Volume 27 · Number 7 · August 2019

Memory

Editors Martin A. Conway and Mark L. Howe

Memory

Routledge
Taylor & Francis Group

ISSN: 0965-8211 (Print) 1464-0686 (Online) Journal homepage: https://www.tandfonline.com/loi/pmem20

How do ordered questions bias eyewitnesses?

Robert B. Michael & Maryanne Garry

To cite this article: Robert B. Michael & Maryanne Garry (2019) How do ordered questions bias eyewitnesses?, Memory, 27:7, 904-915, DOI: 10.1080/09658211.2019.1607388

To link to this article: https://doi.org/10.1080/09658211.2019.1607388

Published online: 17 Apr 2019.

MEMORY
2019, VOL. 27, NO. 7, 904–915
https://doi.org/10.1080/09658211.2019.1607388

Routledge
Taylor & Francis Group

How do ordered questions bias eyewitnesses?

Robert B. Michael [a*] and Maryanne Garry [b*]

[a]Department of Psychology, University of Louisiana Lafayette, Louisiana, USA; [b] Psychology Department, The University of Waikato, Hamilton, New Zealand

ABSTRACT
Background: Suggestive techniques can distort eyewitness memory (Wells & Loftus, 2003, Eyewitness memory for people and events. In A. M. Goldstein (Ed.), *Handbook of psychology: Forensic Psychology*, Vol. 11 (pp. 149–160). Hoboken, NY: John Wiley & Sons Inc). Recently, we found that suggestion is unnecessary: Simply reversing the arrangement of questions put to eyewitnesses changed what they believed (Michael & Garry, 2016, Ordered questions bias eyewitnesses and jurors. *Psychonomic Bulletin & Review*, 23, 601–608. doi:10.3758/s13423-015-0933-1). But why? One explanation might be that early questions set an anchor that eyewitnesses then adjust away from insufficiently. *Methods:* We tracked how eyewitness beliefs changed over the course of questioning. We then investigated the influence of people's need to engage in and enjoy effortful cognition. This factor, "Need for Cognition," (NFC) affects the degree to which people adjust (Cacioppo, Petty, & Feng Kao, 1984, The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48, 306–307. doi:10.1207/s15327752jpa4803_13; Epley & Gilovich, 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). *Results:* In our first two experiments we found results consistent with an anchoring-and-adjustment account. But in Experiments 3 and 4 we found that NFC provided only partial support for that account. *Conclusions:* Taken together, these findings have implications for understanding how people form beliefs about the accuracy of their memory.

ARTICLE HISTORY
Received 20 July 2018
Accepted 5 April 2019

KEYWORDS
Eyewitness; memory; question order; anchoring-and-adjustment

How do ordered questions bias eyewitnesses?

Eyewitnesses play a critical role in the criminal justice system. But more than 40 years of psychological research shows that suggestive techniques such as leading questions or post-identification feedback can distort eyewitness memory and confidence (Douglass & Steblay, 2006; Frenda, Nichols, & Loftus, 2011; Loftus, 2005; Loftus, Donders, Hoffman, & Schooler, 1989). When eyewitnesses are unknowingly wrong, their confidence becomes a problem because jurors find it persuasive (Cutler, Penrod, & Dexter, 1990; Douglass, Neuschatz, Imrich, & Wilkinson, 2010). That persuasiveness is especially troublesome in the context of eyewitness identifications. We now know that about 70% of wrongful convictions involve eyewitnesses who mistakenly identify the wrong perpetrator (Innocence Project, 2018). It is therefore important to investigate the factors that affect eyewitnesses' beliefs about their memory.

Recently, we discovered that eyewitnesses' beliefs about their memory can be manipulated without the use of suggestive techniques: All it took was a simple change to the order of a set of questions. We asked people to watch a simulated crime, and then we asked them questions about what they had seen. When we arranged those questions from the easiest to the most difficult (*easy-to-difficult*), people believed they had answered more questions correctly and were more confident about what they remembered, compared with their counterparts for whom we had arranged those same questions the other way around (*difficult-to-easy*; Michael & Garry, 2016).

These findings are consistent with research investigating the influence of ordered questions within an educational context (Jackson & Greene, 2014; Weinstein & Roediger, 2010, 2012). But despite an accumulating body of research, we know little about the mechanisms underlying these effects (Jackson & Greene, 2014; Michael & Garry, 2016; Michael & Weinstein, 2018; Weinstein & Roediger, 2010, 2012). More specifically, the following question remains largely unanswered: How do ordered questions influence people's beliefs?

CONTACT Robert B. Michael robert.michael@louisiana.edu Department of Psychology, University of Louisiana Lafayette, PO Box 43644, Lafayette, LA, USA

*Both authors contributed to the conception of the research and its design. Robert B. Michael collected and analyzed data. Both authors interpreted the data and contributed to the writing of the article, including drafts and critical revisions. Robert B. Michael and Maryanne Garry approved the final version of the article.

Supplemental data for this article can be accessed at doi:10.1080/09658211.2019.1607388

We know from research that at least two theories provide explanations that are unlikely. The first of these theories – the *affect* heuristic – proposes that people's feelings can quickly and automatically influence their subsequent information processing (for a review, see Slovic, Finucane, Peters, & MacGregor, 2007). Within the context of ordered questions, early easy questions should produce positive affect, while early difficult questions should produce negative affect. These affective states could then influence people's interpretation of the later questions. If this explanation were true, then we should expect people's confidence in their answers to questions to vary depending on where in a sequence those questions appear. For example, if the first few questions were difficult, then people's confidence for a subsequent easy question should be lower than if that same easy question had appeared early on. But it is not. Results from both the educational and eyewitness domains show that confidence ratings for specific questions are similar, regardless of when those questions appear in a sequence (Michael & Garry, 2016; Weinstein & Roediger, 2010, 2012).

The second theory – the *availability* heuristic – proposes instead that people rely on the information that most easily springs to mind when making decisions and evaluations (Tversky & Kahneman, 1973). Within the context of ordered questions, early questions suffer less from a buildup of interference and can be rehearsed more than later questions (Rundus, 1971). Therefore, when people are later asked to estimate their performance on the whole test, we might expect that what springs to mind are the early parts of the test. If this explanation were true, we should see that people can most easily remember the early test questions. But that is not what we see. In fact, what little research there is instead finds that people tend to remember the later questions best (Franco, 2015; Jones & Roediger, 1995). Moreover, other work shows that differences in beliefs develop while people take the test, and not solely afterward as a result of remembering the experience (Weinstein & Roediger, 2012).

The affect and availability heuristic explanations seem inadequate. Where, then, does that leave us? One promising alternative theory – the *anchoring-and-adjustment* heuristic – proposes that in situations of uncertainty, people rely on an initial piece of information as a starting point when providing estimated answers to questions (Tversky & Kahneman, 1974). This "anchor" need not be given explicitly; it can be self-generated. For example, when asked to estimate the freezing point of vodka, what springs to mind for most people who are uncertain of the true answer is the freezing point of water – an anchor that people's estimates are skewed towards (Epley & Gilovich, 2006). But why are adjustments away from these self-generated anchors typically insufficient? Research suggests that the adjustment process is effortful and stops once people reach a plausible value (Epley & Gilovich, 2006). Because a plausible range of values will often include values between the anchor and the true answer, insufficient adjustment becomes a likely outcome.

How would the anchoring-and-adjustment heuristic explain the influence of ordered questions on people's beliefs? We hypothesise that people generate their initial beliefs about test performance and memory confidence based on the ease or difficulty of early test questions. More specifically, that easy-to-difficult subjects hold initial beliefs of relatively good test performance and high confidence in their memory, while difficult-to-easy subjects hold initial beliefs of relatively poor test performance and low confidence in their memory. We further hypothesise that people adjust these beliefs as the test becomes progressively easier or more difficult, but only to the degree that the adjustment is plausible. The result? People's final beliefs about how they performed on the test, or their confidence in their memory, are skewed toward their initial anchor. In other words, despite both question arrangement groups answering the same overall set of questions, their resulting beliefs are not the same.

Some evidence from the existing research fits with the anchoring-and-adjustment theory. As noted earlier, differences in people's beliefs emerge as the test progresses, and not only at the end (Weinstein & Roediger, 2012). But we are still missing a finer-grained examination of how these biases develop. For example, one important but unanswered question is: How do these differences emerge between people who answer easy-to-difficult questions and those who answer difficult-to-easy questions? Is it all over after the very first question, or is some minimum number necessary before these groups start to diverge? In addition, we know nothing about how or why people adjust their beliefs over the course of questioning. One possibility – consistent with the anchoring-and-adjustment explanation – is that people who answer easy-to-difficult questions develop an initial impression that the test is easy and they're performing well, then adjust this belief as the test becomes progressively more difficult. People who answer difficult-to-easy questions might do exactly the opposite. The problem is that we do not know if the theory is correct and that this approach is really how people behave.

To address this problem, we first conducted two experiments (Experiments 1 and 2) in which we asked people to predict, after every test question, how many of the 30 total questions they would answer correctly. Across both experiments, we found initial support for an anchoring-and-adjustment explanation. Then, in an effort to add nuance to this theoretical account, we conducted two additional experiments (Experiments 3 and 4). Specifically, we know that people with a relatively strong desire to engage in effortful thinking tend to make more sufficient adjustments than people with a relatively weak desire (Epley & Gilovich, 2006). We hypothesised that if people's beliefs are indeed the product of an anchoring-and-adjustment heuristic, then the desire to engage in effortful thinking, or *Need For Cognition* (NFC; Cacioppo et al., 1984), should influence the magnitude of those beliefs. Across both experiments, however, we found only partial support for this explanation.

Experiment 1

If the anchoring-and-adjustment explanation is correct, then subjects who answer questions arranged from the easiest to most difficult should initially believe they are doing well, but should then adjust their estimates downward over the course of the test. Conversely, subjects who answer questions arranged from the most difficult to the easiest should show the opposite pattern, initially believing they are doing poorly, but adjusting their estimates upward over the course of the test. In addition, subjects should make insufficient adjustments to these estimates, resulting in group differences even at the end of the test (Epley & Gilovich, 2006). To investigate these predictions, we tracked how subjects' beliefs about their performance changed over the course of questioning. We repeatedly asked subjects to predict how many of the 30 total questions they would answer correctly.

Method

Subjects

In our earlier work, we populated each cell of the experimental design with a minimum of 50 subjects (Michael & Garry, 2016). In line with Cumming's (2012) recommendations, we aimed to boost precision in the current experiment with a sample size of 100 per cell (200 total). We ultimately recruited a total of 218 Amazon Mechanical Turk workers, because Mechanical Turk and Qualtrics – our experimental software – interact such that it is possible to unintentionally collect more data points than requested.

Design

We manipulated Question Order (easy-to-difficult, difficult-to-easy) between subjects.

Procedure

The experiment had four phases. First, we told subjects the study was examining learning styles. Subjects then watched a video of a tradesman who stole items from the unoccupied house in which he was working (Takarangi, Parker, & Garry, 2006).

The second phase began when the video ended. Subjects solved Sudoku number puzzles for 10 min as a filler task.

In the third phase, subjects took a surprise memory test consisting of 30 two-alternative forced choice (2AFC) questions about the video. These questions, drawn from and normed in our earlier work, were arranged sequentially from those that people answer with the lowest confidence to those that people answer with the highest confidence (difficult-to-easy) or vice versa (easy-to-difficult); these arrangements are highly related to reported question difficulty (see Michael & Garry, 2016).[1] Subjects were randomly assigned one of these test versions. For each test question, subjects used a scale from 1 (*Not at all confident*) to 5 (*Very confident*) to report their confidence they had selected the correct answer. This item-confidence measure served primarily as a manipulation check. Critically, between each test question we asked subjects, "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 fourth phase followed the test. Subjects answered two randomly ordered questions. One question asked: "The memory test about Eric the Electrician consisted of 30 questions. How many of those questions do you think you answered correctly?" Subjects responded with a number between 0 and 30. This question, in combination with those asked in the third phase, results in 30 estimates of performance for each subject, staggered across the test. The other question asked: "How confident are you about the accuracy of your memory for the video?" Subjects responded on a scale from 1 (*Not at all confident*) to 5 (*Very confident*).

Results and Discussion

We first carried out a manipulation check by examining mean confidence ratings for individual test questions. These data appear in the bottom panel of Figure 1 and show that our manipulation was successful: the difficult-to-easy subjects were increasingly confident in their test answers ($M_1 = 1.75$, $SD_1 = 1.07$; $M_{30} = 4.76$, $SD_{30} = 0.66$, $r = .52$, 95% CI [.50, .55], $p < .01$), and easy-to-difficult subjects were the opposite ($M_1 = 4.77$, $SD_1 = 0.77$; $M_{30} = 1.83$, $SD_{30} = 1.09$, $r = -.57$, 95% CI [−.59, −.54], $p < .01$). We also found that the order of questions had no meaningful effect on overall test performance, $M_{\text{difficult-to-easy}} = 20.72$, $SD_{\text{difficult-to-easy}} = 3.12$; $M_{\text{easy-to-difficult}} = 20.54$, $SD_{\text{easy-to-difficult}} = 2.52$; $M_{\text{diff}} = 0.17$, 95% CI [−0.58, 0.93], $t(216) = 0.45$, $p = .65$.

We next examined responses to the questions asked in the fourth phase, to determine the extent to which the order of test questions affected how well subjects believed they performed on the test and how confident they felt about the accuracy of their memory for the video. These data showed that difficult-to-easy subjects believed they performed more poorly on the test than easy-to-difficult subjects, $M_{\text{difficult-to-easy}} = 15.08$, $SD_{\text{difficult-to-easy}} = 5.04$; $M_{\text{easy-to-difficult}} = 18.49$, $SD_{\text{easy-to-difficult}} = 5.38$; $M_{\text{diff}} = 3.42$, 95% CI [2.02, 4.81], $t(216) = 4.83$, $p < .01$. Surprisingly, however, these differences did not extend to subjects' reported confidence in the accuracy of their memory for the video, $M_{\text{difficult-to-easy}} = 3.00$, $SD_{\text{difficult-to-easy}} = 0.91$; $M_{\text{easy-to-difficult}} = 3.00$, $SD_{\text{easy-to-difficult}} = 0.98$; $M_{\text{diff}} = 0.00$, 95% CI [−0.25, 0.25], $t(216) = 0.00$, $p = 1.00$.

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

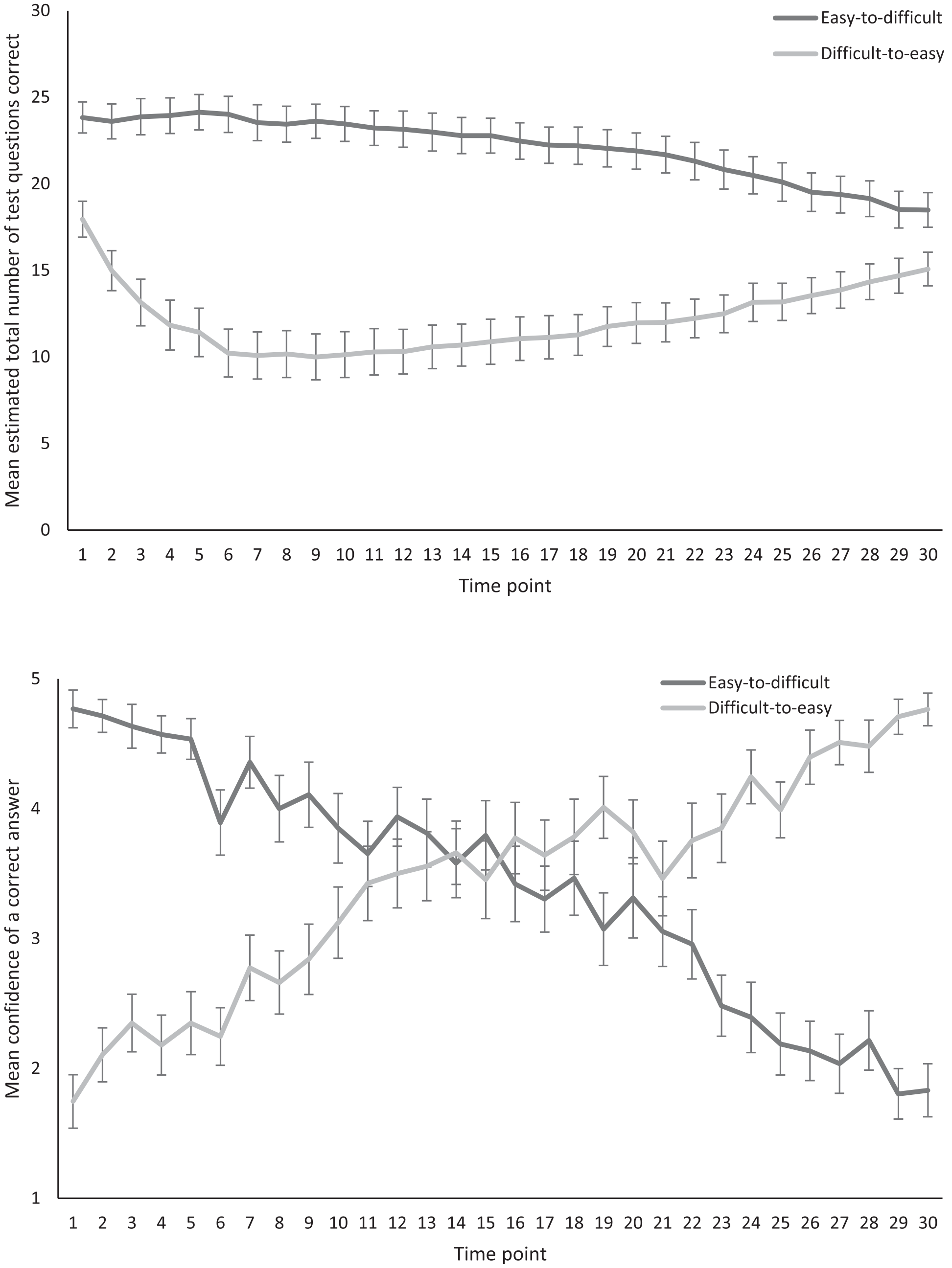

Figure 1. Top panel: Mean estimated total test scores reported after each test question as a function of question arrangement. Bottom panel: Mean confidence of a correct answer for each test question as a function of question arrangement. Error bars represent 95% confidence intervals of means. Data are from Experiment 1.

people's beliefs about their test performance. But on the other hand, we did not replicate our earlier findings with respect to memory confidence. One possible explanation is that people believe their test performance reflects the ease or difficulty of the test questions themselves, rather than reflecting the quality of their memory. That potential difference in attribution fits with research showing that people rely on anchors less as their compatibility with target judgments decreases (Chapman & Johnson, 2002). If this explanation is true, then it is plausible that the arrangement of questions influences estimates of test performance, but does little to influence judgments of memory confidence. Another explanation is that the arrangement of questions has a smaller true influence on confidence than we estimated in our earlier work; these results might therefore reflect ordinary sampling variability.

We now turn to our primary question: How do people adjust their beliefs over the course of questioning? To answer this question, we examined the mean predicted test scores people reported after each test question; these data appear in the top panel of Figure 1.

As the figure shows, the influence of a question depended on the difficulty of that question and when it appeared. After the first question, easy-to-difficult subjects made predictions that were high ($M_1 = 23.83$, $SD_1 = 4.79$) and descended over the course of the test ($M_{30} = 18.49$, $SD_{30} = 5.38$). Put another way, regression analyses showed that their adjustments fit to a straight line[2]: estimate = 25.03–0.19 * Time point; $R^2 = .08$, $F(1, 3358) = 291.26$, $p < .01$. But the difficult-to-easy subjects did not do the opposite; instead, even after the first question their predictions were lower than those of their easy-to-difficult counterparts ($M_1 = 17.95$, $SD_1 = 5.36$). These predictions continued to drop, reaching their lowest point after the ninth question ($M_9 = 10.00$, $SD_9 = 6.87$), at which point they ascended over the remainder of the test ($M_{30} = 15.08$, $SD_{30} = 5.04$). Put another way, regression analyses showed that their adjustments fit to a cubic curve: estimate = 6.28 + 0.27 * Time point + 0.02 * (Time point – 15.5)^2 - 0.002 * (Time point – 15.5)^3; $R^2 = .08$, $F(3, 3176) = 86.28$, $p < .01$.

In addition, a repeated-measures analysis of variance (ANOVA) revealed an interaction between Time point and Question Order, $F(29, 188) = 12.80$, $p < .01$. Follow-up Bonferroni-corrected comparisons (i.e., $\alpha = .05 / 30 = 0.00167$) revealed statistically significant differences between the two groups at every time point. The maximum difference in predictions occurred after the 9th test question, $M_{\text{difficult-to-easy}} = 10.00$, $SD_{\text{difficult-to-easy}} = 6.87$; $M_{\text{easy-to-difficult}} = 23.62$, $SD_{\text{easy-to-difficult}} = 5.26$; $M_{\text{diff}} = 13.62$, 95% CI [11.99, 15.25], $t(216) = 16.48$, $p < .0001$, and the minimum difference in predictions occurred after the final test question, $M_{\text{difficult-to-easy}} = 15.08$, $SD_{\text{difficult-to-easy}} = 5.04$; $M_{\text{easy-to-difficult}} = 18.49$, $SD_{\text{easy-to-difficult}} = 5.38$; $M_{\text{diff}} = 3.42$, 95% CI [2.02, 4.81], $t(216) = 4.83$, $p < .0001$.

Taken together, these findings show that people adjust their beliefs about performance during questioning. Moreover, the narrowing gap between estimates from the easy-to-difficult and difficult-to-easy subjects is consistent with an anchoring-and-adjustment explanation. To determine the extent to which these patterns would replicate and generalise to a different question format, we conducted Experiment 2.

Experiment 2

Method

Subjects

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

Design

The design was the same as Experiment 1.

Procedure

The procedure was the same as Experiment 1, except we converted each 2AFC question into a cued-recall question.

Results and Discussion

We first carried out a manipulation check by examining mean confidence ratings for individual test questions. These data appear in the bottom panel of Figure 2 and show that our manipulation worked: difficult-to-easy subjects were increasingly confident in their test answers ($M_1 = 1.99$, $SD_1 = 1.19$; $M_{30} = 4.35$, $SD_{30} = 1.29$, $r = 0.41$, 95% CI [.38, .44], $p < .01$), and easy-to-difficult subjects were the opposite ($M_1 = 4.39$, $SD_1 = 0.92$; $M_{30} = 1.72$, $SD_{30} = 1.06$, $r = -.45$, 95% CI [−.48, −.42], $p < .01$).

We next 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. In our prior work using the same scoring criteria, we found a high correlation with a blind rater's hand-scores ($r = .96$, $p < .01$; Michael & Garry, 2016). As in Experiment 1, the order of questions had no meaningful influence on overall test performance, $M_{\text{difficult-to-easy}} = 11.95$, $SD_{\text{difficult-to-easy}} = 3.71$; $M_{\text{easy-to-difficult}} = 11.13$, $SD_{\text{easy-to-difficult}} = 3.83$; $M_{\text{diff}} = 0.82$, 95% CI [−0.24, 1.88], $t(196) = 1.53$, $p = .13$.

We next examined the extent to which the order of test questions affected how well subjects believed they performed on the test and how confident they felt about the accuracy of their memory for the video. These data showed that difficult-to-easy subjects believed they performed more poorly on the test than easy-to-difficult subjects, $M_{\text{difficult-to-easy}} = 10.09$, $SD_{\text{difficult-to-easy}} = 4.37$; $M_{\text{easy-to-difficult}} = 13.55$, $SD_{\text{easy-to-difficult}} = 6.54$; $M_{\text{diff}} = 3.46$, 95% CI [1.89, 5.03], $t(196) = 4.33$, $p < .01$. But these differences did not extend to subjects' reported confidence in the accuracy of their memory for the video. We found only a trivial difference in subjects' post-test confidence ratings, $M_{\text{difficult-to-easy}} = 2.59$, $SD_{\text{difficult-to-easy}} = 0.95$; $M_{\text{easy-to-difficult}} = 2.48$, $SD_{\text{easy-to-difficult}} = 0.98$; $M_{\text{diff}} = 0.11$, 95% CI [−0.16, 0.38], $t(196) = 0.83$, $p = .41$. Taken together, these findings are consistent with Experiment 1.

Next, we examined the mean predicted test scores people reported after each test question; these data appear in the top panel of Figure 2. This pattern looks remarkably similar to the pattern in Figure 1. After just one question, the easy-to-difficult subjects made predictions that were high ($M_1 = 21.85$, $SD_1 = 5.66$) and then descended over the course of the test ($M_{30} = 13.55$, $SD_{30} = 6.54$). Put another way, regression analyses showed that their adjustments fit to a straight line: estimate = 22.79–0.28 * Time point; $R^2 = .12$, $F(1, 3088) = 405.44$, $p < .01$. But the difficult-to-easy subjects – after just one question – made predictions that were lower than their easy-to-difficult counterparts ($M_1 = 15.31$, $SD_1 = 5.66$). These predictions continued to drop, reaching their lowest point after the eleventh question ($M_{11} = 6.82$, $SD_{11} = 5.97$), and then ascended over the remainder of the test ($M_{30} = 10.09$, $SD_{30} = 4.37$). Put another way, regression analyses showed that their adjustments fit to a cubic curve: estimate = 4.70 + 0.16 * Time point + 0.02 * (Time point -

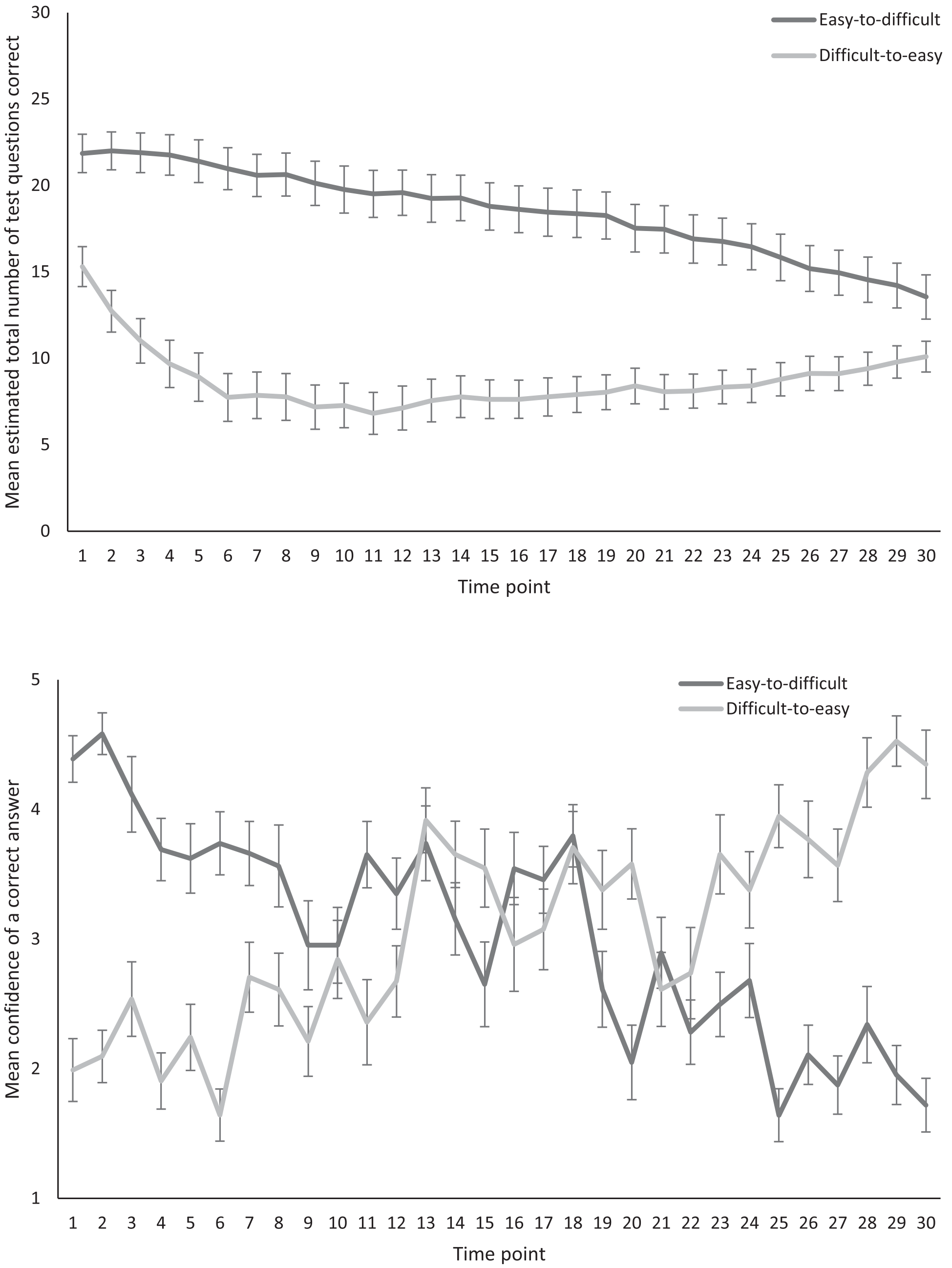

Figure 2. Top panel: Mean estimated total test scores reported after each test question as a function of question arrangement. Bottom panel: Mean confidence of a correct answer for each test question as a function of question arrangement. Error bars represent 95% confidence intervals of means. Data are from Experiment 2.

$15.5)^2$ - 0.001 * (Time point - $15.5)^3$; $R^2 = .08$, $F(3, 2846) = 81.83$, $p < .01$.

In addition, a repeated-measures ANOVA revealed an interaction between Time point and Question Order, $F(29, 168) = 11.58$, $p < .01$. Follow-up Bonferroni-corrected comparisons (i.e., $\alpha = .05 / 30 = 0.00167$) revealed statistically significant differences between the two groups at every time point. The maximum difference in predictions occurred after the 6th test question, $M_{\text{difficult-to-easy}} = 7.74$, $SD_{\text{difficult-to-easy}} = 6.78$; $M_{\text{easy-to-difficult}} = 20.97$, $SD_{\text{easy-to-difficult}} = 6.23$; $M_{\text{diff}} = 13.23$, 95% CI [11.41, 15.06], $t(196) = 14.31$, $p < .0001$, and the minimum difference in predictions occurred after the final test question, $M_{\text{difficult-to-easy}} = 10.09$, $SD_{\text{difficult-to-easy}} = 4.37$; $M_{\text{easy-to-difficult}} = 13.55$, $SD_{\text{easy-to-difficult}} = 6.54$; $M_{\text{diff}} = 3.46$, 95% CI [1.89, 5.03], $t(196) = 4.33$, $p < .0001$.

Taken together, Experiments 1 and 2 are consistent with the idea that people developed beliefs using an anchoring-and-adjustment heuristic. The results suggest that the ease or difficulty with which people experienced the first test question provided an anchoring point that constrained adjustments across the remainder of the test. The end result was a difference in what people believed about their performance – even though everyone answered the same set of questions, and their actual performance was the same.

But the way people adjusted over the course of the test was more interesting than we predicted. Specifically, we predicted that easy-to-difficult subjects would initially believe they were performing well, and would insufficiently adjust their beliefs downward over the course of the test. By contrast, we predicted that difficult-to-easy subjects would show the inverse. Instead, we found a more complex pattern, in which easy-to-difficult subjects behaved as expected, but difficult-to-easy subjects first adjusted down before slowly adjusting back up. Moreover, these patterns were consistent across Experiments 1 and 2. In summary, the two test arrangements produce markedly different experiences.

There are at least two possible explanations for these different experiences. First, given the ambiguous difficulty of upcoming test questions, easy-to-difficult subjects might be somewhat cautious in their initial optimism. Perhaps these subjects anticipated the unlikely situation that they would answer every question correctly; therefore, they rapidly hit a subjective ceiling. The only reasonable adjustment these subjects could then have made early on was downward, or none at all. Second, difficult-to-easy subjects may have been unwilling to initially anchor at a sufficiently low value – perhaps sensibly, given the ambiguous difficulty of upcoming test questions. But as these subjects accrued evidence over the early questions that they were performing poorly, they continued to adjust downward, before eventually recognising that the questions were getting easier. These two explanations are not mutually exclusive.

What other evidence might we look for to support or refute an anchoring-and-adjustment explanation? We suspected that one promising approach would be to examine people's *Need For Cognition*, because people who enjoy effortful thinking adjust more sufficiently than people who do not (Cacioppo et al., 1984; Epley & Gilovich, 2006). In Experiment 3, we set out to replicate Experiment 1 while considering the role of people's NFC. We hypothesised – in accord with the theoretical account – that people with high NFC would make larger adjustments over the course of questioning than people with low NFC. More specifically, we predicted that: (1) easy-to-difficult subjects with high NFC would initially adjust upward beyond the subjective ceiling of their low NFC counterparts, before making larger downward adjustments; (2) difficult-to-easy subjects with high NFC would initially adjust downward beyond their low NFC counterparts, before making larger upward adjustments.

Experiment 3

Method

Subjects

We recruited 400 Mechanical Turk workers.

Design

The design was the same as Experiment 1, except that we additionally split the sample into two groups based on NFC scores (High NFC, Low NFC).

Procedure

The procedure was the same as Experiment 1, except as follows.

First, we removed the confidence ratings subjects reported for their answers to each test question. We removed this manipulation check because (a) we have already established across multiple experiments – both here and in our prior work – that the manipulation is effective, and (b) it is possible the confidence judgments about individual questions were confounding the predictions subjects repeatedly provided in Experiments 1 and 2. If so, then we might expect to see a different pattern of developing beliefs when this confound is removed.

Second, we included an 18-item short form of the Need for Cognition Scale just prior to the end of the experiment. Subjects rated their agreement with each item on a scale from 1 (*Strongly disagree*) to 7 (*Strongly agree*). An example item is "The idea of relying on thought to make my way to the top appeals to me." This form of the Need for Cognition Scale demonstrates good reliability, $\theta = .90$ (here, θ represents a maximised Cronbach's alpha coefficient; Cacioppo et al., 1984).

Results and Discussion

In the analyses that follow, we found virtually identical results when treating NFC as a continuous or categorical variable, so for simplicity, we first split our sample into two groups based on the median NFC score of 4.50 (high: $M = 5.41$, $SD = 0.56$, $n = 198$; low: $M = 3.69$, $SD = 0.70$, $n = 202$; overall: $M = 4.54$, $SD = 1.07$, $n = 400$). We again found that the order of questions had no meaningful influence on overall test performance, $M_{\text{difficult-to-easy}} = 19.64$, $SD_{\text{difficult-to-easy}} = 3.59$; $M_{\text{easy-to-difficult}} = 20.24$, $SD_{\text{easy-to-difficult}} = 3.26$; $M_{\text{diff}} = 0.61$, 95% CI [–0.07, 1.28], $t(398) = 1.76$, $p = .08$. Interestingly, however, people with high NFC answered slightly more questions correctly than their low NFC counterparts, $M_{\text{high}} = 20.83$, $SD_{\text{high}} = 3.11$; $M_{\text{low}} = 19.05$, $SD_{\text{low}} = 3.54$; $M_{\text{diff}} = 1.78$, 95% CI [1.12, 2.43], $t(398) = 5.33$, $p < .01$. We found no statistically significant interaction between the order of questions and NFC, $F(1, 396) = 0.44$, $p = .51$.

Next, we examined subjects' final test estimates and post-test reports of confidence in the accuracy of their memory. For test estimates, we replicated the typical finding in which difficult-to-easy subjects believed they performed more poorly on the test than easy-to-difficult subjects, $M_{\text{difficult-to-easy}} = 14.52$, $SD_{\text{difficult-to-easy}} = 6.12$; $M_{\text{easy-to-difficult}} = 18.55$, $SD_{\text{easy-to-difficult}} = 6.08$; $M_{\text{diff}} = 4.04$, 95% CI [2.84, 5.24], $t(398) = 6.61$, $p < .01$. We found no statistically significant interaction between the order of questions and NFC, $F(1, 396) = 0.03$, $p = .85$, nor a main effect of NFC, $F(1, 396) = 1.80$, $p = .18$. These results remained virtually unchanged when we controlled for the slight difference in test accuracy between people with low and high NFC. For subjects' confidence in the accuracy of their memory, we replicated the null findings from Experiments 1 and 2, finding no statistically significant differences in subjects' post-test confidence ratings, all $ps > .08$.

We now turn to our primary question: How does the desire to engage in effortful cognition influence the adjustments eyewitnesses make to their developing beliefs about performance? To answer this question, we examined the mean predicted test scores people reported after each test question; these data appear in Figure 3.

As the figure shows, the influence of a question again depended on the difficulty of that question and when it appeared. But the figure also reveals that the influence of NFC was more complicated than we predicted. We had anticipated that people with low NFC would make smaller adjustments than their high NFC counterparts. That is, we expected that the lines or curves in Figure 3 for low NFC subjects would look "flatter" than those of the high NFC subjects. But they do not. In fact, in the difficult-to-easy conditions, low and high NFC subjects look virtually identical, adjusting similarly across the test. Put another way, regression analyses showed that both groups' adjustments fit to quadratic curves: $\text{estimate}_{\text{low}} = 12.35–0.01 * \text{Time point} + 0.01 * (\text{Time point - } 15.5)^2$; $R^2 = .02$, $F(2, 3237) = 27.77$, $p < .01$; $\text{estimate}_{\text{high}} = 11.57 + 0.002 * \text{Time point} + 0.02 * (\text{Time point - } 15.5)^2$; $R^2 = .04$, $F(2, 2967) = 54.76$, $p < .01$. The same cannot be said about the easy-to-difficult conditions. Here, NFC mattered. Specifically, people with high NFC consistently reported higher estimates across the test than their low NFC counterparts. Put another way, regression analyses showed that the low NFC group's adjustments fit to a simple line, but the high NFC group's adjustments fit to a quadratic curve: $\text{esitmate}_{\text{low}} = 22.90–0.14 * \text{Time point}$; $R^2 = .02$, $F(1, 2818) = 72.23$, $p < .01$; $\text{estimate}_{\text{high}} = 26.63–0.17 * \text{Time point - } 0.01 * (\text{Time point - } 15.5)^2$; $R^2 = .02$, $F(2, 2967) = 144.00$, $p < .01$.

In addition, a repeated-measures ANOVA revealed a three way interaction, $F(28, 369) = 1.93$, $p < .01$. We decomposed this interaction with two additional repeated-measures ANOVAs, examining the influence of NFC within each question arrangement condition. For the difficult-to-easy subjects, this analysis revealed only a main effect of Time point, $F(28, 178) = 7.27$, $p < .01$. But for the easy-to-difficult subjects, we found a statistically significant interaction between Time point and NFC, $F(28, 164) = 2.02$, $p < .01$. Follow-up Bonferroni-corrected comparisons (i.e., $\alpha = .05 / 30 = 0.00167$) revealed statistically significant differences between the easy-to-difficult low and high NFC groups after questions 9, 16, and 19 only – although we note that the mean is always numerically greater for people with high NFC. The maximum difference in predictions between high and low NFC subjects occurred after the 9th test question, $M_{\text{high}} = 24.66$, $SD_{\text{high}} = 5.31$; $M_{\text{low}} = 21.34$, $SD_{\text{low}} = 7.87$; $M_{\text{diff}} = 3.32$, 95% CI [1.42, 5.21], $t(191) = 3.45$, $p = .0007$, and the minimum difference in predictions occurred after the 28th test question, $M_{\text{high}} = 20.08$, $SD_{\text{high}} = 5.50$; $M_{\text{low}} = 18.93$, $SD_{\text{low}} = 7.09$; $M_{\text{diff}} = 1.16$, 95% CI [−0.64, 2.95], $t(191) = 1.27$, $p = .21$.

How are we to explain these results? On the one hand, the patterns are consistent with Experiments 1 and 2, in that the overall shape of developing beliefs fits with an anchoring-and-adjustment explanation. Moreover, the consistently higher predictions from people with high NFC in the easy-to-difficult condition fits with our earlier idea about people hitting a subjective ceiling – one that is slightly higher for people with high NFC, who are more capable adjusters. But on the other hand, we did not anticipate the lack of any meaningful differences according to NFC in the difficult-to-easy conditions, and it is difficult to reconcile that finding with an anchoring-and-adjustment explanation.

One possible problem with interpreting these data is that in asking people to repeatedly predict their test performance, we altered their behaviour from how it would unfold in the absence of these repeated requests. Specifically, the repeated requests for predictions might have encouraged people to more carefully monitor and think effortfully about their ongoing performance, reducing their reliance on the anchoring-and-adjustment heuristic (Simmons, LeBoeuf, & Nelson, 2010). To address this issue, we conducted Experiment 4 in an effort to examine the influence of NFC when people are asked to provide only one final estimate of their test performance. We hypothesised – in accord with the theoretical account – that people with high NFC would adjust more sufficiently than their low NFC counterparts. We therefore predicted that: (1) in the easy-to-difficult condition, people with high NFC would report a smaller final test estimate than people with low NFC; (2) in the difficult-to-easy condition, people with high NFC would report a larger final test estimate than people with low NFC.

Experiment 4

Method

Subjects

We aimed to recruit 400 Mechanical Turk workers, and ultimately recruited 408.

Design

The design was the same as Experiment 3.

Procedure

The procedure was the same as Experiment 3, except that we no longer asked people to predict their test performance after every test question. Instead – as in our earlier

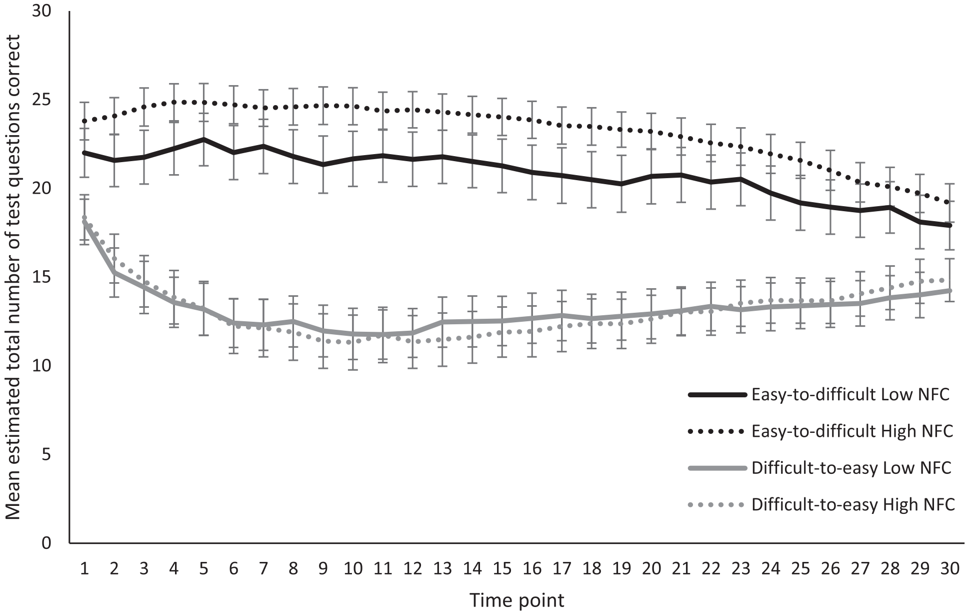

Figure 3. Mean estimated total test scores reported after each test question as a function of question arrangement and Need For Cognition (NFC). Error bars represent 95% confidence intervals of means. Data are from Experiment 3.

work – we asked people to estimate their test performance only once, at the end of the test. We know from this earlier work that reliable differences emerge in beliefs about test performance as a function of question arrangement (Michael & Garry, 2016).

Results and Discussion

Recall that, as in Experiment 3, our primary question of interest is the extent to which NFC influences the use of an anchoring-and-adjustment heuristic in producing the question arrangement effect. To answer that question, we first split our sample into two groups based on the median NFC score of 4.61 (high: $M = 5.45$, $SD = 0.57$, $n = 201$; low: $M = 3.79$, $SD = 0.71$, $n = 207$; overall: $M = 4.61$, $SD = 1.05$, $n = 408$). Consistent with Experiments 1-3, we found that the order of questions had no meaningful influence on overall test performance, $M_{\text{difficult-to-easy}} = 20.25$, $SD_{\text{difficult-to-easy}} = 2.94$; $M_{\text{easy-to-difficult}} = 20.64$, $SD_{\text{easy-to-difficult}} = 3.23$; $M_{\text{diff}} = 0.39$, 95% CI [−0.20, 1.00], $t(406) = 1.29$, $p = .20$. As in Experiment 3, however, people with high NFC answered slightly more questions correctly than their low NFC counterparts, $M_{\text{high}} = 21.01$, $SD_{\text{high}} = 2.97$; $M_{\text{low}} = 19.90$, $SD_{\text{low}} = 3.12$; $M_{\text{diff}} = 1.12$, 95% CI [0.52, 1.71], $t(406) = 3.70$, $p < .01$. We found no statistically significant interaction between the order of questions and NFC, $F(1, 404) = 0.64$, $p = .42$.

Next, we examined subjects' final test estimates and post-test reports of confidence in the accuracy of their memory. Overall, subjects behaved similarly, regardless of differences in NFC. More specifically, for test estimates, we replicated only the typical finding wherein difficult-to-easy subjects believed they performed more poorly on the test than easy-to-difficult subjects, $M_{\text{difficult-to-easy}} = 14.25$, $SD_{\text{difficult-to-easy}} = 5.54$; $M_{\text{easy-to-difficult}} = 17.61$, $SD_{\text{easy-to-difficult}} = 5.03$; $M_{\text{diff}} = 3.37$, 95% CI [2.34, 4.40], $t(406) = 6.43$, $p < .01$. For confidence in the accuracy of their memory, difficult-to-easy subjects were also less confident than easy-to-difficult subjects, $M_{\text{difficult-to-easy}} = 2.73$, $SD_{\text{difficult-to-easy}} = 0.88$; $M_{\text{easy-to-difficult}} = 3.10$, $SD_{\text{easy-to-difficult}} = 0.88$; $M_{\text{diff}} = 0.37$, 95% CI [0.20, 0.54], $t(406) = 4.25$, $p < .01$.

In other words, for subjects' final test estimates we found no statistically significant interaction between the order of questions and NFC, $F(1, 404) = 0.01$, $p = .91$, nor a main effect of NFC, $F(1, 404) = 1.49$, $p = .22$. As in Experiment 3, these results remained virtually unchanged when we controlled for the slight difference in test accuracy between people with low and high NFC. For subjects' reports of confidence in the accuracy of their memory, we found no statistically significant interaction, $F(1, 404) = 1.16$, $p = .28$, nor a main effect of NFC, $F(1, 404) = 0.02$, $p = .90$.

Overall, these results are consistent with our earlier work and show that the biasing influence of question arrangement happens both when people make repeated predictions during testing, and when they make a single post-test prediction (Michael & Garry, 2016). The patterns depicted in Figures 1–3 may therefore represent how people's beliefs develop implicitly. But importantly, we found no meaningful moderation in the size of the question arrangement effect due to NFC. This unexpected result is, as in Experiment 3, difficult to reconcile with an anchoring-and-adjustment explanation. Finally, the difference in post-test memory confidence could suggest that question arrangement only influences this judgment when people are not making explicit, repeated predictions about their performance. Of course, the alternative explanation – that the bouncing around of this small effect reflects ordinary sampling variability – is still viable.

General Discussion

Across four experiments, we aimed to determine what drives the finding that the order in which we ask eyewitnesses questions about an event can shape how well those eyewitnesses believe they answered those questions. To achieve this aim, in Experiments 1 and 2 we repeatedly asked subjects to report how well they thought they would perform on an eyewitness memory test, tracking how this belief changes over the course of questioning. We found that even with two different test formats, flipping the order of questions does not simply flip the pattern of beliefs people develop. Instead, the two orders produce markedly different experiences.

In Experiments 3 and 4, we further aimed to identify the role of Need For Cognition, an individual difference measure known to affect the extent to which people make adjustments to numerical estimates (Cacioppo et al., 1984; Epley & Gilovich, 2006). We anticipated that people high in NFC would make greater adjustments to their estimates than their low NFC counterparts, both when people repeatedly provided estimates over the course of the test (Experiment 3) and when people provided only one estimate after the test (Experiment 4). Such findings, if present, would fit with the idea that people rely on an anchoring-and-adjustment heuristic when forming beliefs about their performance. But instead, both experiments produced results that are difficult to reconcile with an anchoring-and-adjustment explanation.

In Experiment 3, people with high NFC adjusted differently compared to people with low NFC only when the test was arranged from the easiest to most difficult question. And, in Experiment 4, we found no evidence that NFC affected people's single, post-test estimates of performance – estimates that were now free of the potential influence of repeated test score predictions. Across both experiments, we had anticipated instead that people with high NFC would adjust more than their low NFC counterparts, reducing the difference in final test estimates between the question arrangement conditions (see, e.g., Epley & Gilovich, 2006). Overall, the results from these two experiments suggest that effortful thinking may not protect people from the influence of ordered questions. But we state this suggestion only tentatively, because an alternative explanation is that there are, in fact, small differences in adjustment due to NFC that require greater precision to detect.

Considered as a package, a critic might wonder if these four experiments have value, given that they do not support firm conclusions about the mechanisms responsible for the influence of ordered questions. On the contrary, we think they do. In particular, the patterns of developing beliefs in Experiments 1 and 2 raise an important question: Why do these beliefs develop in a qualitatively different way, when everyone ultimately sees the same set of questions? Put another way, why is it that difficult questions dramatically change people's beliefs about test performance when encountered first, but those exact same questions produce almost no change in beliefs about test performance when encountered last? Our results also add to the small but growing body of literature investigating explanations for the influence of question arrangement. The available evidence to date suggests that a number of other explanations are unlikely, including the possibility that people remember the first test questions best (Franco, 2015); their affect changes across the test (Weinstein & Roediger, 2010, 2012); and their attention declines across the test (Michael & Garry, 2016).

In line with our prior work, we consistently found that eyewitnesses who first answered easy questions believed they answered more questions correctly than eyewitnesses who first answered difficult questions. That finding replicated across all four experiments, and fits with research investigating the influence of question arrangement in an educational paradigm (Jackson & Greene, 2014; Weinstein & Roediger, 2010, 2012). But in contrast to our previous work, we found in three of the four experiments that eyewitnesses who first answered easy questions were just as confident in the accuracy of their memory as eyewitnesses who first answered difficult questions. This finding is at odds with our previous work (Michael & Garry, 2016).

How are we to explain this disconnection between judgments of test performance and memory confidence? We suspect that it may be due to different attributions people make across these two judgments. More specifically, test performance is a consequence of both the quality of memory and the nature of the test questions. If initially asked difficult questions that virtually no one could answer correctly, people might develop an impression that their test performance is poor – but not because of a shaky memory. Instead, that poor performance can be attributed to some unfairly difficult questions. A similar difference in attribution could arise if initially asked easy questions that virtually everyone could answer correctly. One way to test this speculative explanation would be to ask people to explain their test performance and memory confidence judgments. If our hypothesised explanation is correct, we would expect that people attribute their test performance to the ease or difficulty of the test, rather than the quality of their memory. As we acknowledged earlier, however, an alternative explanation – one that is simpler, but perhaps less interesting – is that the true size of this effect is smaller than we estimated in our prior work (Michael & Garry, 2016).

Our research adds nuance to the literature because it shows that a seemingly trivial and non-suggestive manipulation can influence eyewitness metacognition (Wells &

Loftus, 2003). Moreover, the results have implications for the mechanisms responsible for the effects that occur when people answer questions arranged in certain orders (Weinstein & Roediger, 2012). As a whole, the theory of effortful adjustment seems an inadequate explanation for our results (Epley & Gilovich, 2006). But very recently, a new paper appeared providing empirical support for an alternative theory that may prove fruitful in future investigations. This theory proposes that anchoring effects are the result of an aversion to extreme adjustments (Lewis, Gaertig, & Simmons, 2018).

It is also worth noting a methodological difference between the work presented here and other investigations of the anchoring phenomenon. In our paradigm, people provide an estimate of their performance after a series of questions. In other work, people typically provide an estimate in response to a single question (Epley & Gilovich, 2006; Tversky & Kahneman, 1974). Perhaps the serial nature of our paradigm reduces the reliance on an anchoring-and-adjustment heuristic because it provides people with multiple retrieval cues that can lead to recall of event details, reducing the necessity of relying on other information – like how easy or difficult it feels to answer questions (Greifeneder, Bless, & Pham, 2011).

What recommendations could we make – if any – for applied contexts, such as eyewitness interviewing? We know that best practice interviewing techniques often recommend an initial rapport-building phase that could be construed as a set of easy questions before the "real," more difficult questioning begins (Collins, Lincoln, & Frank, 2002). So it is plausible that question arrangement may have some influence when interviewing eyewitnesses. But we state this possibility cautiously, because a rapport-building technique differs in a number of ways from the serially ordered question manipulation we used, and thus might not meaningfully bias eyewitnesses at all. Furthermore, we also know that best practice techniques typically recommend that the types of questions we asked should be used only toward the end of interviewing, after extensive free report procedures (Paulo, Albuquerque, & Bull, 2013). We therefore also don't know, yet, whether question arrangement would make any appreciable difference in people's beliefs if those people have already had an opportunity to engage in extensive recall. Finally, it is difficult to see how forensic interviewers could possibly know *a priori* the difficulty of their questions. Perhaps the only reasonable conclusion to draw, then, is that we may need to think more carefully about how the experience of difficulty changes for eyewitnesses over the course of questioning, because that experience can plausibly distort what people believe.

Notes

1. In our prior work we established that reported confidence closely aligns with reported difficulty, suggesting that confidence is a good proxy for subjective difficulty ($r = -.82$, 95% CI [$-.66$, $-.91$]; Michael & Garry, 2016).
2. We present these line and curve data because they are intuitively understandable. But the careful reader will note they are statistically problematic due to autocorrelation. We therefore ran additional regression analyses that included a lag variable of the estimates, and in each case this approach improved model fit and successfully removed autocorrelation. These data can be found in Table 1 of the Supplementary Materials.

Disclosure of interest

The authors report no conflict of interest.

Data availability statement

The data for all four experiments reported in this manuscript are available from the Open Science Framework at the following address: https://osf.io/8hkmj/

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Robert B. Michael http://orcid.org/0000-0001-5275-7636

References

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307. doi:10.1207/s15327752jpa4803_13

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 120–138). Cambridge, UK: Cambridge University Press.

Collins, R., Lincoln, R., & Frank, M. G. (2002). The effect of rapport in forensic interviewing. *Psychiatry, Psychology and Law*, *9*, 69–78. doi:10.1375/pplt.2002.9.1.69

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

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

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

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.

Frenda, S. J., Nichols, R. M., & Loftus, E. F. (2011). Current issues and advances in misinformation research. *Current Directions in Psychological Science*, *20*, 20–23. doi:10.1177/0963721410396620

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*(2), 107–141. doi:10.1177/1088868310367640

Innocence Project. (2018). The Causes of Wrongful Conviction. Retrieved from https://www.innocenceproject.org/causes/eyewitness-misidentification/.

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

Jones, T. C., & Roediger, H. L. (1995). The experiential basis of serial position effects. *European Journal of Cognitive Psychology*, *7*, 65–80. doi:10.1080/09541449508520158

Lewis, J., Gaertig, C., & Simmons, J. P. (2018). Extremeness aversion is a cause of anchoring. *Psychological Science*, *30*, 1–15. doi:10.1177/0956797618799305

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., 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

Michael, R. B., & Garry, M. (2016). Ordered questions bias eyewitnesses and jurors. *Psychonomic Bulletin & Review*, *23*, 601–608. doi:10.3758/s13423-015-0933-1

Michael, R. B., & Weinstein, Y. (2018). The influence of ordered question difficulty: A meta-analysis of two paradigms. *Manuscript in preparation*.

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

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*, 63–77. doi:10.1037/h0031185

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology*, *99*, 917–932. doi:10.1037/a002140

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

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

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

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

Weinstein, Y., & Roediger, 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

Wells, G. L., & Loftus, E. F. (2003). Eyewitness memory for people and events. In A. M. Goldstein (Ed.), *Handbook of psychology: Forensic Psychology, Vol. 11* (pp. 149–160). Hoboken, NY: John Wiley & Sons Inc.