision-making, a model in which juror's verdicts are influenced by the stories they construct to make sense of events (Pennington & Hastie, 1986, 1988, 1992). I conduct a more thorough investigation of the mechanisms responsible for the biasing influence of question arrangement in Part 2.

Our findings have implications for eyewitnesses' metacognition, because they suggest that the order of questions influences eyewitnesses' ability to evaluate what they know about an event. Similarly, our findings are reminiscent of other suggestive techniques that manipulate eyewitness beliefs, such as subtle changes to the wording of questions, or direct feedback about lineup identifications (Douglass & Steblay, 2006; Loftus & Palmer, 1974; Loftus & Zanni, 1975). But in contrast, we have manipulated what eyewitnesses and jurors believe about memory without using suggestive techniques.

Our findings also raise interesting questions. For instance, does the order of questions influence other related judgments, such as eyewitnesses' estimates of how well they saw the perpetrator? We know that positive post-identification feedback enhances eyewitnesses' beliefs about their memory for a crime, including how well they could see a suspect's face and how much attention they paid (Wells & Bradfield, 1998). Perhaps an initial experience of subjectively easy questions causes similar enhancements. It would also be useful to know if the order of questions produces lasting changes in beliefs or if the influence is fleeting. Finally, it is worth

considering that we ordered questions in our experiments either by subjective confidence or subjective difficulty. Earlier work has ordered questions by objective difficulty, calculated as the mean proportion of people who answer a question correctly (Jackson & Greene, 2014; Weinstein & Roediger, 2010, 2012). Our results suggest that the subjective experience of difficulty may underpin the influence of question order—but a future experiment teasing apart subjective and objective difficulty could provide information about their relative contributions.

Moreover, what makes a question easy or difficult? In the trivia studies, difficulty was operationalized as the proportion of people who answer a question correctly, based on prior norms. Our alternative is the subjective experience of ease or confidence. But these definitions tell us only *which* questions people are likely to answer correctly or experience as easy, not *why*. There are a multitude of reasons why questions vary in difficulty. Questions are typically easy when they assess information: we know well; that is emotional; that received more attention; or even that simply feels easy to remember (Cahill & McGaugh, 1995; Christianson & Loftus, 1991; Oppenheimer, 2008). The questions used in my experiments likely contain a complex mix of these characteristics, and I hypothesise that the subjective experience of difficulty is influential irrespective of the underlying question characteristics producing it. But it would be useful in future work to disentangle these properties, in order to better understand the nature of questions in a forensic setting.

Eyewitnesses play an undeniably important role in the justice system. But justice requires that we protect the integrity of eyewitness memory as much as possible. That integrity is called into question when eyewitnesses and jurors are swayed by something as trivial as the order in which they answer questions. It is therefore crucial that we gain a better understanding of the underlying psychological mechanisms responsible for the biasing influence of question arrangement. That is the focus of Part 2.

Part 2

Chapter 1

In Part 1 I documented, over a series of six experiments, the influence of question arrangement on eyewitnesses' and jurors' beliefs about eyewitness memory. Those findings are novel and alarming because they show that there is a new cause for concern where memory intersects with the law. You do not need to use suggestive techniques to change what people believe about their memory, or even what people believe about someone else's memory. Instead, the findings suggest that something as seemingly trivial as starting an interview with a few easy questions could be enough to sway an eyewitness into thinking that, overall, she did a great job—even though she did not.

But we have yet to answer the question of *how* the arrangement of questions exerts its influence. That is the focus of Part 2. Naturally, there are a number of candidate explanations. I will begin by addressing those explanations covered in prior research that have little-to-no evidence in their favour. Then I will address more promising explanations and outline a series of experiments investigating the mechanism(s) responsible for the biasing influence of question order.

THE PRIMACY EFFECT

One explanation for how the arrangement of questions influences people relates to how people encode and remember information from a series. One of the most famous and robust findings in memory research—the *serial position effect*—is the relative advantage in recalling information encountered early and late in a series, compared to recalling information encountered in the middle (Ebbinghaus, 1913). The relative advantage in recalling early information is known as the *primacy effect*, and the relative advantage in recalling late information is known as the *recency effect*. Primacy effects typically occur because people have more opportunity to rehearse early information, which strengthens its transfer into long-term memory. Recency effects typically occur because late information is still available in short-term (or working) memory (Glenberg et al., 1980; Marshall & Werder, 1972; Rundus, 1980)[5].

[5] There are, in fact, a number of explanations for recency effects that differ primarily according to the length of time information is retained. But the focus here is on an

Perhaps, then, when people are asked to evaluate the quality of their memory after taking a memory test—either by giving an estimate for how well they've done or by reporting how confident they feel about the accuracy of their memory—what they remember best is the earliest information. For people who first answered easy questions, this primacy effect would result in evaluations skewed towards higher estimates and confidence. For people who first answered difficult questions, this primacy effect would result in evaluations skewed towards lower estimates and confidence.

But there are a number of reasons why this explanation is unlikely to be correct, or at the least is insufficient. First, the evaluations people make happen immediately after the memory test. We should therefore see the influence of a recency effect in addition to a primacy effect. But the pattern of results across these experiments fits only with a primacy effect and not a recency effect. Second, in trivia question studies where people are asked immediately following the test to report the questions they remember, the questions they tend to report are the most recent—not the earliest (Franco, 2015). That finding of a recency effect is the exact opposite of what we should see if the primacy explanation were true. Third, this explanation suggests that the influence of question arrangement arises only when people recall the questions while making their evaluations. But research shows that the influence of question arrangement arises during the test experience, and not merely afterward when making evaluations (Weinstein & Roediger, 2012).

Taken together, an explanation that relies purely on a memory-based primacy effect cannot adequately explain the influence of question arrangement—both theoretically and empirically.

THE AFFECT HEURISTIC

Another explanation for how the arrangement of questions influences people relates to the use of a particular mental shortcut that people employ when making decisions. The use of mental shortcuts—or heuristics—in general is adaptive (Gigerenzer, 1996; Gigerenzer & Goldstein, 1996). When situations arise where difficult decisions or judgements must be made with speed and efficiency, it makes

---

explanation for the influence of question arrangement that relies on a primacy effect, and so I will not expand further on different recency effect explanations.

little sense to rely on a process that requires a lengthy, exhaustive search for information. Instead, we can take mental shortcuts, using quick processes that incorporate whatever information is readily available to simplify decisions and judgements (Shah & Oppenheimer, 2008). Typically, there is little to no cost in using these heuristics. But in certain circumstances our mental shortcuts can lead us astray, resulting in a number of different cognitive biases (Finucane, Alhakami, Slovic, & Johnson, 2000; Tversky & Kahneman, 1973, 1974).

One source of information that is readily available is our emotional response, or *affect*. We can use the positive or negative feelings that we rapidly—and typically involuntarily—generate in response to stimuli to quickly make judgements and decisions (Finucane et al., 2000; Shiv & Fedorikhin, 1999; Winkielman, Zajonc, & Schwarz, 1997; Zajonc, 1980). Perhaps, then, when people are asked to evaluate the quality of their memory after taking a memory test, their feelings influence their evaluations. If people feel relatively positive, their evaluations will be skewed towards higher estimates and confidence, but if people feel relatively negative, their evaluations will be skewed towards lower estimates and confidence.

There are a number of problems with this explanation, too. First, everyone answers the same overall set of questions, so there is no good reason to expect that people who begin with easy questions will feel more positive after the test than people who end with those same easy questions. Second, and relatedly, the explanation would require not only that people employ the affect heuristic, but also that the early questions evoke stronger emotional responses than later questions. If that were true, we should expect to see higher reports of confidence in people's answers to easy questions when they appear early rather than late, and lower reports of confidence in people's answers to difficult questions when they appear early rather than late. But research using the trivia question paradigm has found no differences in confidence ratings in response to individual questions (Weinstein & Roediger, 2010). Third, research shows that people's reports of how much they are enjoying a trivia test varies according to question difficulty, but not order (Weinstein & Roediger, 2012). In other words, on average people find the test similarly enjoyable, regardless of the order of questions.

Taken together, an explanation that relies purely on the affect heuristic—much like the preceding primacy explanation—cannot adequately explain the influence of question arrangement, both theoretically and empirically.

THE ANCHORING-AND-ADJUSTMENT HEURISTIC

Another explanation for how the arrangement of questions influences people relates to the use of a different mental shortcut—the *anchoring-and-adjustment heuristic*. Just as we can draw on our current and readily available feelings when making decisions or judgements, so too can we draw on our readily available knowledge about numbers, dates, and values (Tversky & Kahneman, 1974).

Anchoring-and-adjustment is a two-step process. In the first step of this process, we receive or generate an anchor—a piece of numeric information—from or in response to a question about which our answer is uncertain. Sometimes that anchor is given to us in an initial, comparative question (Tversky & Kahneman, 1974). For example, the anchor 5,000 in the question, "*Is the population of Timaru greater or less than 5,000?"* would precede the question *"What is the population of Timaru?"* At other times, we provide our own anchor (Epley & Gilovich, 2001). For example, when encountering the question *"What is the population of Timaru?"* a person living in Wellington might spontaneously generate the population of Wellington as an anchor; an answer known to be wrong but that nonetheless comes easily to mind. In the second step of this process, we adjust away from the anchor before giving an answer. But because adjustments require effort and attention, we typically adjust only until we reach a plausible value (Crano, 1977; Epley & Gilovich, 2006). Our adjustments are therefore frequently insufficient, and lead to answers that are biased towards the initial anchor.

Perhaps, then, when people are asked to evaluate the quality of their memory after taking a memory test, those evaluations are a product of the repeated use of the anchoring-and-adjustment heuristic that occurred over the course of the test. More specifically, the first test questions set a subjective anchor—e.g., "I'm getting 100% right" for the people who start with easy questions, or "I'm getting 0% right" for the people who start with difficult questions—and because adjustments are insufficient, the end result is evaluations that are biased towards those anchors.

Data from research using the trivia-question paradigm are consistent with this explanation. In one experiment, subjects answered a series of 100 trivia questions arranged either from the easiest through to the most difficult, or vice versa. After each block of 10 questions, subjects estimated how many of those 10 questions they had answered correctly. On every block, subjects who first answered easy questions thought they answered more questions correctly than subjects who first answered difficult questions—a pattern of results suggesting that subjects anchored to an initial, self-generated value that they failed to adequately adjust away from (Weinstein & Roediger, 2012).

There are at least two problems with this explanation, however. First, an anchor is a number, either generated in response to a question or given in the question itself. But in our paradigm, there are no anchors in the questions, and the questions themselves are not about numbers. We might therefore expect that people do not generate an anchor. Of course, subjects could generate their own—but even if they do, recall from Experiment 1 that people's insight into their own accuracy varies considerably; people do not necessarily know when they are right and when they are wrong. We might expect then, that even if subjects do generate their own anchors, that they will vary widely. Alternatively, people might generate an anchor that is not a number per se, but rather a general belief about their performance. If that is the case, however, the explanation more closely resembles an alternative explanation based on how people form impressions—an explanation I return to in more depth later. Second, this explanation is incomplete, or at the least relies on an unusual definition of the anchoring-and-adjustment heuristic. Typically, experiments investigating the influence of the heuristic use a single anchor—given or generated in response to a single question—which subjects adjust away from. But in our experiments and those using the trivia-questions paradigm, people are given multiple questions. It is unclear what role each new question would play. For instance, is each question considered a new anchor, or only the first, with later questions instead merely being a cue used in adjustment? The former implies a repeating anchoring-and-adjustment process, while the latter implies a single anchoring process with repeated adjustment processes. In either case, the formulation is not typical of the paradigm. While it is at least plausible to modify the

explanation to fit within our paradigm, these issues highlight one of the key problems plaguing the heuristics approach: They can at once explain everything and nothing (Gigerenzer, 1996; Gigerenzer & Goldstein, 1996).

IMPRESSION FORMATION

One final explanation for how the arrangement of questions influences people relates to the way we form impressions about others and ourselves. We have long known that when it comes to our developing beliefs about others, not all information is created equal. Research shows that people can rapidly form impressions of other entities—typically people or groups of people—and these impressions seem to have a "sticky" quality; they are frequently resistant to change (Anderson, 1965; Anderson & Barrios, 1961; Asch, 1946; DeCoster & Claypool, 2004; Hamilton & Sherman, 1996; McConnell, Sherman, & Hamilton, 1994, 1997). Take, for example, a classic paradigm demonstrating this effect. Subjects are read a list of adjectives that describe a person, and are asked to form an impression of that person. These adjectives might progress, for instance, from positive through to negative descriptors, as in the following list: *brilliant, creative, kind, unattractive, opinionated, shallow*. For half the subjects, the adjectives are read in reverse order. When subjects then give a description of the person or rate them on a number of evaluative dimensions, a consistent pattern emerges: The first adjectives are most influential in the development of the overall, resulting impression. For example, subjects who first heard *brilliant* will have a more favourable impression of the person than subjects who first heard *shallow*—even though both groups heard the same entire list (Anderson, 1965; Anderson & Barrios, 1961; Asch, 1946). This phenomenon is summed up by the colloquial phrase "first impressions last."

Why are first impressions so influential? There are at least three explanations. One explanation—the *attention decrement* hypothesis—supposes that attention declines over time, such that early information receives the most attention and is therefore most influential (Crano, 1977). A second explanation—the *change-of-meaning* hypothesis—suggests that early information establishes an expectation that later information will be consistent with that early information. As such, the meaning of later information is changed in an attempt to make it fit more closely with the early information (Asch, 1946; Hamilton & Zanna, 1974; Zanna & Hamilton,

1977). The third, related explanation—the *inconsistency discounting* hypothesis—also maintains that early information establishes an expectation, but rather than changing the meaning of later information, this explanation instead supposes that people discount later information when it is inconsistent with their expectation (Anderson & Jacobson, 1965). Evidence for each of these three candidate explanations is mixed, and still debated today—but regardless, the phenomenon of a primacy effect[6] in impression formation is robust (DeCoster & Claypool, 2004; Uleman & Kressel, 2013).

If we can rapidly form impressions about other people and groups, it seems plausible that we could also rapidly form impressions about ourselves—including how well we're performing on a memory test, and the quality of our memory for an event. Perhaps, then, when people are asked to evaluate the quality of their memory after taking a memory test, those evaluations are a product of a global impression formed over the course of the test experience. More specifically, the first test questions could be most influential because they receive the most attention, or because they set an expectation of ease or difficulty through which the rest of the test is filtered, either by changing the meaning of later information, or by discounting it.

But we have already shown that decreasing attention is an unlikely explanation for the influence of question arrangement. The change of meaning hypothesis is also an unlikely explanation. If people changed the meaning of later information to fit more closely to the meaning of early information, then we should expect that confidence ratings for easy questions would be higher when those questions appear first than when they appear last, and that confidence ratings for difficult questions would be lower when those questions appear first than when they appear last. But in the trivia questions paradigm, no consistent differences emerge in confidence ratings for individual questions according to their order (Weinstein & Roediger, 2010, 2012). Moreover, our own data provide at best only weak support for this prediction. For example, the average item confidence from Experiment 1 in

[6] Note that a primacy effect as described in the social cognitive literature on impression formation is not the same phenomenon as a primacy effect as described in the memory literature. Specifically, the primacy effect in memory refers to relatively better retrieval of information encountered early in a series than information encountered in the middle. The primacy effect in impression formation refers to evaluations of a target that are consistent with information encountered early in a series.

Part 1 was only slightly higher for subjects who began with easy questions, M_{Diff} = 0.23, 95% CI [-0.01, 0.48], $t(100) = 1.91$, $p = .06$. We cannot rule out the possibility of the third explanation, however: That people discount later information when it is inconsistent with their expectations.

OVERVIEW OF EXPERIMENTS

The data from Part 1 cannot distinguish between two candidate mechanisms: anchoring-and-adjustment, and impression formation. In addition, the cognitive processes purportedly responsible for both these candidate mechanisms provide insufficient explanations. Specifically, in each case an additional process or "tweak" must be made to the explanation for it to work with our data. Nonetheless, the question remains: How does the order of questions shape eyewitnesses' impressions of test performance and beliefs about the accuracy of their memory?

We first took an exploratory approach in addressing that question. In Experiment 1, we asked subjects to make a prediction after every test question; estimating how many of the 30 total test questions they believed they would answer correctly. Where previously we had only one time point—that is, at the end of the test—this repeated questioning procedure instead lets us see how people's beliefs about their performance develop over the course of the test. The results showed that difficult questions produced large changes in beliefs when encountered early, but the same questions produced virtually no change in beliefs when encountered late.

In Experiment 2, we show that the beliefs people develop over the course of the test are not dependent on a particular test format. Similar to Experiment 2 from Part 1, the influence of question arrangement remained the same when we switched from a recognition test to a more difficult cued recall test format.

The patterns of results from these two experiments were difficult to reconcile with an anchoring-and-adjustment explanation. We therefore next examined the extent to which impression formation was a viable explanation for the biasing influence of question arrangement. One of the assumptions underlying the process of impression formation is that people expect other individuals to have consistency in their behaviour (Asch, 1946; Hamilton & Sherman, 1996). That assumption fits with a mechanism whereby people change the meaning of later information to more closely align with earlier information, or a mechanism whereby people discount

later information if it is inconsistent with earlier information. But when it comes to the behaviour of a group of individuals, we do not have this same expectation, because it is reasonable to expect that members of a group can be quite different from one another. We are therefore not as predisposed to discounting or changing the meaning of incoming information when it pertains to a group. To put it another way, we expect a degree of *coherence* in the behaviours or traits of an individual that we do not expect in a group (Hamilton & Sherman, 1996; McConnell et al., 1994, 1997). We attempted to capitalise on this difference in Experiments 3 and 4.

In Experiment 3, we manipulated the test so that it appeared to be our standard individual test of 30 questions, or so that it appeared to be a group of 3 tests consisting of 10 questions each. If the processes producing primacy effects in impression formation are reduced or eliminated when the target is a group rather than an individual, then the influence of question arrangement might be diminished when the test is made to appear as though it is a group of 3 tests. But the results showed that the influence of question arrangement was similar regardless of whether the test appeared to be a single test or a group of 3 tests.

In Experiment 4, we tested a counter-explanation for the pattern of results from Experiment 3. Rather than manipulating the appearance of the test itself, we manipulated the ostensible source of the test questions. In one version, we told subjects that all the questions came from a single individual. In the other version, we told subjects that each question came from a different individual—all of whom were part of a group. We expected that this manipulation would make the test seem like a single test when all the questions came from a single individual, but would make the test seem more like a group of 30 tests when each question came from a different individual. But the results again showed that the influence of question arrangement was similar regardless of whether the test questions ostensibly came from 1 or 30 people.

With scarce evidence to support an impression formation explanation, we re-focused our attention on an anchoring-and-adjustment explanation. In Experiment 5 we examined the extent to which anchoring-and-adjustment is a viable explanation for the biasing influence of question arrangement. We gave some subjects an initial high or low anchor before the test, telling them that most people get 90% (or 10%) of

the questions correct. If anchoring-and-adjustment is a viable explanation for the influence of question arrangement, we should see people's estimates of test performance and ratings of memory confidence skewed towards these anchors. Alternatively, if impression formation—and in particular the discounting of inconsistent information—is a viable explanation for the influence of question arrangement, we should see a diminished or reversed influence of question arrangement when the first questions are inconsistent with the anchor. The results suggest that people are sensitive to an initial anchor, and adjust away from it insufficiently.

Chapter 2

Experiment 1

Method

Subjects. We aimed to collect 100 observations per between subjects cell and ultimately recruited 219 Mechanical Turk workers.

Design and Procedure. The design and procedure was the same as Experiment 1 from Part 1, except as follows. After answering each test question and giving a confidence rating for that question, subjects were asked the following: "This test consists of 30 questions total. How many of those questions do you think you will get correct?" Subjects responded with a number between 0 and 30. The only time this question did not appear was after the final test question. Here, we instead asked our two standard post-test questions: the retrospective estimate and the memory confidence rating.

Results and Discussion

How does the order of questions shape the beliefs people form of their own performance? To answer this question, we examined the mean predicted test scores people reported after each test question; these data appear in Figure 3. The figure shows that the influence of a question on a person's beliefs about their test performance depends on the difficulty of that question, and when that question appears. High-to-low confidence subjects' initial estimates were high (M_{Time1} = 23.83, SD_{Time1} = 4.79) and descended steadily over the course of the test (M_{Time30} = 18.49, SD_{Time30} = 5.38). But for low-to-high confidence subjects the pattern was markedly different. More specifically, the pattern was not just the inverse of the high-to-low confidence subjects. Instead, low-to-high confidence subjects' initial estimates were already much lower than the high-to-low confidence subjects' (M_{Time1} = 17.95, SD_{Time1} = 5.36) and continued to drop until reaching their lowest point after the ninth question (M_{Time9} = 10.00, SD_{Time9} = 6.87), at which point they ascended steadily over the remainder of the test (M_{Time30} = 15.08, SD_{Time30} = 5.04).

To determine the extent to which these patterns would replicate and generalize to the more real-world situation of open-ended questions, we conducted Experiment 2.

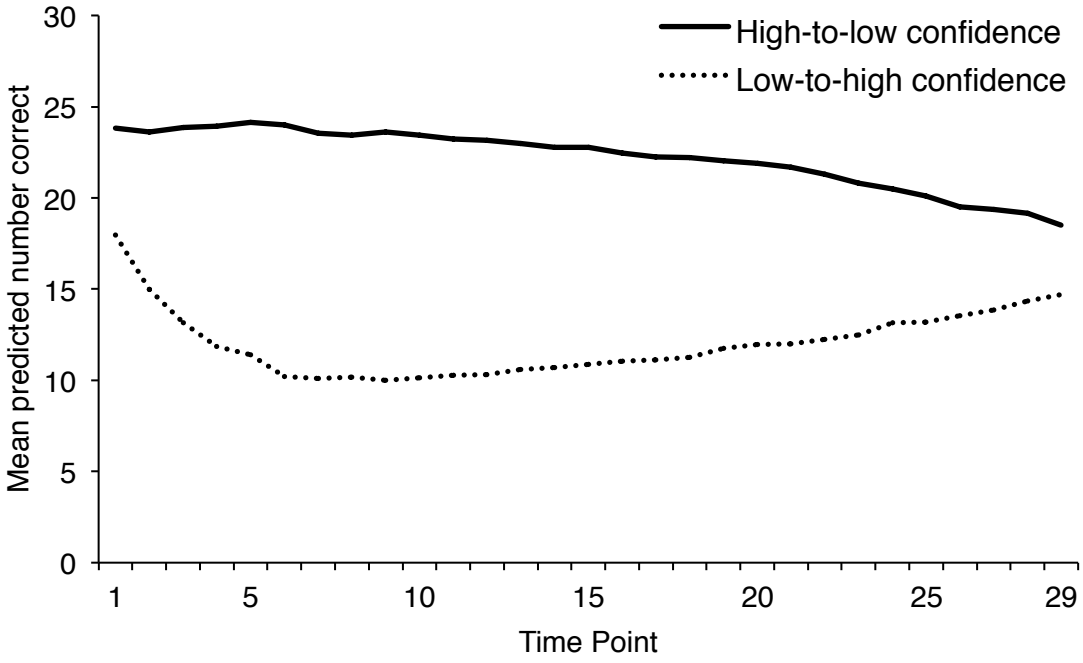

Figure 3. Mean predicted test scores on a recognition test as a function of time and question arrangement.

Experiment 2

Method

Subjects. We recruited 200 Mechanical Turk workers. Two subjects were excluded due to missing data.

Design and Procedure. The design and procedure was the same as Experiment 2 from Part 1—using a cued recall test rather than a recognition test—except that we also incorporated the continuous prediction questions from Experiment 1.

Results and Discussion

We again examined the mean predicted test scores people reported after each test question; these data appear in Figure 4. This figure looks remarkably similar to Figure 3 from Experiment 1, bolstering the claim that the influence of a question depends on the difficulty of that question, and when that question appears. Moreover, these data show that this influence is consistent across different question formats. That consistency is important for three reasons. First, it suggests reliability. Second, it shows that question arrangement wields influence even when overall difficulty changes, because cued recall is a more difficult task than recognition.

Third, it suggests that the influence of question arrangement is plausible in field environments that typically rely on cued recall rather than recognition tests.

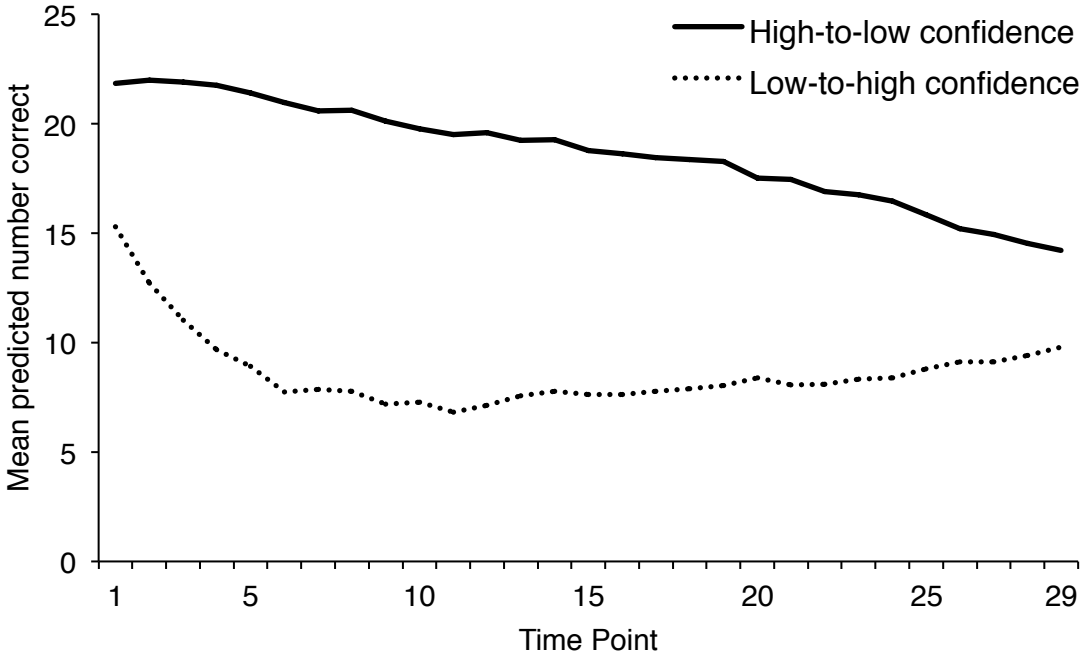

Figure 4. Mean predicted test scores on a cued recall test as a function of time and question arrangement.

As in Experiment 1, High-to-low confidence subjects' initial estimates were high (M_{Time1} = 21.85, SD_{Time1} = 5.66) and descended steadily over the course of the test (M_{Time30} = 13.55, SD_{Time30} = 6.54). But for low-to-high confidence subjects the pattern was different, consistent with Experiment 1. Low-to-high confidence subjects' initial estimates were already much lower than the high-to-low confidence subjects' (M_{Time1} = 13.31, SD_{Time1} = 5.66) and continued to drop until reaching their lowest point after the eleventh question (M_{Time11} = 6.82, SD_{Time11} = 5.97), at which point they ascended steadily over the remainder of the test (M_{Time30} = 10.09, SD_{Time30} = 4.37).

Taken together, the patterns in Experiments 1 and 2 are difficult to reconcile with an explanation relying on the anchoring-and-adjustment heuristic. According to the heuristic, the ease or difficulty of early questions provides an anchor that constrains the adjustments people make to their evaluations over the remainder of the test. But that mechanism alone cannot account for the finding that an identical question produces a large change in people's evaluations when it is encountered

early yet almost no change when it is encountered late. More specifically, an explanation that relies purely on the anchoring-and-adjustment heuristic would not predict the asymmetric patterns of changing estimates visible in Figures 3 and 4.

The patterns are instead reminiscent of a process purportedly responsible for the effects seen in impression formation, whereby early information sets an expectation about future information, which tends to be discounted when it does not fit with the expectation. If that process is indeed responsible for the influence of question arrangement, then we would expect that people who first answer easy questions will discount the informational value of the later difficult questions, and that people who first answer difficult questions will discount the information value of the later easy questions. More concretely, people who first answer easy questions would develop an initial impression that they are performing well and, due to discounting, the later difficult questions cannot divorce them fully from that impression. Similarly, people who first answer difficult questions would develop an initial impression that they are performing poorly and, due to discounting, the later easy questions cannot divorce them fully from that impression.

How might we test the extent to which impression formation processes are responsible for the biasing influence of question arrangement? One way would be to vary the *coherence* of the information people receive. We know from the literature that the processes responsible for primacy effects in impression formation are diminished when the same set of information is attributed to a group of people rather than a single individual (Hamilton & Sherman, 1996; McConnell et al., 1994, 1997). That difference is due to our underlying assumptions about the nature of individuals and groups. When it comes to an individual, we have learned to expect a strong degree of consistency—or coherence—to their behaviour (Schneider, 1973; Todd & Rapporport, 1964). For example, if I meet John for the first time and discover that he is friendly, then I expect John's other attributes will fit with that trait; it would not make sense for John to also be callous. But when it comes to a group, this expectation of coherence is not as strong, because we have learned that groups can consist of individuals who vary in their behaviours and traits (Hamilton & Sherman, 1996). For example, if I discover that John is a member of my gym, then I do not necessarily expect that when I meet Jane—who is also a member of my gym—that

she will have traits similar to John's; it is entirely reasonable for Jane to be callous, because the members of a gym probably have little in common other than wanting a place to exercise.

This difference in our expectations about the coherence of individuals and groups leads us to process incoming information about these two types of targets differently. With an expectation of coherence in individuals, we form an initial impression and attempt to integrate later information into that impression. But because this expectation is either not present or not as strong for groups, we encode incoming information about the group but do not need to form an integrated impression. If we are then asked to report our impressions of individuals or groups, our response is a consequence of these processing differences: For individuals, our impression—formed "on-line" as we integrate incoming information—typically shows a primacy bias; For groups, our impression—formed from memory when we are asked to report—typically shows a recency bias (Hamilton & Sherman, 1996; Hastie & Park, 1986; McConnell et al., 1994, 1997).

If the influence of question arrangement is a consequence of an integrative impression formation process, then here we have a potential manipulation that could diminish or eliminate that process: Modifying the test so that it appears as though it is a group of tests, rather than a single individual test. If people form impressions about a group of tests similar to how they form impressions about a group of people, we should expect that the pattern of results will no longer resemble a primacy bias, and might instead resemble a recency bias. We ran Experiment 3 to test this idea.

Experiment 3

Method

Subjects. We aimed to collect 400 data points, and ultimately recruited 419 Mechanical Turk workers.

Design and Procedure. We used a 2 (Question Order: high-to-low confidence, low-to-high confidence) x 2 (Test Coherence: individual, grouped) between subjects design. The procedure was the same as Experiment 1 from Part 1, except as follows. For half the subjects, we manipulated the test so that it now appeared to be three separate tests of 10 questions, rather than a single test of 30 questions. First, we

changed the wording of the instructions preceding the test from "You will now take a memory test", to "You will now take three separate memory tests: Test 1, Test 2, and Test 3." Second, we added a heading above each individual test question that read "TEST 1" (or 2, or 3, as appropriate). Third, we added a section break between Tests 1 and 2, and Tests 2 and 3, which read "Thank you. That's the end of Test [1/2]. Please click Next to start Test [2/3]." We further distinguished the 3 tests by displaying each in a unique combination of font face, colour, and style. Specifically, Test 1 appeared in the font Arial, the colour red, and in bold (e.g., **TEST 1**); Test 2 appeared in the font Times New Roman, the colour green, and in italics (e.g., *TEST 2*); Test 3 appeared in the font Verdana, the colour blue, and in both bold and italics (e.g., ***TEST 3***).

As a manipulation check, we asked subjects the following question at the end of the experiment: "How many memory tests were there about Eric the Electrician?" Subjects responded with a number. We also requested—via an optional textbox—that subjects tell us anything else they noticed about the memory tests.

Results and Discussion

We first carried out a manipulation check by examining subjects' responses to the question about the number of memory tests. Initial casual inspection of these data revealed that just over half the subjects (n = 217, 52.04%) misunderstood the question. These subjects all responded with the number 30, probably because they incorrectly thought the question was asking, "How many individual questions were on the memory test?" A much smaller proportion of subjects gave a clearly incorrect answer (n = 27, 6.48%). The remainder (n = 173, 41.49%) gave the correct answer. Because the pattern of results was consistent when including or excluding subjects who misinterpreted the question or simply answered it incorrectly, we included all subjects in our analyses. We also found that question arrangement and test appearance only trivially influenced overall test performance, and the two factors did not interact. The accuracy differences across the cells in the design ranged from virtually nothing (0.0009) to a maximum of approximately half a question (0.48); All Fs < 1.77.

We now turn to our primary question: How does manipulating the coherence of a test change the influence of question arrangement? To answer that question, we

once again examined subjects' bias scores and reports of memory confidence, classified according to the order of questions and the appearance of the test. We display these data in Figure 5. As the left side of the top panel of the figure shows, we replicated the typical finding whereby question arrangement influences people's beliefs about their test performance, M_{diff} = 1.51, 95% CI [0.11, 2.92]. But recall that we expected that manipulating the test so that it appeared as though it was a group of 3 tests would diminish the influence of question arrangement. The right side of the top panel of the figure shows that, if anything, the opposite was true—the influence of question arrangement was larger when the test appeared as though it was a group of 3 tests rather than a single test, M_{diff} = 3.08, 95% CI [1.61, 4.56]. We state this finding with caution, however, because the confidence intervals for these two results overlap considerably, meaning that zero is included in the range of reasonable estimates for the true size of the difference. In null-hypothesis terms, we found only a main effect of Question Order: M_{Diff} = 2.30, 95% CI [1.28, 3.31]; $t(417)$ = 4.45, $p < .001$.

The bottom panel of the figure displays people's reported memory confidence, and provides additional, albeit limited support for the counterintuitive idea that question arrangement is more influential when a test appears as though it is a group of tests. The left part of the panel shows that—unlike the results from our previous experiments—question arrangement did not reliably influence people's confidence in their memory, M_{diff} = 0.03, 95% CI [-0.23, 0.30]. The right part of the panel shows that question arrangement was slightly more influential when the test appeared as though it was a group of tests, M_{diff} = -0.18, 95% CI [-0.44, 0.08]. We state this finding with caution too, because the confidence intervals all include zero as a plausible value. In addition, we found that people reported greater confidence in their memory when the test appeared as a single test than when the test appeared to be a group of 3 tests, M_{diff} = 0.19, 95% CI [0.00, 0.37]. In null-hypothesis terms, we found only a main effect of Coherence; $t(417)$ = 1.97, p = .050.

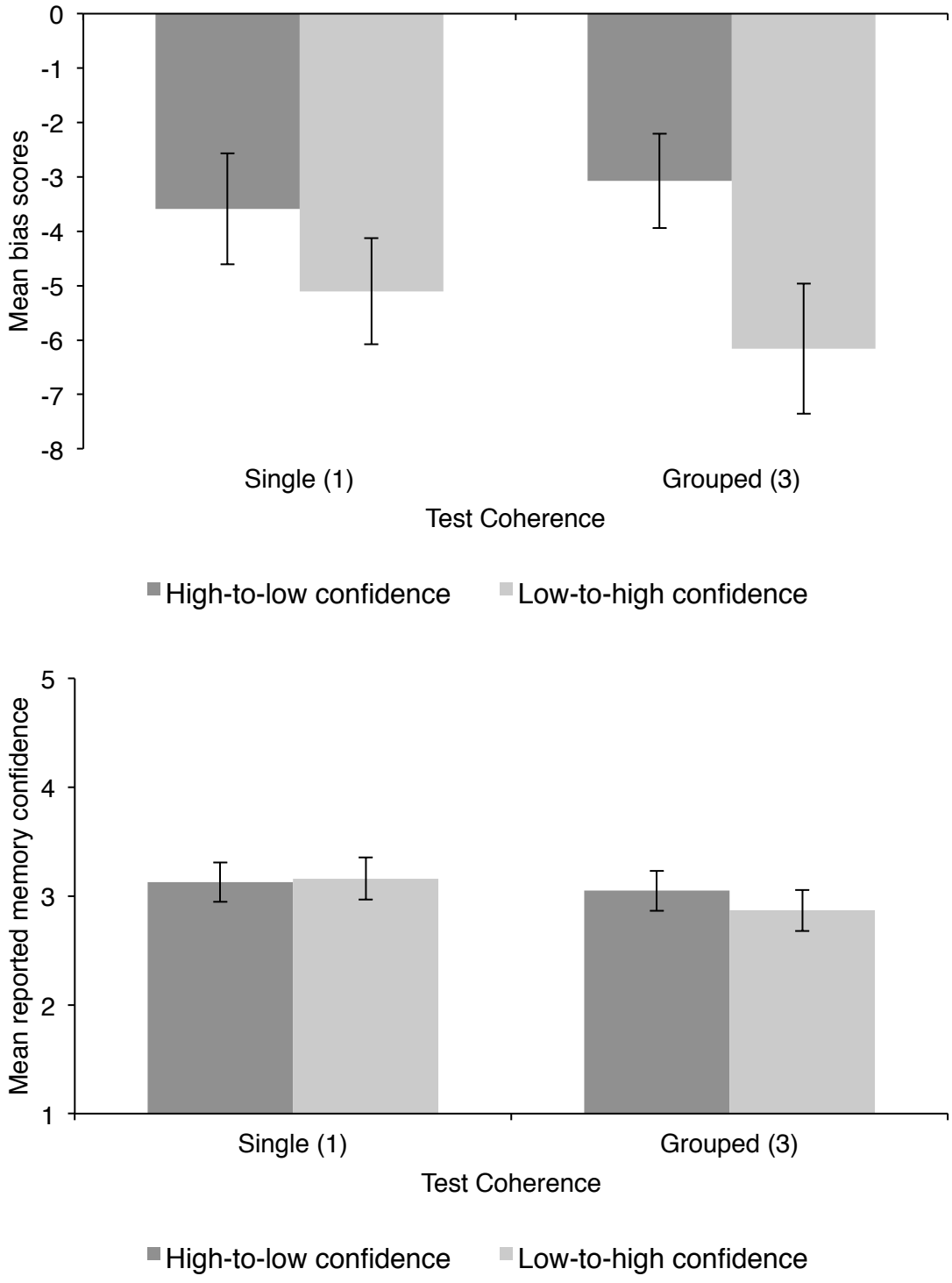

Figure 5. *Top panel:* Mean bias scores classified by Question Order (High-to-low confidence, Low-to-high confidence) and Test Coherence (Single, Grouped). *Bottom panel:* Mean reported memory confidence classified by Question Order (High-to-low confidence, Low-to-high confidence) and Test Coherence (Single, Grouped).

What are we to make of these results? On the one hand, the findings partially replicate our earlier experiments, showing that question arrangement influences

people's beliefs about their test performance. But on the other hand, we did not replicate our earlier findings with respect to memory confidence. How can we explain that null result? One possibility is that the confidence judgement is less prone to the influence of question arrangement, because it is less tightly coupled to that manipulation than the retrospective test estimate (Greifeneder, Bless, & Pham, 2011). Confidence differences might therefore be less reliable, resulting in the occasional sample with a null result. Alternatively, the result could simply reflect sampling variability. Regardless of the true explanation, the more alarming result is the apparent backfiring of our coherence manipulation. Contrary to what we expected, question arrangement seemed to be slightly *more* influential when the test appeared as though it was a group of tests rather than an individual test. How can we explain those results? One possibility is that our manipulation was not strong enough, such that subjects did not perceive the grouped version of the test as three distinct entities that were different from one another. If so, then subjects may have continued to form an integrated impression and would therefore show the typical pattern of results that is consistent with a primacy bias. In Experiment 4, we used an alternative manipulation in an effort to address this counter-explanation.

Experiment 4

Method

Subjects. We aimed to collect 600 data points and ultimately recruited 609 Mechanical Turk workers.

Design and Procedure. We used a 2 (Question Order: high-to-low confidence, low-to-high confidence) x 3 (Sources: Unspecified, One, Thirty) between subjects design. The procedure was identical to that of Experiment 1 from Part 1, except as follows. The two Unspecified source groups served as replication conditions of the standard question arrangement manipulation. For the remaining groups, we provided subjects with an additional piece of information in the instructions before the test. We told the One source groups, "All the questions on this test were written by one person." We told the Thirty sources groups, "Each question on this test was written by a different person." These statements appeared in bold in an effort to make them more noticeable. In addition, for these groups of subjects every question

on the memory test was prefaced by the name of a person who was the ostensible source of the question. For example, subjects in the One source groups saw "Michael Thomas asks:" directly above each test question. Subjects in the Thirty sources groups also received this source information above each question, but every question was attributed to a different person. To construct the thirty names required for the Thirty sources groups, we searched Wolfram|Alpha for the 30 most popular first names (half male, half female) and 30 most popular surnames in the United States (Wolfram|Alpha, 2016). Surnames were then randomly assigned to first names, to create a final list of 30 names. For the complete list of these names, see Appendix E. Finally, after subjects in the One or Thirty sources groups had reported their estimated test scores and memory confidence, we asked them the following additional question: "How many people constructed the memory test about Eric the Electrician?" Subjects responded with a number.

Results and Discussion

We first carried out a manipulation check by examining subjects' responses to the question about the number of people who constructed the memory tests. As in Experiment 3, initial casual inspection of these data revealed that approximately half the subjects either misinterpreted the question or simply answered it incorrectly (n = 193, 47.42%). Closer inspection suggested that when people were uncertain, they merely guessed a number, because there were no clear patterns to the incorrect responses. Because the pattern of results was consistent when including or excluding these subjects, we included all subjects in our analyses. We also found that the order of questions and the number of sources had little effect on overall test performance, and the two factors did not interact. The accuracy differences across the cells in the design ranged from virtually nothing (0.01) to a maximum of less than one question (0.83); All Fs < 1.51.

We now turn to our primary question: How does manipulating the coherence of a test—by making the questions appear to come from one single individual or from a group of 30 individuals—change the influence of question arrangement? To answer that question, we once again examined subjects' bias scores and reports of memory confidence, classified according to the order of questions and the number of sources. We display these data in Figure 6. As the left side of the top panel of the

figure shows, we replicated the typical finding whereby question arrangement influences people's beliefs about their test performance, M_{diff} = 2.82, 95% CI [1.29, 4.35]. We expected that when the test questions all ostensibly came from one person a similar pattern would emerge, but that when each test question ostensibly came from a different person the influence of question arrangement would be diminished. But as the rest of the top panel of the figure shows, the results are not entirely consistent with our predictions. On the one hand, the Thirty sources conditions showed a smaller influence of question arrangement, M_{diff} = 1.87, 95% CI [0.45, 3.29]. But on the other hand, the One source conditions showed an even smaller influence, M_{diff} = 1.22, 95% CI [-0.38, 2.82]. Moreover, confidence intervals across these differences overlapped considerably, suggesting that the influence of source was plausibly negligible. In null-hypothesis terms, we found only a main effect of Question Order, M_{diff} = 1.97, 95% CI [1.10, 2.84]; $t(607) = 4.45$, $p < .001$.

As the bottom panel of the figure shows, the findings with respect to reports of memory confidence were also somewhat inconsistent with our predictions. The left part of the panel shows that we found a small difference in confidence using our standard question arrangement paradigm, where question source is unspecified, M_{diff} = 0.18, 95% CI [-0.08, 0.44]. But the confidence interval suggests that the true size of this difference might plausibly be zero, and so this finding must be interpreted cautiously. Consistent with our predictions, the Thirty sources conditions showed a smaller influence of question arrangement, M_{diff} = 0.04, 95% CI [-0.20, 0.28]. But as with bias, and inconsistent with our predictions, the One source conditions showed an even smaller influence—in fact, on average, no difference at all—M_{diff} = 0.00, 95% CI [-0.27, 0.27]. In null-hypothesis terms, we found no statistically significant effects; All Fs < 2.48.

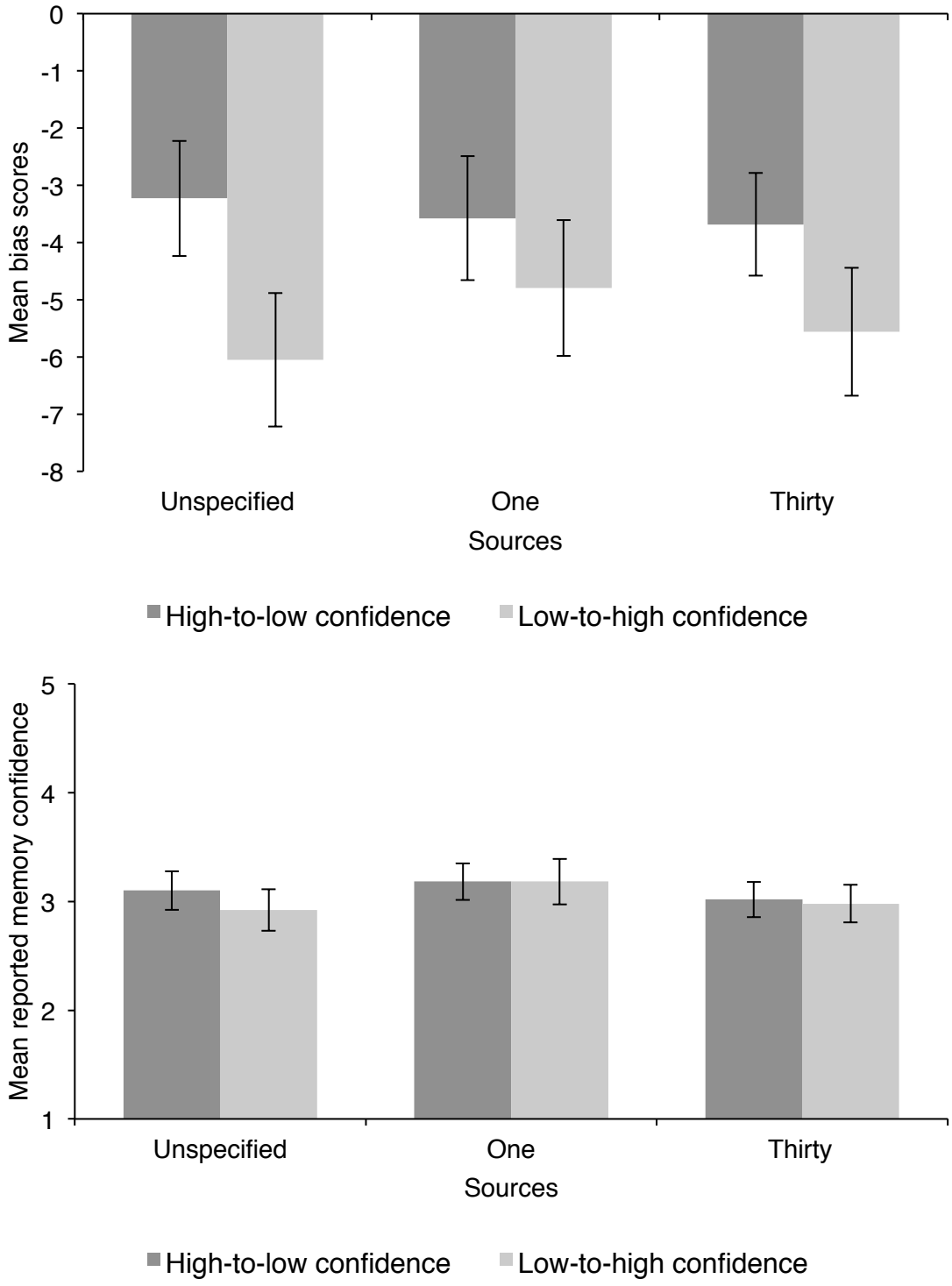

Figure 6. *Top panel:* Mean bias scores classified by Question Order (High-to-low confidence, Low-to-high confidence) and Sources (Unspecified, One, Thirty). *Bottom panel:* Mean reported memory confidence classified by Question Order (High-to-low confidence, Low-to-high confidence) and Sources (Unspecified, One, Thirty).

Taken together, this set of findings partially replicates our earlier work. Question arrangement influenced people's beliefs about their test performance—a

finding that appears to replicate fairly consistently. But question arrangement showed a weaker influence—and plausibly no influence at all—on people's confidence in their memory. As discussed in Experiment 3, that weakened influence could be because people's impressions about the test are less relevant to a judgement about the quality of their memory than they are to a judgement about test performance (Greifeneder et al., 2011). Specifically, the test experience is probably informative about test performance, but might not be informative about the quality of memory. For example, imagine a memory test where all the questions are about minute details that most people pay no attention to. The test experience will be informative about your test performance, but it says little about the quality of your memory, because you were never questioned about most of what you remember. In summary, the degree of relevance between the impression people develop over the course of the test and a judgement they make could moderate the impression's influence on that judgement.

In contrast to Experiment 3, but somewhat consistent with our predictions, we found that a less coherent test—in this instance, a test where each question came from a different person—resulted in a somewhat weaker influence of question arrangement on people's beliefs about their test performance and confidence in their memory. Those findings could suggest that manipulating the number of sources was more successful in making the test seem like a group than manipulating the appearance of the test, and therefore diminished impression formation processes. But there are at least two reasons why we should not draw any strong conclusions from these findings. First, the confidence intervals around the differences across source conditions overlap considerably, and thus the manipulation plausibly does nothing at all. Second, both source conditions showed reduced influence of question arrangement relative to an unspecified source condition. If our hypothesis was correct—that impression formation processes are diminished or eliminated when the test appears more like a group than an individual—then we should only see a reduction in the influence of question arrangement in the Thirty sources conditions, and not in the One source conditions. Third, the One source conditions showed the greatest reduction in the influence of question arrangement—precisely the opposite of what we would expect to see.

Taken together, the findings from Experiments 3 and 4 provide at best only limited support for the explanation that the influence of question arrangement is due to an impression formation process, whereby incoming information is integrated into an initial expectation. We are left with two possibilities: [1] The influence of question arrangement is not the result of an impression formation process, and is instead the result of alternative mechanism(s), or [2] The influence of question arrangement is the result of an impression formation process, but our manipulation failed to adequately break down that process.

In our final experiment, we used a different manipulation that we predicted would moderate the degree to which people integrate incoming information. Specifically, we manipulated people's initial expectations about test performance. We told some subjects that they should expect to perform extremely well on the test, and we told others that they should expect to perform extremely poorly. If this information sets an initial expectation, then incoming information will be integrated more when it fits with that expectation, and less when it does not. For example, if I expect to get most questions right and the initial questions feel very easy, then I will integrate that information and rapidly form an impression that I am performing excellently—discounting later, difficult questions because they do not fit with my developing impression. If, however, I expect to get most questions right and the initial questions feel very difficult, I will do the opposite—discounting early, difficult questions and integrating later, easier questions.

Manipulating people's expectations about their upcoming performance can also be thought of as providing people with an explicit anchor. If people use the expectation of performance as an anchor, then we should see people's estimates of test performance and reported memory confidence skewed towards these anchors. Specifically, people who are told to expect a high level of performance should show greater optimism in their test estimates and higher reported confidence in their memory than people who are told to expect a low level of performance.

In summary, this manipulation leads to different predictions depending on the putative mechanism underlying the influence of question arrangement. If people rely on the anchoring-and-adjustment heuristic, then the influence of question arrangement should remain constant—the manipulation should simply skew all

people equally towards the given anchor. But if people rely on impression formation processes, then the influence of question arrangement should be reduced or even eliminated when people are given an initial anchor, because that anchor sets an expectation and changes which information gets integrated.

Experiment 5

Method

Subjects. We aimed to recruit 600 subjects, and ultimately recruited 625 Mechanical Turk workers.

Design and Procedure. We used a 2 (Question Order: high-to-low confidence, low-to-high confidence) x 3 (Expectation: Unspecified, Low, High) between subjects design. The procedure was identical to that of Experiment 1 from Part 1, except as follows. The two Unspecified expectation groups served as replication conditions of our standard question arrangement manipulation. For the remaining groups, we provided subjects with an additional piece of information in the instructions before the test. We told the Low expectation groups, "Please note: We find that people answer only about 10% of these questions correctly." We told the High expectation groups, "Please note: We find that people answer just about 90% of these questions correctly." These statements appeared in bold in an effort to make them more noticeable. Finally, after subjects in the Low or High expectation groups had reported their estimated test scores and memory confidence, we asked them the following additional question: "Can you recall what percentage of the questions about the video people normally answer correctly? If you're unsure, please just take a guess." We asked subjects in the Unspecified expectation groups the following alternative question: "What percentage of the questions about the video do you think people normally answer correctly?" Subjects responded with a number between 0 and 100.

Results and Discussion

We first carried out a manipulation check by examining subjects' responses to the question about the percentage of test questions normally answered correctly. Of those subjects who were told this number in the instructions, 191 (30.66%) answered incorrectly, with answers widely distributed. Because the pattern of results was

consistent when including or excluding these subjects, we included all subjects in our analyses. We also found that the order of questions and the anchor had little effect on overall test performance, and the two factors did not interact. The accuracy differences across the cells in the design ranged from virtually nothing (0.01) to a maximum of less than one question (0.85); All *F*s < 2.16.

We now turn to our primary question: How does manipulating an initial expectation of test performance change the influence of question arrangement? To answer that question, we once again examined subjects' bias scores and reports of memory confidence, classified according to the order of questions and the anchor. We display these data in Figure 7. As the left side of the top panel of the figure shows, we replicated the typical finding whereby question arrangement influences people's beliefs about their test performance, M_{diff} = 2.72, 95% CI [1.25, 4.20]. The middle portion of the top panel shows that question arrangement continued to influence people's beliefs about their test performance even when they were expecting to perform poorly, M_{diff} = 1.83, 95% CI [0.35, 3.30]. The right portion of the top panel shows that the same is true when people were expecting to perform well, M_{diff} = 4.15, 95% CI [2.67, 5.63]. In null-hypothesis terms, we found a main effect of Question Order: M_{diff} = 2.90, 95% CI [2.05, 3.75]; $t(621) = 6.68$, $p < .001$.

In addition, we found that bias scores were skewed towards anchors. Subjects given the low expectation were most pessimistic, M_{Low} = -5.54, 95% CI [-4.80, -6.29]. Subjects given no expectation were slightly less pessimistic, $M_{Unspecified}$ = -4.36, 95% CI [-3.60, -5.12]. Subjects given the high expectation were least pessimistic, M_{High} = -4.18, 95% CI [-3.39, -4.97]. In null-hypothesis terms, we found a main effect of Expectation: $F(2, 621) = 3.80$, $p = .023$. Follow-up Tukey tests showed that only the difference between Low and High expectation subjects was statistically significant, M_{diff} = 1.35, 95% CI [0.10, 2.60], $p = .030$.

As the left side of the bottom panel of the figure shows, we also replicated the typical finding whereby question arrangement influences people's confidence in their memory, M_{diff} = 0.34, 95% CI [0.10, 0.58]. This pattern was not apparent when people were expecting to perform poorly, as shown in the middle portion of the panel, M_{diff} = 0.03, 95% CI [-0.21, 0.26]. The right portion of the panel, however, shows that question arrangement continued to influence people's reports of memory

confidence when they were expecting to perform well, M_{diff} = 0.33, 95% CI [0.07, 0.59]. Given the considerable overlap of these confidence intervals, these findings must be interpreted cautiously. Collapsing across question arrangement, we found no strong evidence that reports of memory confidence were skewed towards anchors—all confidence intervals overlapped considerably. In null-hypothesis terms, we found only a main effect of Question Order: M_{diff} = 0.23, 95% CI [0.09, 0.38]; $t(623) = 3.23$, $p = .001$.

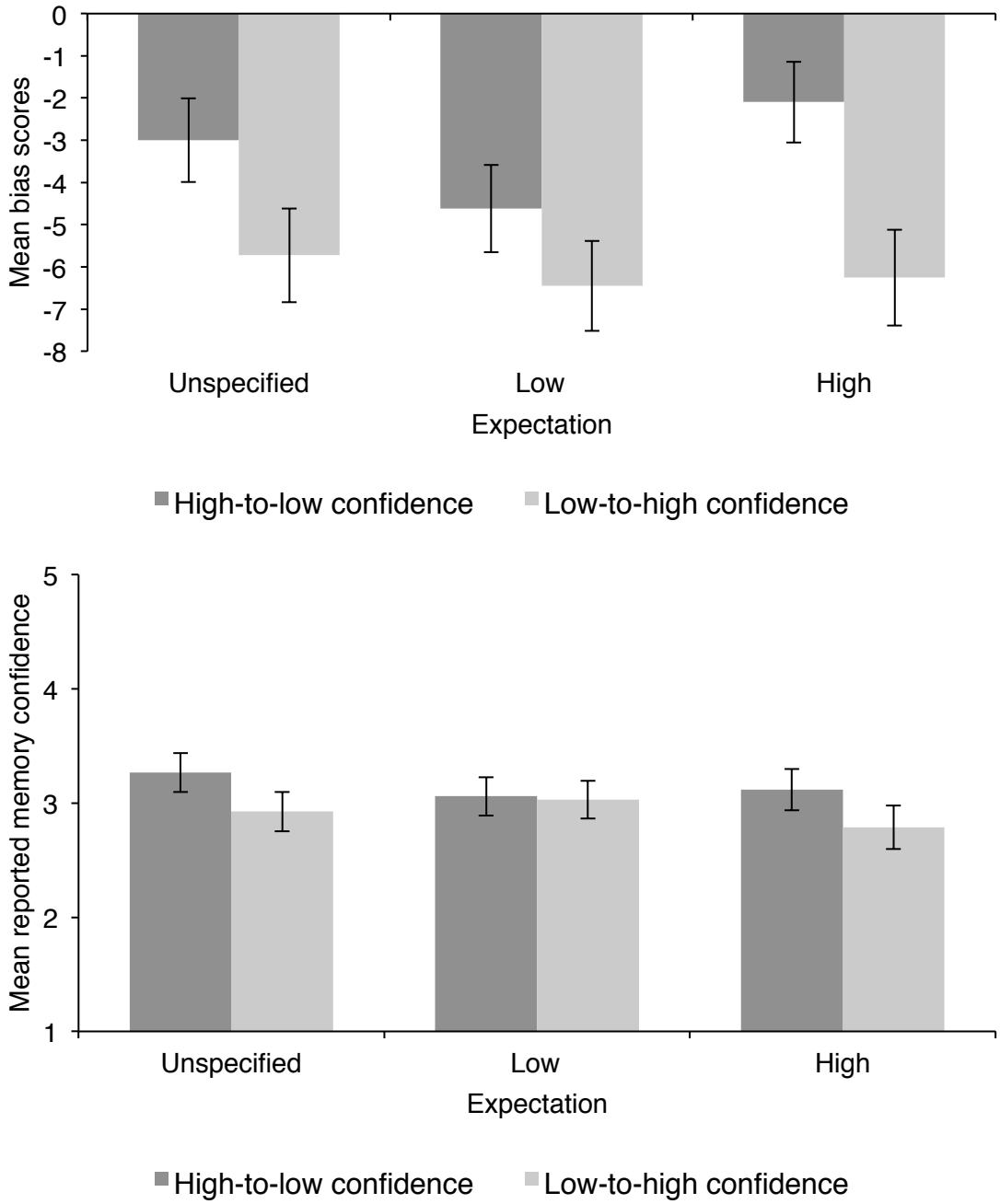

Figure 7. *Top panel:* Mean bias scores classified by Question Order (High-to-low confidence, Low-to-high confidence) and Expectation (Unspecified, Low, High). *Bottom panel:* Mean

reported memory confidence classified by Question Order (High-to-low confidence, Low-to-high confidence) and Expectation (Unspecified, Low, High).

Taken together, these results are more consistent with an anchoring-and-adjustment explanation than an impression formation explanation. However, the findings are far from conclusive. Lending weight to the anchoring-and-adjustment explanation is the fact that—for bias scores—the influence of question arrangement was present regardless of the presence or absence of an anchor, and that the scores were skewed towards anchors. But on the other side of the scale, we note that these same patterns were not consistently present in people's reports of memory confidence. Moreover, visual inspection of the top panel of Figure 7 suggests that the anchoring-and-adjustment explanation may be inadequate, because people who first answer easy questions seem more prone to the influence of an anchor than people who first answer difficult questions. Note, however, that we found at best only weak evidence for this interaction, Question Order x Expectation: $F(2, 619) = 1.99$, $p = .137$.

Chapter 3

Across 5 experiments, we examined the mechanisms responsible for the influence of question arrangement on people's beliefs about their memory. In Experiment 1, we showed that when the order of questions on a memory test is rearranged symmetrically, it produces asymmetrically developing beliefs about test performance. Subjects who began with easy questions initially believed they were performing well, and made only minor adjustments to this belief over the course of the test. But subjects who were given the same questions in the opposite order did not show a simple reversal of that pattern. Instead, subjects who began with difficult questions made dramatic changes to their initial beliefs about their test performance, before returning to a pattern of minor adjustments. We replicated these findings in Experiment 2, using a cued recall test. In Experiments 3 and 4, we examined the extent to which processes involved in forming impressions were responsible for the biasing influence of question order. We used two manipulations in an effort to make the test seem more like a group rather than an individual—manipulations that have been shown to prevent people from forming an impression that is unduly influenced by early information (Hamilton & Sherman, 1996; McConnell et al., 1994, 1997). But the results from both experiments are difficult to reconcile with an impression formation explanation. In Experiment 5, we used a manipulation that leads to different predictions for two hypothesised explanations of the biasing influence of question arrangement: the anchoring-and-adjustment heuristic, and impression formation. We gave some subjects an expectation that they would perform extremely well on the test, and other subjects an expectation that they would perform extremely poorly. The results were more consistent with an anchoring-and-adjustment explanation than impression formation.

Taken together, this package of experiments lends some support to an anchoring-and-adjustment explanation for the influence of question arrangement on people's beliefs about their memory. An initial anchor—provided by the experimenter, or generated from the subject's initial experience—guides the adjustments people make over the remainder of the test. Because accurate adjustments take time and effort, people tend to rely on a mental shortcut, adjusting only until they reach a plausible value (Epley & Gilovich, 2006; Fiske & Taylor,

2013). The resulting beliefs are therefore skewed towards the anchor. The findings from Experiment 5—where we manipulated anchors and found shifts towards those anchors—were somewhat consistent with this explanation.

But there are at least two reasons why this explanation remains incomplete. First, the findings from Experiment 5 were not entirely conclusive. Although the high-to-low confidence subjects displayed a pattern of results that supported an anchoring-and-adjustment explanation, the pattern was less apparent for the low-to-high confidence subjects. Second, an anchoring-and-adjustment explanation—in its typical formulation—would not predict the asymmetric patterns of developing beliefs in Experiments 1 and 2. Together, these findings suggest that an early experience of difficult questions is particularly influential.

But why are difficult questions influential only when they appear early? One reason could be that people have an expectation from learned experience that tests typically begin with easy questions. Difficult questions would then be especially surprising when encountered early, receiving relatively more cognitive processing than when those same difficult questions are encountered late. That relative boost in processing could explain why the same questions produce different degrees of adjustment depending on when they are encountered. Although we have no direct evidence that people expect tests to begin with easy questions, education research shows that it is commonly recommended to arrange tests this way—both to boost student confidence early on, and to ensure students under time pressure do not miss the difficult questions—even though evidence is mixed with respect to whether question arrangement has any real influence on test scores (Aamodt & McShane, 1992; Hambleton & Traub, 1974; Newman, Kundert, Lane Jr, & Bull, 1988; Sax & Cromack, 1966).

Of course, a slightly modified anchoring-and-adjustment explanation is not the only plausible explanation. Recall from Chapter 1 that there are a number of counter-explanations that would predict the typical pattern of results that we find.

The first counter-explanation is as follows: When making judgements about test performance and memory confidence, people might scan their memory of the test, and the information that comes to mind most easily would influence their judgements. Because people tend to rehearse early questions most, they could

preferentially recall those questions. This primacy effect could explain the influence of question arrangement. But research using the trivia questions paradigm suggests that this counter-explanation is unlikely, because people actually remember the last few questions best, and moreover, biases develop as the test progresses—not merely at the end (Franco, 2015; Weinstein & Roediger, 2012). However, it is possible that the differences between the trivia questions paradigm and our own make a primacy explanation more viable in our case. One key difference, for example, is the number of questions that subjects are asked. In the trivia paradigm, subjects typically answer 100 questions, but in our paradigm, subjects answer only 20 or 30 questions. Our smaller number means there is less opportunity—both in terms of time and amount of material—for later questions to compete with early questions in memory. This reduced *retroactive interference* makes a primacy effect more likely (Dey, 1969; Ecker, Brown, & Lewandowsky, 2015; McGeoch & McDonald, 1931). Because we never asked our subjects to recall test questions, we do not know whether a primacy effect is present. One simple future experiment could test this counter-explanation, by asking subjects to recall test questions as soon as they reach the end of the test. If the primacy effect is a valid explanation for the influence of question arrangement on eyewitness beliefs, then the questions subjects should most readily remember are the earliest ones.

The second counter-explanation is as follows: The influence of question arrangement could be the result of a particular mental shortcut—the affect heuristic—where people rely on their feelings to inform their judgements. We find this explanation unlikely, because everyone answers the same overall pool of questions. The explanation is therefore incomplete—it would need to state why early questions are stronger manipulators of affect than later questions. Moreover, research in the trivia questions paradigm suggests that subjects find both arrangements of the test equally enjoyable (Weinstein & Roediger, 2010). But because we never asked subjects for affective ratings, we cannot rule out this counter-explanation. Again, a simple future experiment could test this counter-explanation by asking subjects to rate how much they enjoyed the test, or how positive or negative they feel before and after the test. If the affect heuristic is a valid explanation for the influence of question arrangement on eyewitness beliefs, then

subjects who begin with easy questions should find the test more enjoyable, or feel more positively, than subjects who begin with difficult questions.

The third counter-explanation is as follows: Early questions set an expectation about performance, and people attempt to integrate later information in with this expectation, building an impression that is influenced most by early, consistent information. In fact, the pattern of developing beliefs in Experiments 1 and 2 led us to suspect that this explanation was likely. Those patterns looked as though early questions rapidly set an initial impression about test performance, and later questions that did not fit with that impression were discounted—an explanation that can account for the asymmetric results of Experiments 1 and 2. But when we attempted to test this explanation further in Experiments 3 and 4, the results were inconsistent with what we would expect if the explanation were valid. Naturally, we must exercise caution with any inferences we draw from those studies, because an "absence of evidence is not evidence of absence" (Oliver & Billingham, 1971, p5). More specifically, it is possible that the explanation is wrong, but it is also possible that the manipulation simply did not work. Anecdotally at least, it seems that the manipulation functioned as intended, because a number of people attributed the increasing or decreasing difficulty of questions to the separate tests, leaving comments like: "Test 3 was much harder than Test 1!" Finally, the results of Experiment 5 were more consistent with an anchoring-and-adjustment explanation.

Considered as a whole, the experiments here represent a novel contribution toward understanding the underlying mechanisms responsible for the influence of question arrangement on eyewitness beliefs. But at the same time, the puzzle is far from solved, and new questions have arisen. For instance, to be sufficient, modifications need to be made to the proposed explanations for these effects. What would those modifications tell us about the cognitive processes involved? Do they suggest multiple, additive processes, or processes that interact? Why does question arrangement influence some judgements consistently, like test performance—but other judgements less consistently, like memory confidence? Does that difference suggest anything about mechanism?

Overall, it appears that the most fruitful avenue for future exploration of mechanism lies with the anchoring-and-adjustment heuristic. After all, Experiment 5

suggests that people are sensitive to an initial anchor, and adjust away from it insufficiently. The reduced influence of the anchor for subjects who began with difficult questions suggests that these early, difficult questions compete with the anchor—possibly because they are surprising and highly influential. A future experiment could test that idea, by incorporating the design from Experiments 1 and 2 where subjects repeatedly predict their test performance. If the early, difficult questions compete with the anchor, we should see that low-to-high confidence subjects' predictions begin at the different anchor points, but come together rapidly in response to these highly influential questions. High-to-low confidence subjects, however, should maintain a degree of separation in their predictions across the course of the test.

This package of experiments has implications for our understanding of the processes underlying eyewitness metacognition and metamemory. Broadly, the findings hint at potential boundaries to the influence of the anchoring-and-adjustment heuristic. In particular, it could be that salient information encountered early enough interacts with the heuristic, strengthening or weakening its use. Moreover, our paradigm—using a series of questions—extends the literature on anchoring-and-adjustment, suggesting that adjustments can be made continuously as more and more information is encountered. Alternatively, the findings could hint at potential boundaries to the influence of an impression formation process. In particular, that the coherence or *entitativity* of a target might be more than a product of its constituent parts; a group of tests seemed to be treated as though it was still a single cohesive test (Hamilton et al., 2015). Our paradigm extends the literature on impression formation too, suggesting the possibility that the processes responsible for the impressions we develop of others could extend to the impressions we develop of ourselves.

Broadly, the findings extend what we know about factors that influence eyewitness memory. In tandem with the results from the experiments in Part 1, we can see that suggestive techniques are not a necessary component in the manufacturing of distorted eyewitness beliefs. Our seemingly trivial manipulation—merely flipping the order of a set of questions—produces changes in eyewitness and juror beliefs that are similar in magnitude to more heavy-handed manipulations.

Considered together, the results are reminiscent of other literatures that investigate the influence of the seemingly trivial on human behaviour, including the ease of processing (Alter & Oppenheimer, 2009), feelings more generally, (Greifeneder et al., 2011), the persuasiveness of neuro-jargon (Michael, Newman, Vuorre, Cumming, & Garry, 2013; Weisberg, Keil, Goodstein, Rawson, & Gray, 2008), and even our own expectations (Michael, Garry, & Kirsch, 2012).

Our findings could see potential application in field contexts, like police interviewing procedures. On the one hand, we might expect that current best-practice interviewing procedures, which encourage an early, rapport-building phase—particularly with children—are similar to starting a test with easy questions (Geiselman et al., 1984; Köhnken, Milne, Memon, & Bull, 1999; Memon, Cronin, Eaves, & Bull, 1993; Memon, Meissner, & Fraser, 2010). That practice might inadvertently inflate interviewees' beliefs about the quality of their memory. If so, it would be necessary to revise these best-practice techniques. Worse still, the results from Experiment 5 hint that this rapport-building practice might have its largest influence when people expect questions will be easy. But on the other hand, we know that building rapport helps interviewers extract more information from interviewees, and so it would be unwise to prematurely recommend any revision to current practice. Moreover, a rapport-building phase might be considered distinct from questions pertaining to a witnessed event, and might therefore have no influence at all.

Of course, it would be unwise to make premature recommendations. This research represents a first step in examining the influence of question order in an eyewitness context, and accordingly features a number of limitations. First, the controlled linearity of question order as it appears in these experiments is unlikely in a forensic setting, where questions shift more dynamically as an interview progresses. To the extent that such linearity produces biased judgements, it is possible that our findings overstate the influence of question order in field settings. Second, real jurors see an eyewitness during examination. Our jurors, in contrast, read an eyewitness's interview report. To the extent that the influence of order differs according to whether it takes place at interview or examination, we might expect to see an entirely different pattern of results in a field setting; perhaps by the

time the eyewitness takes the stand, it is a case of "too little too late." Finally, we used an eyewitness event that is relatively innocuous. But many witnessed events are highly emotional, and might be associated with an initial level of confidence that is resistant to the influence of a subtle manipulation like question order.

There are a number of important and interesting questions to address in future research. For example, just how far does the influence of question arrangement extend? We know from other eyewitness research that positive feedback about lineup decisions is dangerous. Not only does it boost people's confidence in their lineup decisions, it also causes people to re-evaluate their memory, reporting that they got a better view, paid more attention, saw the suspect for longer, and more (Douglass & Steblay, 2006; Wells & Bradfield, 1998). Worse still, jurors are persuaded by these artificially superior eyewitnesses (Douglass et al., 2010). On the one hand, then, we might expect that the influence of question arrangement will similarly cause eyewitnesses to re-evaluate their memory. But on the other hand, recall that—particularly in the experiments in Part 2—people's reports of memory confidence were not consistently affected. That instability could suggest that the influence of question arrangement depends on how closely the judgement matches the manipulation itself. Nonetheless, it would be useful to test this idea empirically.

Another important question relates to *the misinformation effect*—the extent to which people incorporate misleading information encountered after an event into their memories (Loftus, 2005). How would the influence of question arrangement affect people's propensity to the misinformation effect? Imagine an experiment using the basic paradigm from Experiment 1 in Part 1, but with the following key change: Between watching the events in the video and taking the memory test, subjects read another eyewitness's report about the events in the video. Included in that report are some misleading details. For example, if Eric drank a can of Coke, the report might claim that he drank a Pepsi. The question then, is how would the arrangement of questions on the memory test change people's propensity to incorrectly choose those misleading details as their answers? One possibility is that an early experience of easy questions lulls people into a false sense of security, believing that the questions are simple and require little thought. If so, then we might expect that high-to-low

confidence subjects would engage in less effortful monitoring of the source of information that comes to mind when answering the questions, and would therefore be more prone to the influence of misleading information than their low-to-high confidence counterparts (Johnson et al., 1993). An alternative possibility is that an early experience of difficult questions could make people think their memory is terrible, second-guessing themselves. If so, then we might expect that low-to-high confidence subjects would defer their answers in favour of whatever they read in the report, and would therefore be more prone to the influence of misleading information than their high-to-low confidence counterparts (Dodd & Bradshaw, 1980; Vornik, Sharman, & Garry, 2003). The implications in either hypothesised scenario are alarming, and so the question is worthy of attention.

In conclusion, the collection of results in Part 2 fit with those in Part 1, painting a worrying picture of eyewitness and juror beliefs about memory. The way we think about what we remember is prone to the influence of a manipulation that—at face value—is trivial. While we have yet to pin down the precise processes responsible for the influence of question arrangement, this work represents a first step in examining the underlying mechanisms. Perhaps most importantly, the package of experiments together provides strong evidence for the reliability of the influence of question arrangement, and paves the road for future exploration.

References

Aamodt, M. G., & McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management, 21*, 151-160. doi:10.1177/009102609202100203

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*, 129-133. doi:10.1037/0033-2909.97.1.129

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

Anderson, N. H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology, 2*, 1-9. doi:10.1037/h0021966

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., & Jacobson, A. (1965). Effect of stimulus inconsistency and discounting instructions in personality impression formation. *Journal of Personality and Social Psychology, 2*, 531-539. doi:10.1037/h0022484

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

Bell, B. E., & Loftus, E. F. (1988). Degree of detail of eyewitness testimony and mock juror judgments. *Journal of Applied Social Psychology, 18*, 1171-1192. doi:10.1111/j.1559-1816.1988.tb01200.x

Bell, B. E., & Loftus, E. F. (1989). Trivial persuasion in the courtroom: the power of (a few) minor details. *Journal of Personality and Social Psychology, 56*, 669-679.

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology, 72*, 691-695. doi:10.1037/0021-9010.72.4.691

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23*, 1208-1214. doi:10.1177/0956797612441217

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*, 11-30. doi:10.1037/1076-898X.12.1.11

Cahill, L., & McGaugh, J. L. (1995). A novel demonstration of enhanced memory associated with emotional arousal. *Consciousness and cognition, 4*, 410-421. doi:10.1006/ccog.1995.1048

Christianson, S., & Loftus, E. F. (1991). Remembering emotional events: The fate of detailed information. *Cognition and Emotion, 5*, 81-108. doi:10.1080/02699939108411027

Clarke, C., Milne, R., & Bull, R. (2011). Interviewing suspects of crime: The impact of PEACE training, supervision and the presence of a legal advisor. *Journal of Investigative Psychology and Offender Profiling, 8*, 149-162. doi:10.1002/jip.144

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

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*: Routledge.

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1990). Juror sensitivity to eyewitness identification evidence. *Law and Human Behavior, 14*, 185-191. doi:10.1007/Bf01062972

DeCoster, J., & Claypool, H. M. (2004). A meta-analysis of priming effects on impression formation supporting a general model of informational biases. *Personality and Social Psychology Review, 8*, 2-27. doi:10.1207/S15327957PSPR0801_1

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence. *Law and Human Behavior, 4*, 243-260. doi:10.1007/BF01040617

DeSoto, K. A., & Roediger, H. L. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science, 25*, 781-788. doi:10.1177/0956797613516149

Dey, M. K. (1969). Retroactive inhibition as a function of similarity of meaning in free-recall learning. *Psychologische Forschung, 33*, 79-84. doi:10.1007/Bf00424618

Dodd, D. H., & Bradshaw, J. M. (1980). Leading questions and memory: Pragmatic constraints. *Journal of Verbal Learning and Verbal Behavior, 19*, 695-704. doi:10.1016/S0022-5371(80)90379-5

Douglass, A. B., Neuschatz, J. S., Imrich, J., & Wilkinson, M. (2010). Does post-identification feedback affect evaluations of eyewitness testimony and identification procedures? *Law and Human Behavior, 34*, 282-294. doi:10.1007/s10979-009-9189-5

Douglass, A. B., & Steblay, N. (2006). Memory distortion in eyewitnesses: a meta‐analysis of the post‐identification feedback effect. *Applied Cognitive Psychology, 20*, 859-869. doi:10.1002/acp.1237

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*: University Microfilms.

Ecker, U. K., Brown, G. D., & Lewandowsky, S. (2015). Memory without consolidation: Temporal distinctiveness explains retroactive interference. *Cognitive Science, 39*, 1570-1593. doi:10.1111/cogs.12214

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*, 391-396. doi:10.1111/1467-9280.00372

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

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making, 13*, 1-17. doi:10.1002/(Sici)1099-0771(200001/03)13:1<1::Aid-Bdm333>3.0.Co;2-S

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture*: Sage.

Franco, G. (2015). *The order of questions on a test affects how well students believe they performed.* (Unpublished doctoral thesis), Victoria University of Wellington, Wellington, New Zealand.

French, L., Garry, M., & Mori, K. (2011). Relative–not absolute–judgments of credibility affect susceptibility to misinformation conveyed during discussion. *Acta Psychologica, 136*, 119-128. doi:10.1016/j.actpsy.2010.10.009

Geiselman, R. E., Fisher, R. P., Firstenberg, I., Hutton, L. A., Sullivan, S. J., Avetissian, I. V., & Prosk, A. L. (1984). Enhancement of eyewitness memory: An empirical evaluation of the cognitive interview. *Journal of Police Science and Administration, 12*, 74-80.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*, 592-596. doi:10.1037/0033-295X.103.3.592

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650-669. doi:10.1037/0033-295X.103.4.650

Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., Gretz, A. L., . . . Turpin, B. M. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 355-369. doi:10.1037/0278-7393.6.4.355

Greifeneder, R., Bless, H., & Pham, M. T. (2011). 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

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *The Journal of Experimental Education, 43*, 40-46. doi:10.1080/00220973.1974.10806302

Hamilton, D. L., Chen, J. M., Ko, D. M., Winczewski, L., Banerji, I., & Thurston, J. A. (2015). Sowing the seeds of stereotypes: Spontaneous inferences about groups. *Journal of Personality and Social Psychology, 109*, 569-588. doi:10.1037/pspa0000034

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review, 103*, 336-355. doi:10.1037/0033-295X.103.2.336

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

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

Innocence Project. (2016). The Causes of Wrongful Conviction. Retrieved from http://www.innocenceproject.org/causes-wrongful-conviction

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

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3-28. doi:10.1037/0033-2909.114.1.3

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. doi:10.1037/h0026818

Köhnken, G., Milne, R., Memon, A., & Bull, R. (1999). The cognitive interview: A meta-analysis. *Psychology, Crime and Law, 5*, 3-27. doi:10.1080/10683169908414991

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

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*, 361-366. doi:10.1101/lm.94705

Loftus, E. F. (2013, June). How reliable is your memory? [Video file]. Retrieved from https://www.ted.com/talks/elizabeth_loftus_the_fiction_of_memory?language=en

Loftus, E. F., Donders, K., Hoffman, H. G., & Schooler, J. W. (1989). Creating new memories that are quickly accessed and confidently held. *Memory & Cognition, 17*, 607-616. doi:10.3758/Bf03197083

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585-589. doi:10.1016/S0022-5371(74)80011-3

Loftus, E. F., & Zanni, G. (1975). Eyewitness testimony: The influence of the wording of a question. *Bulletin of the Psychonomic Society, 5*, 86-88. doi:10.3758/BF03336715

Marshall, P. H., & Werder, P. R. (1972). The effects of the elimination of rehearsal on primacy and recency. *Journal of Verbal Learning and Verbal Behavior, 11*, 649-653. doi:10.1016/S0022-5371(72)80049-5

McConnell, A. R., Sherman, S. J., & Hamilton, D. L. (1994). On-line and memory-based aspects of individual and group target judgments. *Journal of Personality and Social Psychology, 67*, 173-185. doi:10.1037/0022-3514.67.2.173

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

McGeoch, J. A., & McDonald, W. T. (1931). Meaningful relation and retroactive inhibition. *The American Journal of Psychology, 43*, 579-588. doi:10.2307/1415159

Memon, A., Cronin, O., Eaves, R., & Bull, R. (1993). The cognitive interview and child witnesses. *Issues in Criminological and Legal Psychology, 20*, 3-9.

Memon, A., Meissner, C. A., & Fraser, J. (2010). The Cognitive Interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law, 16*, 340-372. doi:10.1037/a0020518

Michael, R. B., Garry, M., & Kirsch, I. (2012). Suggestion, cognition, and behavior. *Current Directions in Psychological Science, 21*, 151-156. doi:10.1177/0963721412446369

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

Newman, D. L., Kundert, D. K., Lane Jr, D. S., & Bull, K. S. (1988). Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty. *Applied Measurement in Education, 1*, 89-97. doi:10.1207/s15324818ame0101_8

Oliver, B. M., & Billingham, J. (1971). *Project Cyclops: A design study of a system for detecting extraterrestrial intelligent life.* Paper presented at the The 1971 NASA/ASEE Summer Fac. Fellowship Program (NASA-CR-114445).

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*, 237-241. doi:10.1016/j.tics.2008.02.014

Paulo, R. M., Albuquerque, P. B., & Bull, R. (2013). The enhanced cognitive interview: Towards a better use and understanding of this procedure. *International Journal of Police Science and Management, 15*, 190-199. doi:10.1350/ijps.2013.15.3.311

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology, 51*, 242-258. doi:10.1037//0022-3514.51.2.242

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 521-533. doi:10.1037/0278-7393.14.3.521

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology, 62*, 189-206. doi:10.1037//0022-3514.62.2.189

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

Qualtrics. (2016). Qualtrics [survey software]. Provo, UT.

Risinger, D. M. (2007). Innocents convicted: An empirically justified factual wrongful conviction rate. *The Journal of Criminal Law and Criminology, 97*, 761-806.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775-777. doi:10.1037/0003-066X.45.6.775

Rundus, D. (1980). Maintenance rehearsal and long-term recency. *Memory and Cognition, 8*, 226-230. doi:10.3758/BF03197610

Sax, G., & Cromack, T. R. (1966). The effects of various forms of item arrangements on test perormance. *Journal of Educational Measurement, 3*, 309-311. doi:10.1111/j.1745-3984.1966.tb00896.x

Schneider, D. J. (1973). Implicit personality theory: A review. *Psychological Bulletin, 79*, 294-309. doi:10.1037/h0034496

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

Shiv, B., & Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research, 26*, 278-292. doi:10.1086/209563

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the US population. *PLOS ONE, 6*, e22757. doi:10.1371/journal.pone.0022757

Simons, D. J., & Chabris, C. F. (2012). Common (mis) beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *PLOS ONE, 7*, e51876. doi:10.1371/journal.pone.0051876

Takarangi, M. K., Parker, S., & Garry, M. (2006). Modernising the misinformation effect: The development of a new stimulus set. *Applied Cognitive Psychology, 20*, 583-590. doi:10.1002/acp.1209

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science, 18*, 46-50. doi:10.1111/j.1467-9280.2007.01847.x

Todd, F. J., & Rapporport, L. (1964). A cognitive structure approach to person perception: A comparison of two models. *The Journal of Abnormal and Social Psychology, 68*, 469-478. doi:10.1037/h0043314

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131. doi:10.1126/science.185.4157.1124

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): Oxford University Press.

Vornik, L., Sharman, S., & Garry, M. (2003). The power of the spoken word: Sociolinguistic cues influence the misinformation effect. *Memory, 11*, 101-109. doi:10.1080/741938170

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 and Cognition, 38*, 366-376. doi:10.3758/Mc.38.3.366

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

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*, 470-477. doi:10.1162/jocn.2008.20040

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360-376. doi:10.1037//0021-9010.83.3.360

Winkielman, P., Zajonc, R. B., & Schwarz, N. (1997). Subliminal affective priming resists attributional interventions. *Cognition and Emotion, 11*, 433-465. doi:10.1080/026999397379872

Wolfram|Alpha. (2016). Retrieved 13th November 2015, from Wolfram Alpha LLC http://www.wolframalpha.com

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

Index

Appendix A

Table 3. Demographic information about subjects in each experiment in Part 1.

Experiment	Condition	Males (%)	Females (%)	Mean age (SD)
1	High-to-low	21 (40)	31 (60)	32.31 (11.17)
	Low-to-high	21 (42)	29 (58)	32.76 (12.33)
2	High-to-low	47 (43)	63 (57)	34.93 (12.67)
	Low-to-high	34 (31)	76 (69)	34.53 (12.23)
3	High-to-low	42 (42)	57 (58)	33.13 (9.96)
	Low-to-high	41 (39)	65 (61)	34.40 (11.10)
4	High-to-low	59 (46)	70 (54)	30.58 (10.73)
	Low-to-high	54 (41)	77 (59)	29.66 (9.82)
5	High-to-low	63 (41)	90 (59)	34.01 (12.62)
	Low-to-high	67 (43)	89 (57)	33.58 (12.12)
6	High-to-low	64 (41)	91 (59)	35.71 (12.56)
	Low-to-high	68 (43)	91 (57)	32.76 (11.10)

Table 4. Demographic information about subjects in each experiment in Part 2.

Experiment	Condition	Males (%)	Females (%)	Mean age (SD)
1	High-to-low	52 (46)	60 (54)	34.49 (12.83)
	Low-to-high	44 (42)	62 (58)	33.52 (12.79)
2	High-to-low	37 (36)	66 (64)	31.61 (10.67)
	Low-to-high	30 (32)	65 (68)	32.75 (10.82)
3	High-to-low, 1 test	33 (32)	69 (68)	36.05 (11.88)
	Low-to-high, 1 test	32 (30)	75 (70)	36.94 (12.23)
	High-to-low, 3 tests	39 (38)	65 (63)	35.29 (10.66)
	Low-to-high, 3 tests	37 (35)	69 (65)	37.08 (13.09)
4	High-to-low, 1 source	34 (34)	65 (66)	33.92 (11.53)
	Low-to-high, 1 source	40 (40)	59 (60)	33.92 (11.94)
	High-to-low, 30 sources	36 (33)	72 (67)	33.74 (10.87)
	Low-to-high, 30 sources	38 (37)	64 (63)	34.38 (12.64)
5	High-to-low, No anchor	42 (40)	63 (60)	34.44 (10.82)
	Low-to-high, No anchor	41 (39)	64 (61)	34.85 (12.24)
	High-to-low, Low anchor	31 (30)	72 (70)	33.77 (12.13)
	Low-to-high, Low anchor	41 (39)	64 (61)	35.19 (11.98)
	High-to-low, High anchor	37 (36)	66 (64)	36.14 (12.45)
	Low-to-high, High anchor	33 (32)	71 (68)	35.39 (12.11)

Appendix B

Table 5. Question information for the 30-item test. Questions are listed in the order they appeared in the High-to-low confidence version; this order is reversed for the Low-to-high confidence version. Answer options in bold represent correct answers. For the eight items with no bolded answer, the correct answer depended on the version of the video subjects watched.

Question	Option 1	Option 2	Confidence Mean (SD)
Eric ate _____	**an apple**	a banana	4.79 (0.63)
Eric played a _____	Video	**CD**	4.73 (0.73)
In the bathroom Eric stole _____	**pills**	perfume	4.69 (0.82)
Eric was wearing _____	Overalls	**jeans**	4.62 (0.82)
Eric stole _____ in the second bedroom	Money	**a ring**	4.55 (0.98)
The jewelery that Eric stole in the first bedroom was _____	**Earrings**	a necklace	4.03 (1.28)
In the second bedroom, Eric tested a _____	**power point**	light fitting	4.00 (1.06)
Eric found the house key under a _____	door mat	**flower pot**	3.93 (1.41)
In the lounge the picture Eric looked at was the _____Tower	Eiffel	Leaning	3.91 (1.50)
In the lounge Eric looked through a _____	Journal	**photo album**	3.85 (1.37)
The bed in the first bedroom was _____	made	unmade	3.79 (1.29)
The tool that Eric used in the kitchen was _____	pliers	**screwdriver**	3.75 (1.30)
In the second bedroom, Eric tried on a _____ cap	black	blue	3.64 (1.39)
Eric read the note from the homeowner in the _____	Kitchen	**hallway**	3.64 (1.35)
The magazine that Eric read was _____	Time	Newsweek	3.56 (1.54)
When Eric closed the living room doors, it was the _____ door that he closed.	**left**	right	3.51 (1.59)
The curtains in the room where Eric worked on the light fitting were _____	**open**	closed	3.46 (1.25)
Eric drank a can of _____	coke	pepsi	3.33 (1.52)
Eric checked the time _____	on his watch	on the wall clock	3.18 (1.39)
The name of Eric's company was _____	AJ's electricians	RJ's electricians	2.91 (1.58)
The color of Eric's van was _____	**blue**	red	2.82 (1.41)

Question	Option 1	Option 2	Confidence Mean (SD)
The color of the flowers where Eric retrieved the key were _____	yellow	**pink**	2.82 (1.45)
When Eric sat down to watch the television, the book on the coffee table was _____	open	**closed**	2.59 (1.27)
When Eric played the CD, the candelabra was to his _____	left	**right**	2.30 (1.24)
Eric rummaged through papers that were next to a _____ mug	yellow	white	2.20 (1.13)
There were _____ remote controls on the coffee table.	two	**three**	2.08 (1.10)
The color of the rubbish bin in the kitchen was _____	**white**	grey	2.01 (1.11)
The fireplace in the second bedroom Eric visited was	**covered**	uncovered	1.99 (1.05)
The total number of pillows and cushions on the bed in the first room Eric visited was _____	four	**six**	1.83 (1.05)
There were _____ toothbrushes in the bathroom.	two	**three**	1.73 (1.06)

Appendix C

Table 6. Question information for the cued-recall variant of the 30-item test. Questions are listed in the order they appeared in the High-to-low confidence version; this order is reversed for the Low-to-high confidence version. Answers were marked correct when they featured a keyword.

Question	Keyword(s)
What did Eric eat?	apple
What type of media did Eric play?	cd
What did Eric steal from the bathroom?	pill
What style of trousers was Eric wearing?	jeans
What did Eric steal in the second bathroom?	ring
What type of jewelry did Eric steal in the first bedroom?	ear, ring
What did Eric test in the second bedroom?	socket, outlet, plug
What did Eric find the house key under?	plant, pot, flower
What was on the picture that Eric looked at in the lounge?	eiffel, leaning
What did Eric look over in the lounge?	photo, album
What state was the bed in the first bedroom in?	unmade, made
What tool did Eric use in the kitchen?	screwdriver
What color was the cap Eric tried on in the second bedroom?	blue, black
Where did Eric read the note from the homeowner?	hall
What magazine did Eric read?	news, time
Which side of the living room doors did Eric close?	left
In the room where Eric worked on the light fitting, what state were the curtains in?	open
What did Eric drink a can of?	pepsi, coke
What did Eric use to check the time?	watch, clock
What was the name of Eric's company?	aj, rj
What color was Eric's van?	blue
What color were the flowers where Eric retrieved the key?	pink
What state was the book on the coffee table in when Eric sat down to watch television?	closed
Which side of Eric was the candelabra on when he played the CD?	right
What color was the mug that was next to the papers Eric rummaged through?	yellow, white
How many remote controls were on the coffee table?	three, 3

Question	Keyword(s)
What color was the rubbish bin in the kitchen?	white
In the second bedroom Eric visited, what state was the fireplace in?	uncovered, covered
What was the total number of pillows and cushions on the bed in the first room Eric visited?	six, 6
How many toothbrushes were in the bathroom?	three, 3

Appendix D

Table 7. Question information for the 20-item test. Questions are listed in the order they appeared in the High-to-low confidence version; this order is reversed for the Low-to-high confidence version. Answer options in bold represent correct answers. For the eight items with no bolded answer, the correct answer depended on the version of the video subjects watched.

Question	Option 1	Option 2	Difficulty Mean (SD)
What sort of bag did Chad bring to the party?	Satchel	**Backpack**	1.61 (1.07)
What did Chad sort through at the small table in the lounge?	Cutlery	**CD's**	1.79 (1.16)
What did Chad add to the drink he made for the woman?	A squeeze of lemon	A vial of liquid	1.79 (1.29)
Where did Chad go when he was trying to find the toilet?	Laundry	Bedroom	1.84 (1.16)
Where did Chad put the wallet he found on the kitchen counter?	In his back pocket	Back where he found it	1.96 (1.37)
When Chad knocked over the drink, what did he clean up the spill with?	Paper towels	Dish cloth	2.13 (1.31)
What drink did Chad take out of his brown paper bag?	Vodka	Wine	2.23 (1.29)
What decoration was hanging over the doorway to the lounge?	Tinsel	Happy Birthday Banner	2.24 (1.44)
What did Chad eat on his way back from the toilet?	**Chips**	Carrots	2.62 (1.22)
What colour cup did the person who opened the front door have?	Pink	Blue	2.65 (1.51)
Where was Chad when his cell phone rang?	**Leaving the kitchen**	Leaving the bathroom	2.66 (1.15)
How did Chad get into the house?	Knocked on the door	Rang on the doorbell	3.04 (1.41)
When Chad found it, was the toilet seat up or down?	Up	**Down**	3.08 (1.40)
Were the curtains in the lounge closed or open?	**Closed**	Open	3.48 (1.37)
What colour balloons were hanging over the bathroom door?	Blue	Red	3.58 (1.47)
How many doors did Chad close at the party?	Three	**Four**	3.68 (1.16)
What colour was the kitchen counter?	**Yellow**	Green	4.00 (1.09)
What colour were the lounge walls painted?	**Peach**	Pink	4.02 (1.05)

Question	Option 1	Option 2	Difficulty Mean (SD)
What was the painting hanging over the drinks table of?	Daffodils	**Sunflowers**	4.21 (1.26)
How many jacket hooks were there beside the drinks table?	Four	**Five**	4.67 (0.74)

Appendix E

Table 8. List of names attributed as sources of the questions in Experiment 4.

Emma Smith
Noah Johnson
Olivia Williams
Liam Brown
Sophie Jones
Mason Miller
Isabelle Davis
Jacob Baker
Ava Young
William Wilson
Mia Allen
Ethan Anderson
Emily Taylor
Michael Thomas
Abigail Wright
Alexander Moore
Madison Martin
James Jackson
Charlotte Thompson
Daniel White
Harper King
Elijah Lee
Sofia Scott
Benjamin Harris
Avery Clark
Logan Lewis
Elizabeth Robinson
Aiden Walker
Amelia Green
Jayden Hall

Appendix F

Table 9. List of questions assessing compliance with general instructions. Each question required a Yes or No response.

Did you maximize the size of your web browser so that it covers your entire screen?
Did you complete the experiment in a single session, without stopping?
Did you pause or leave the experiment to engage in other tasks, even if they were other computer tasks?
Did you use your web browser's back or refresh buttons at any point during the experiment?
Did you complete the experiment in an environment that is free of noise and distraction?
Did you complete the experiment without anyone helping you?
Did you speak with anyone at any time during the experiment?
Please tell us whether you used a search engine at any point during the experiment to look anything up.