Empirical Research Paper

Conflating Temporal Advancement and Epistemic Advancement: The Progression Bias in Judgment and Decision Making

Personality and Social Psychology Bulletin
1–17

DOI: 10.1177/0146167219838542
journals.sagepub.com/home/pspb
SAGE

Haotian Zhou[1], Xilin Li[2], and Jessica Sim[3]

Abstract

When seeking out the truth about a certain aspect of the world, people frequently conduct several inquiries successively over a time span. Later inquiries usually improve upon earlier ones; thus, it is typically rational to expect the finding of a later inquiry to be closer to the truth than that of an earlier one. However, when no meaningful differences exist between earlier and later inquiries, later findings should not be considered epistemically superior. However, in these cases, people continue to regard findings from later inquiries as closer to the truth than earlier ones. In 10 experiments, when later inquiries conflicted with—but did not epistemically improve upon—earlier ones, participants' global judgments about the truth aligned more with later findings than earlier ones, an effect referred to as progression bias. The liability to progression bias may have severe ramifications for the well-being of the society and its members.

Keywords

progression bias, schema, overgeneralization, judgment and decision making

Received December 20, 2017; revision accepted January 11, 2019

Introduction

A psychologist recently related an incident involving a friend whose frustration many readers can empathize with (Corker, 2015):

> A friend posted a question to a group of research colleagues recently: "Three weeks ago, I ran a 100 person two-condition study on MTurk. Result: $t = 2.95$, $p = .004$. Today I ran another 100 person two-condition study on MTurk, using the identical measure. No differences in what came before that measure. Result? $t = 0.13$, $p = .89$." The friend was exasperated and didn't know what to do.

Would the friend have been *less* frustrated if the results of the two studies had been reversed? We suspect that the answer is probably yes. If Study 2 actually improved over Study 1 (e.g., better operationalization, higher power), the friend might be justified in feeling more confident about the project when Study 2 rather than Study 1 yielded the supportive data. However, given that Study 2 was apparently an exact replication of Study 1, allowing it to have more influence on one's global assessment than by an identical study conducted earlier seems to lack rational ground.

Like the psychologist's friend who needed to assess the plausibility of her hypothesis based on mixed results, people from all walks of life often have to form global judgments that integrate results of multiple inquiries conducted successively over time. For example, a dieter may step on a scale several times to see whether her diet regimen is working, a doctor may order multiple tests to verify whether a tumor is cancerous, a consumer may peruse multiple reviews to decide whether a new movie is worth checking out, and a detective may interrogate a suspect multiple times to gauge whether he is truly innocent. Whether the finding of a later inquiry deserves more consideration than that of an earlier one depends on whether any substantive changes occurred between the two inquiries. When later inquiries do not improve upon earlier ones, faulty judgments can occur if the later inquiries are nonetheless treated as epistemically superior to those of earlier ones, a mistake we term *progression bias*. Ascertaining whether and why people are prone to

[1]ShanghaiTech University, China
[2]The University of Chicago, IL, USA
[3]Elmhurst College, IL, USA

Corresponding Author:
Haotian Zhou, ShanghaiTech University, 199 Huanke Road, Shanghai, China, 201210.
Email: zhouhaotian@gmail.com

progression bias could have broad social and economic implications.

Is there any reason to think that people are liable to be *unduly* influenced by the more recent inquiries when forming global judgments? The answer we think is affirmative. Cognitive errors and biases often result from people unreflectively applying methods, rules of thumb, or mental schemas that are appropriate in one context to another ostensibly similar but fundamentally incompatible context (e.g., Gigerenzer & Gaissmaier, 2011; Kernis, Brockner, & Frankel, 1989; Tomasello & Herron, 1988; Zebrowitz & Rhodes, 2004). The conjunction fallacy (better known as the Linda problem) vividly illustrates how overgeneralization of an otherwise adaptive mental shortcut, that is, representative heuristic, to an inappropriate situation can go awry (Tversky & Kahneman, 1983). We argue that progression bias can result from the overgeneralization of what we term *sequential-inquiry schema*, a mental schema for the type of common human activities alluded to earlier, that is, seeking truth through several inquiries conducted sequentially over a time span.

Mental schemas are often automatically activated based on superficial cues. For instance, seeing bats, a mammalian species can activate the schema for birds. Once activated, they can affect what people attend to, how people interpret things, and how people act (Fiske & Linville, 1980; Rumelhart, 1978). More importantly, a mental schema can filter out or distort the characteristics of an entity that violates the schema's specifications (Kleider, Pezdek, Goldinger, & Kirk, 2008). In Brewer and Treyens's (1981) seminal study, participants were asked to wait in a room identified as a professor's study. When later probed about the room's contents, participants recalled seeing books even though none were present. Participants' schema for a professor's study—a room full of books—led them to mentally "normalize" the inconsistent room to fit the schema.

Schemas typically originate from people's firsthand experience as well as shared cultural knowledge. As a result, we think that a key specification of sequential-inquiry schema is that later inquiries are better equipped to reveal truths than earlier ones. When people make multiple attempts to investigate a truth, they usually tweak each successive inquiry to improve the odds of uncovering the truth. The epistemology of science familiar to most laypeople also implicates the progressive elimination of falsehoods, leaving behind a narrative of monotonic progression, with each successive generation of scientists inching closer to the nature of reality (Kuhn, 2012)—an idea capsulated by the cultural metaphor of dwarfs standing on the shoulders of giants. Not unlike other schemas, sequential-inquiry schema can be activated by aberrant cases where later inquires do not improve upon earlier ones. Yet, once activated, it sets up the expectation that each successive inquiry should be epistemically superior to its predecessor, which could obscure the fact that the changes from an earlier inquiry to a later one are essentially random variations.

Aside from the overgeneralization of sequential-inquiry schema, several alternative accounts could also explain the underserved "allure" of later inquires. A situation that involves someone repeatedly probing a truth might activate the lay theory that practice makes perfect, which we refer to as *practice-effect heuristic*. Activation of this heuristic could also set up the expectation that later inquiries should be epistemically superior to earlier ones because they are supposedly executed more competently. As a result, it would be difficult, in many situations, to determine whether a partiality toward later inquiries should be attributed to sequential-inquiry schema or practice-effect heuristic. In fact, both could be working in tandem to tip the scale unfairly toward later inquiries. Conceivably, practice-effect heuristic can also be misapplied to situations where skill has already plateaued, as in the case of experts. As a result, simply because progression bias could be accounted for by practice-effect heuristic does not make it rational. Now, a case can be made that when a different agent administers each successive inquiry, practice-effect heuristic would probably be either dormant or overshadowed by sequential-inquiry schema. Thus, in a sense, the scope of practice-makes-perfect heuristic is more restrictive than sequential-inquiry schema.

A second plausible account of progression bias involves *novelty*, which is defined as the discrepancy between what is known and what is discovered and has been found to arouse interest and behavior (Mather, 2013). Infant studies show that attention wanes in response to repeated stimulation but recovers to novel stimuli (Fantz, 1964; Thompson & Spencer, 1966). Furthermore, research in cognition shows that people are more eager to learn from, and are typically rewarded for responding to, novel events (Kumaran & Maguire, 2007; Ranganath & Rainer, 2003). Given novelty's definition, the findings of later inquiries typically are more novel than earlier inquiries and therefore would be more privileged by attention, receiving more consideration as a result (Sternberg, Roediger, & Halpern, 2007). Yet, from the perspective of a person making global judgments, the novelty level of an inquiry is determined by when she is informed of its findings rather than when it is conducted. For instance, people today probably find the medieval star-alignment theory of flu more novel than the modern virus theory. Thus, in cases where later inquiries are learned before earlier inquiries, the novelty account would predict an anti-progression bias.

Another possible alternative explanation traces progression bias to a person's general outlook on life, that is, *dispositional optimism*. To the extent that people hold the conviction that things typically get better (Peterson, 2000; Scheier & Carver, 1985), they should also expect later inquiries to be better than earlier one. A straightforward corollary of the dispositional-optimism account is pessimistic individuals should be more liable to anti-progression bias, unduly trusting earlier rather than later inquiries. Given that our society is dominated by pessimists (Roser & Nagdy, 2018)—only 6% of Americans in a recent survey believed the world

was improving (Dahlgreen, 2016), it would be practically impossible to observe progression bias if it were indeed exclusively rooted in dispositional optimism. Thus, empirical evidence of progression bias would be the best counter-evidence to this alternative account.

In sum, we hypothesize that due to the activation of sequential-inquiry schema, people will expect later inquiries to be more valid than earlier ones, even when the temporal order of successive inquiries is not predictive of their relative strength. As a result, when earlier and later findings contradict, we predicted that global judgment about a truth would be disproportionately influenced by later, rather than earlier, findings. Across 10 experiments, we tested our central prediction, assessed the relative merit of our hypothesis versus other viable alternative accounts, and illustrated the potential implications of progression bias for important judgments in real life.

Based on a pilot study, we anticipated a medium effect size between 0.5 and 0.6 as measured by Cohen's *d* for our manipulation of temporal order. Along with a desired statistical power of 0.80, we determined ex ante that the sample size for each cell in all the experiments would be around 50 or more. Readers can refer to the Stimulus Materials file for the exact wording of all instructions, manipulations, and questions.

Experiment 1: Surveillance Camera

The goal of this experiment was to test our central prediction that people would still exhibit progression bias when later inquiries did not meaningfully improve upon earlier ones.

Participants

One hundred eighteen participants (65 males; M_{age} = 33) recruited from the crowdsourcing platform Amazon Mechanical Turk (i.e., MTurk) completed the experiment in exchange for monetary compensation. MTurk proves to be a quality data source for noncomplicated psychological experiments (Buhrmester, Kwang, & Gosling, 2011; Hauser & Schwarz, 2016).

Materials and Procedure

Participants were randomly assigned to either the *later-positive* (n = 60) or *later-negative* condition (n = 58) condition. Later-positive participants read the following vignette:

> One night, a police officer fired at and injured an armed man in a dimly lit parking lot. Before the shooting, the officer had ordered the man to put down to the ground the handgun he was wielding over his head. The officer claimed that he acted in self-defense when he saw the man was trying to shoot him. However, the hospitalized man claimed that he merely did what the officer commanded with no malintent.
>
> The committee investigating the incident sought help from the handful of buildings in the vicinity. Specifically, the committee asked each building's management to find out if any of its outdoor surveillance cameras had had unobstructed views of the shooting scene.
>
> A building east of the parking lot first identified such a camera and turned its footage over to the committee for examination. From this camera's angle, the victim appeared to intend to surrender his weapon when lowering it down.
>
> At a later time, a nearby building also reported a find. But the footage of this camera, which was at an angle to the early one, revealed a different story upon examination. From this second camera's angle, the victim appeared to intend to aim his weapon at the officer's chest when lowering it down.

Later-negative participants read the same vignette except that the apparent intentions of the victim shown on the two footages were reversed. All participants were asked to rate the extent to which they thought that the victim actually intended to harm the officer on a 1-to-100 sliding scale where 1 = *not at all intended* and 100 = *definitely intended*.

In this vignette, there was no obvious a priori reason to think that one footage was of a higher quality than another. After all, the investigative committee was not at all involved in the creation of either footage. Furthermore, as the committee did not even know that the second footage existed until it was brought to them, participants should not have assumed that the analysis of the second footage was prompted by concerns about the inadequacy of either the first footage or its analysis.

Results and Discussion

The later-positive participants were more convinced that the victim intended to harm the officer (M = 58.72, SD = 21.48, 95% confidence interval [CI] = [53.17, 64.26]) than the later-negative participants (M = 42.86, SD =26.51, 95% CI = [35.89, 49.83]), $t(116) = 3.58$, $p < .001$, Cohen's d = 0.66. Consistent with our prediction, participants seemed to be unduly influenced by the result of the second inquiry although there were no meaningful differences between the first and second inquiries. On the contrary, the data contradict the dispositional-optimism account of progression bias, which would predict anti-progression bias in an American sample.

In this vignette, the two inquiries were conducted by the same agent (i.e., the investigative committee), which raises at least two concerns that might affect how the findings should be interpreted.

First, practice-effect heuristic might have been activated, leading participants to the surmise that the committee probably became more adept at analyzing footage the second time around. In other words, both sequential-inquiry schema and practice-effect heuristic could account for the progression

bias observed in this experiment. However, the fact that the undue influence of the later inquires could be accounted for by practice-effect heuristic does not mean the progression bias observed in this experimental was rational. On one hand, a committee charged with investigating such a serious matter supposedly consisted of established forensic experts. Experts are people whose skills do not typically benefit from additional practices. On the other hand, even if the committee was unfamiliar with analyzing intent from surveillance footage initially, it could have just gone over the first footage multiple times to get acquainted, especially when it had no idea whether it would have another footage in the first place. In sum, the conjecture that followed from practice-effect heuristic, that is, the committee probably became more adept at analyzing footage the second time around, lacked rational support in this vignette.

The second concern is that the committee was aware of the finding of the first inquiry at the outset of the second one. Participants might have reasoned that given human beings' predilection for consistency, the evidence in the second footage must have been particularly strong for the committee to draw an opposite conclusion. As a result, it was not at all irrational to give more consideration to a conclusion supported by stronger evidence. However, this interpretation presumes a combination of a deep understanding of human psychology and sophisticated reasoning, probably a tall order for a typical layperson.

Nonetheless, a better way to address these two concerns is to test our prediction with a vignette where successive inquiries were conducted by different agents, who did not communicate with each other, an approach we adopted in the next experiment.

Experiment 2: Sniffer Dogs

Participants

One hundred twenty-three MTurk workers participated in the experiment for monetary compensation. The demographic questions, including gender and age, were left out of the survey by accident.

Materials and Procedure

Participants were randomly assigned to either the *later-positive* ($n = 63$) or *later-negative* condition ($n = 60$). Later-positive participants read the following vignette:

> One day morning, a suspicious aluminum suitcase was found in the waiting area of an airport. The local FBI office requested the two sniffer-dog training schools nearby to each send their best sniffer dogs to check if the suitcase contained explosive materials.
>
> At around noon, the first sniffer dog arrived at the scene. After sniffing at the suitcase for a few moments, the dog leisurely strolled away from the suitcase without a whimper. Half an hour later, the second sniffer dog arrived. After sniffing at the suitcase for a few moments, the dog, unlike the first one, started barking at the suitcase furiously.

Later-negative participants read the same vignette except that the reactions of the two sniffer dogs were reversed. All participants were asked to rate the extent to which they believed that there were explosives concealed in the suitcase on a 1-to-100 sliding scale, where 1 = *not at all likely* and 100 = *extremely likely*.

Similar to the vignette in Experiment 1, in this vignette the decision to probe the suitcase with a second sniffer dog was contingent on neither the characteristics nor the behavior of the first dog, eliminating the possibility that the existence of the second inquiry was taken as a sign of the deficiencies of the first inquiry. In addition, the two inquiries were conducted by two separate agents (i.e., two sniffer dogs), which lacked neither the opportunity nor the ability to exchange knowledge about the suitcase.

Results and Discussion

We again found evidence of progression bias. The later-positive participants were more convinced that the suitcase contained explosives ($M = 58.50$, $SD = 18.77$, 95% CI = [53.76, 63.22]) than their later-negative counterparts ($M = 50.91$, $SD = 16.47$, 95% CI = [46.66, 55.17]), $t(121) = 2.38$, $p = .019$, Cohen's $d = 0.43$. More importantly, by having two different agents, who could not have communicated with each other, responsible for the two inquiries, we ruled out practice-effect heuristic as a likely alternative for the progression bias documented in this experiment, thereby lending indirect support to our hypothesis that progression bias is rooted in the overgeneralization of sequential-inquiry schema.

Experiment 3: Chinese Antique

In the previous two experiments, we had assumed that the differences between the successive inquiries excluding temporal order (e.g., the two cameras were at an angle to each other in Experiment 1, and different schools trained the two dogs in Experiment 2) were not indicative of quality disparity between them. The validity of this assumption is critical for adjudicating the irrationality of progression bias. Therefore, in the current experiment, we included a pretest to evaluate whether this assumption held for the inquiries involved.

Participants

One hundred forty-two MTurk workers (84 males, $M_{age} = 34$) participated in the main experiment for monetary compensation.

Materials and Procedure

Participants of the main experiment were randomly assigned to either the *later-positive* ($n = 70$) or *later-negative* condition ($n = 72$). Later-positive participants read the following vignette:

> A celebrity who collects Chinese art pieces is considering purchasing a ceramic vase from an art dealer. The dealer claims that the vase is a genuine Chinese antique made some 400 years ago.
>
> Seeking to verify the seller's claim before consummating the deal, the celebrity consults his collector friends about reliable antique-authentication services and narrows the choice down to two highly regarded agencies, each with its own distinct authentication method. Unable to judge the relative merit of the two approaches, the celebrity decides to have the two agencies inspect the ceramic vase in turn.
>
> The vase first stays a week at the agency whose authentication method capitalizes on the fact that water in the atmosphere tends to be absorbed by and chemically bonded with ceramics as ions. The agency's analyses indicate that the vase is actually a forgery made not long ago.
>
> The celebrity then arranges to have the vase sent to the other agency, where it stays another week. This second agency's method capitalizes on the fact that environmental radioactivity tends to cause electric charges to accumulate in ceramics. Contrary to the first agency, the second agency's analyses indicate that the vase is indeed a genuine antique crafted centuries ago.

Later-negative participants read the same vignette except that the findings of the two agencies were reversed. All participants were asked to rate the extent to which they thought this vase was a genuine Chinese antique on a 1-to-100 sliding scale, where 1 = *not at all genuine* and 100 = *almost certainly genuine*.

Note that the two authentication principles described in the vignette are simple redescriptions of the underlying mechanisms of two actual ceramic-dating techniques. By having the two agencies specialize in two distinct, scientific-sounding methods, we intended to make the story in the vignette more realistic and less contrived. However, this feature entailed a possible drawback, that is, we could not safely assume that participants would see these two methods as epistemic equals. Therefore, we conducted a pretest on an independent group of 66 MTurk workers (38 males, $M_{age} = 35$), who read the following vignette, a redacted version of the vignette from the main experiment:

> A celebrity who collects Chinese art pieces is interested in purchasing a ceramic vase from an art dealer. The dealer claims that the vase is a genuine Chinese antique made some 400 years ago.
>
> Seeking to verify the seller's claim before consummating the deal, the celebrity consults his collector friends about reliable antique-authentication services and narrows the choice down to two highly regarded agencies, each with its own distinct authentication method.
>
> While the method of one agency (A) capitalizes on the fact that water in the atmosphere tends to be absorbed by and chemically bonded with ceramics as ions, the method of the other agency (B) capitalizes on the fact that environmental radioactivity tends to cause electric charges to accumulate in ceramics.
>
> Based on your intuitive understanding of these two methods as summarized above, which method do you think is the superior one?
>
> The pretest participants answered this question on a −3 to +3 bipolar, where −3 = *definitely Method (A)* and +3 = *definitely Method (B)*.

Results and Discussion

Pretest participants did not seem to hold any a priori assumption that one authentication method was superior to the other ($M = -0.09$, $SD = 1.90$, 95% CI = [−0.56, 0.38]), $t(65) = -0.39$, $p = .70$. Note that the vignette in the main experiment made it clear that the celebrity himself considered the two methods comparable. Therefore, the two inquiries in the main experiment should be considered equivalent in terms of reliability.

Nonetheless, progression bias surfaced again in the main experiment. The later-positive participants were more convinced the genuineness of the ceramic vase ($M = 55.46$, $SD = 25.86$, 95% CI = [49.99, 60.92]) than their later-negative counterparts ($M = 42.33$, $SD = 16.47$, 95% CI = [36.26, 48.41]), $t(140) = 3.20$, $p = .002$, Cohen's $d = 0.54$.

Experiment 4: White Blood Cell Count

In each of the first three experiments, how exactly an inquiry was conducted was shrouded in mystery. It is quite unlikely an average MTurk worker would know precisely how forensic experts analyze footage from a surveillance camera, how sniffer dogs detect explosives, or how archeologists date ceramic antiques. Hence, it is possible that progression bias is confined to situations where people do not know the mechanics of the inquiries in question. To evaluate this postulate, the vignette in the current experiment lays out the steps involved in the inquiries.

Participants

One hundred nineteen MTurk workers (70 males, $M_{age} = 35$) participated in the main experiment for monetary compensation.

Materials and Procedure

Participants were randomly assigned to either the *later-positive* ($n = 61$) or *later-negative* condition ($n = 58$). All participants read the following vignette:

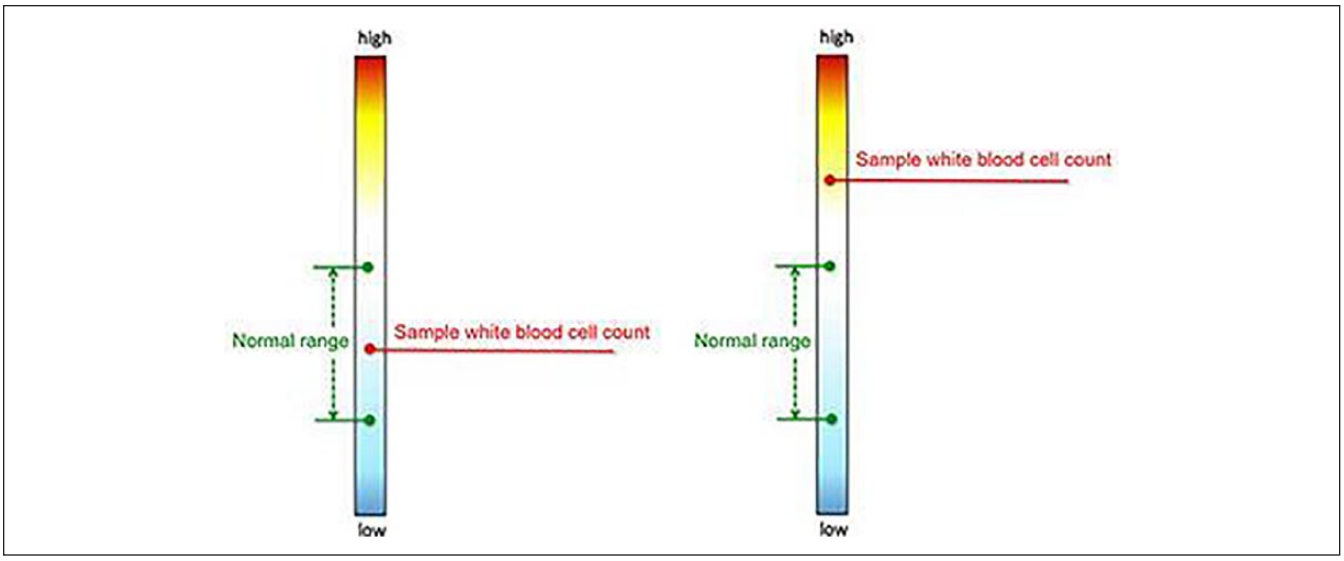

Figure 1. The results of the morning and afternoon tests shown to the later-positive participants.
Note. The positions of the two images were reversed for the later-negative participants.

> Mr. Jackson's company recommends all the employees to undergo the routine annual physical checkup. Mr. Jackson decides to comply with the company's recommendation. One item in the checkup is the white blood cell count because above-normal white blood cell count often indicates the presence of inflammation in the body. The doctor orders two blood tests for the next day, one in the morning and one in the afternoon just to be sure.
>
> The next morning, the nurse on duty draws 10 cc blood sample from Mr. Jackson and counts the white blood cells in the sample under the microscope. The white blood cell count in this first blood sample is shown in the image on the LEFT. When Mr. Jackson comes back for the second test in the afternoon, the nurse on duty has changed. Nonetheless, Mr. Jackson undergoes the same procedure as in the morning. The white blood cell count in this second blood sample is shown in the image on the RIGHT.

Unlike the previous experiments where the findings of the inquiries were verbally described, the results of the two blood tests were presented graphically. Figure 1 reproduces the graph shown to the later-positive participants. The positions of the two images in the graph were reversed for the later-negative participants. By graphically displaying the results of the two tests side by side, we sought to make the processing of the vignette less cognitively demanding, thereby providing a more stringent test of progression bias. All participants rated the extent to which they believed that Mr. Jackson had inflammation on a 1-to-100 sliding scale, where 1 = *definitely had no inflammation* and 100 = *definitely had inflammation*.

Similar to the Sniffer Dogs study, in the current experiment, two agents (i.e., nurses on duty), who were unlikely to have exchanged their knowledge about the patient, were individually responsible for the two inquiries. In addition, the rationale for ordering a second test was completely decoupled from the circumstances of the first test. More critically, the precise steps involved in the inquiries were described to participants in accessible terms—namely, drawing a small sample of blood, counting a particular type of cells in it, and comparing the count with a predetermined criterion.

Astute readers probably have noticed that the two inquiries did differ on a dimension that could potentially be indicative of test accuracy. Specifically, the first blood test was conducted in the morning and the second in the afternoon. To gauge whether this difference could constitute a rational ground for overvaluing the result of the second blood test, we ran a pretest on an independent group of 68 MTurk workers (37 males, M_{age} = 36), who read the following vignette, a redacted version of the vignette from the main experiment:

> Mr. Jackson's company recommends all the employees to undergo the routine annual physical checkup. Mr. Jackson decides to comply with the company's recommendation. One item in the checkup is the white blood cell count because above-normal white blood cell count often indicates the presence of inflammation in the body.
>
> To test whether a person's white blood cell count is above the normal level, a nurse would draw 10 cc blood from that person and count the white blood cells in the sample under a microscope.
>
> Suppose that Mr. Jackson can choose to have this blood test either in the morning or the afternoon. According to your intuition, is morning or afternoon a better time to have this blood test?

Pretest participant responded to the question on a −3 to +3 bipolar scale, where −3 = *definitely in the morning* and +3 = *definitely in the afternoon*.

Results and Discussion

Pretest participants had the a priori assumption that the morning test tends to be more accurate ($M = -1.09$, $SD = 1.92$, 95% CI = [–1.62, −0.56]), $t(67) = -4.11$, $p < .001$, which would work actually against progression bias, and therefore could not account for progression.

Despite the a priori assumption that the second inquiry was conducted during the less ideal part of the day (i.e., afternoon), participants' global judgment was still consistent with progression bias. Specifically, the later-positive participants were more convinced that Mr. Jackson had inflammation ($M = 63.91$, $SD = 19.06$, 95% CI = [58.90, 68.93]) than the later-negative participants ($M = 50.64$, $SD = 25.64$, 95% CI = [44.07, 57.21]), $t(117) = 3.19$, $p < .01$, Cohen's $d = 0.59$.

Experiment 5: Home Pregnancy Test

The four experiments reported so far not only consistently demonstrate people's liability to progression bias but also lend credence to our hypothesis that this bias is due to the overgeneralization of sequential-inquiry schema to speciously compatible situations by ruling out both the dispositional-optimism account and the practice-effect heuristic account. However, the novelty account remained a viable contender. In each of the four experiments, the order in which the two inquiries were presented to the participants (i.e., presentation order) was always positively correlated with the order in which the two inquiries were conducted (i.e., inherent temporal order). As a result, when participants learned about the second inquiry, they had already known the first inquiry, making the second inquiry more novel by the definition of novelty, that is, the discrepancy between what is known and what is discovered.

One way to appraise the merit of the novelty account relative to the hypothesized sequential-inquiry schema account is to pit presentation order against inherent temporal order by creating a vignette that describes a later inquiry before an earlier one, making the earlier inquiry more novel. If novelty were responsible for the progression bias we observed so far, then with such a vignette we would expect participants to rely more on the earlier inquiry, resulting in anti-progression bias. We adopted this strategy in this experiment.

Participants

One hundred forty-four MTurk workers (81 males, $M_{age} = 37$) participated in the experiment for monetary compensation.

Materials and Procedure

Participants were randomly assigned to either the *later-positive* ($n = 70$) or *later-negative* condition ($n = 74$). Later-positive participants read the following vignette:

> Janet's period was late for about a week. She was worried that she might have been pregnant. There was no home pregnancy test left in the medicine cabinet. So she bought a new box and did one test by urinating on the test stick. Since the test required a 10-min waiting time before displaying the result, Janet decided to make a cup of tea and read some book to distract herself. Just then, she received a phone call from her mom. When she hung up, she had forgotten where she placed the test stick.
>
> So she had to take out another stick and did a second test. This time she made a mental note of where the stick was before she went into the kitchen to make tea. When the time was up, she checked the stick; and the result indicated that she was pregnant. Later that day, Janet accidentally found the first stick on the bookshelf in her study. However, unlike the second test, the first test indicated that she was not pregnant.

Later-negative participants read the same vignette except that the results of the two tests were reversed. All participants indicated on a 1-to-100 sliding scale the extent to which they thought Janet was pregnant, where 1 = *definitely not pregnant* and 100 = *definitely pregnant*.

Note that although the same agent (i.e., Janet) administered both inquiries, the vignette is relatively immune to concerns that might afflict typical single-administrator situations. First, Janet was ignorant of the finding of the first test when she conducted the second test. Second, Janet was no novice when it came to administering pregnancy test as implied by this part of the vignette: "There was no home pregnancy test left in the medicine cabinet. So she bought a new box."

More importantly, unlike previous experiments where presentation order is positively correlated with inherent temporal order, the two orders are negatively correlated in the current experiment, allowing us to directly assess the merits of the novelty account vis-à-vis our hypothesis. If the progression bias in the previous experiments were due to the later inquiries being more novel, we would expect to observe anti-progression bias in the current experiment. Specifically, we would expect later-positive participants to be less convinced of Jane's pregnancy than their later-negative counterparts because the former read about the negative result second, while the latter read about the positive result second.

Results and Discussion

Contrary to the novelty account but consistent with our hypothesis, later-positive participants were still more convinced that Janet was pregnant ($M = 65.47$, $SD = 19.55$, 95% CI = [60.94, 70.00]) than later-negative participants ($M = 55.19$, $SD = 25.32$, 95% CI = [49.15, 61.22]), $t(142) = 2.74$, $p < .01$, Cohen's $d = 0.46$.

Experiment 6: Headphone Reviews

An alternative way of assessing the merit of the novelty account is to remove inherent temporal order altogether and

see if presentation order alone can lead to the more recently presented inquiries being unduly influential in the formation of global judgments. We adopted this approach in the current experiment to provide corroborating evidence against the plausibility of the novelty account. Note that with verbal materials, there seems to be no conceivable way of removing presentation order.

A second purpose of the current experiment was to test our hypothesis with a measure more proximal to the hypothesized process than global judgments. After all, global judgments being unduly influenced by later inquiries is not equivalent with later inquiries being regarded epistemically superior. Therefore, in this experiment, aside from global judgment, we also directly asked participants to choose between an earlier and a later inquiry on the quality basis.

Participants

Two hundred forty MTurk workers (123 males; M_{age} = 36) participated in the experiment for monetary compensation.

Materials and Procedure

Participants were randomly assigned to one of four conditions in a 2 (*Publication Order*: sequential vs. parallel) × 2 (*Rating Sequence*: later-higher vs. later-lower) between-participant design. Participants in the sequential condition read the following scenario:

> Imagine that you are interested in purchasing a pair of noise-canceling headphones for use in noisy environments such as airplane, cafeteria etc. You are drawn to a model released last December by a new company specialized in audiovisual equipment. Before you take the plunge, you want to get some expert opinions on this product.
>
> You find two rather extensive review articles of this headphone model by two highly reputable magazines. Magazine A's review appeared in its February issue while Magazine B's review appeared in its March issue. Both reviews are available in digital format but are only accessible to the magazines' annual subscribers. However, both magazines offer the nonsubscribers the option to pay a small fee to view an article of their choices. The fees are about the same for both magazines.

Participants in the parallel condition read the same scenario except that both Magazine A and B's reviews appeared in their respective March issues. All participants were asked to choose which magazine's review they would prefer if they could only pay for one article. Compared with global judgment, this dependent variable constituted a more direct measure of whether one inquiry was viewed as epistemically superior to another. If what matters was the presentation order of the inquiries rather than their inherent temporal order as prescribed by the novelty account, then participants in both sequential and parallel conditions would show a preference for Magazine B's review over A's. On the contrary, our hypothesis, which involves sequential-inquiry schema, predicted that only participants in the sequential condition would show a preference for Magazine B's review over A's. Note that up till this point, the experiment involved only one factor, that is, Publication Order.

After submitting their choices, participants in the *later-higher* condition read a "sequel" to the scenario above:

> As you are about to pay for the review article you choose, you notice that although you cannot view the full articles without paying, you can actually read some excerpts for free for both reviews. From the excerpts, you see that Magazine A's review gave the headphone model a 6.7 out of 10 rating whereas Magazine B's March gave it an 8.4 out of 10 ratings.

Participants in the later-lower condition read the same sequel except that the two magazines' ratings were reversed. All participants were then asked to report, given these ratings, how much they were interested in purchasing this headphone model on a 1-to-100 sliding scale, where 1 = *not at all interested* and 100 = *extremely interested*.

Results and Discussion

Preferred review. When Magazine B published its review later than Magazine A (i.e., the sequential condition), 81.67% (95% CI = [73.73%, 87.64%]) of the participants chose to pay for B's review. However, when Magazine B published its review at the same time as Magazine A (i.e., the parallel condition), only 17.50% (95% CI = [11.67%, 25.35%]) of the participants chose to pay for B's review, $\chi^2(1) = 96.27$, $p < .001$. These results suggest that it was sequential-inquiry schema rather than novelty that led participants to overvalue the later inquiries.

However, there seemed to be a trivial explanation for these results. To wit, in the sequential condition, participants might have assumed that Magazine B's reviewer probably had more time with the headphones and therefore had learned more about them. Having said that, this assumption could also merely be a post hoc rationalization for their affinity for Magazine B's review that stemmed from rooted some other factor such as the activation of sequential-inquiry schema. To find out if this *more-time-therefore-more-information* assumption was a post hoc rationalization or a possible cause of participants' preference, we conducted a supplemental study on an independent group of 57 MTurk workers (33 males, M_{age} = 36), who answered the following question:

> Imagine that you are interested in purchasing a pair of noise-canceling headphones for use in noisy environments. You are drawn to a model released last December by a new company specialized in audiovisual equipment. Before taking the plunge, you decide to read up on some expert reviews of this product.

> An expert can form a well-grounded opinion about an electronic product only after testing it for an extended period. A review based on one day's experience with a product tends to be less reliable than one based on three days' experience. However, after a certain point, spending extra time with the product is unlikely to improve the assessment.
>
> In the case of the noise-canceling headphones mentioned above, after how many days of testing the product do you intuitively feel an expert could NO longer make his or her review materially better by spending more time with it?

It turned out that participants did not seem to believe a priori that the additional 30 days (i.e., from February to March) Magazine B's reviewer might have had with the product was going to make any differences in the quality of his review. The median response to this question was just 3 days and the mean was 6.78 days (SD = 9.78, 95% CI = [4.18, 9.18]). In fact, even the maximum response was only 60 days. According to these results, the reviewer should have had enough time with this product already by February. It seemed that the assumption that Magazine B's reviewer likely had learned more thanks to the extra time he had with the product was more of a post hoc rationalization than an a priori reason for participants' more favorable attitude toward B's review.

Purchase intention. We conducted a 2 × 2 ANOVA on the expressed interest in purchasing the headphones. The analysis revealed a significant two-way interaction between *Publication Order* (sequential vs. parallel) and *Rating Sequence* (later-higher vs. later-lower), $F(1, 236) = 5.31$, $p = .02$, $\eta_p^2 = .02$. In the presence of inherent temporal order (i.e., when the two reviews were published in successive months), those in the later-higher condition (M = 65.02, SD = 18.02, 95% CI = [60.51, 69.52]) were more interested in purchasing the product than those in the later-lower condition (M = 49.04, SD = 23.57, 95% CI = [42.72, 55.35]), $t(118) = 4.20$, $p < .001$, Cohen's $d = 0.77$. In contrast, in the absence of inherent temporal order (i.e., when the two reviews were published in the same months), those in the later-higher condition (M = 59.27, SD = 20.95, 95% CI = [53.81, 64.73]) were not more interested in purchasing the headphones than those in the later-lower condition (M = 55.75, SD = 21.16, 95% CI = [50.34, 61.17]), $t(118) = 0.92$, $p = .36$, Cohen's d = 0.17. The finding that presentation order per se had no impact on the global judgment is again not consistent with the novelty account for progression bias but fully compatible with our hypothesis.

Experiment 7: Dot Estimation

Thus far, we have documented consistent evidence of people falling prey to progression bias in low-stakes, hypothetical scenarios. To test whether people overgeneralize sequential-inquiry schema even when their financial well-being is at stake, we conducted the current incentive-compatible experiment where participants should be motivated to avoid mistakes.

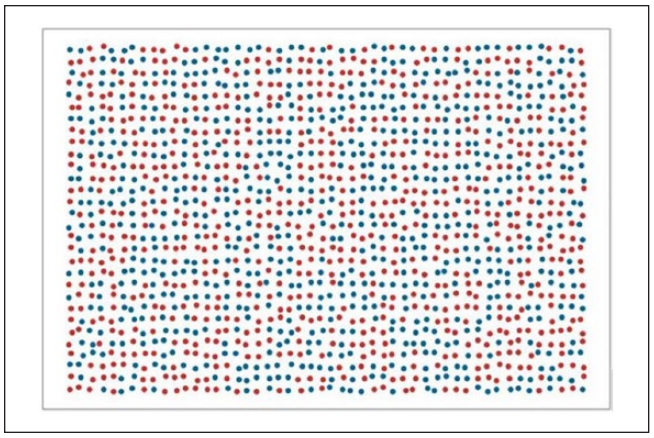

Figure 2. The blue–red dot plot.
Note. Participants were asked whether there were more blue dots or red dots in the plot. In reality, there are 745 blue dots and 755 red dots. Color figures are featured online.

Participants

One hundred MTurk workers (52 males, M_{age} = 34) participated in the experiment for monetary compensation.

Materials and Procedure

Participants were randomly assigned to either the *later-red* (n = 52) or *later-blue* condition (n = 48). All participants were shown the same image, with 745 blue dots and 755 red dots positioned within the boundary of a gray rectangle as shown in Figure 2. Participants were told that the number of blue dots and red dots differed and they were to estimate whether there were more blue dots or more red dots. Participants were promised an 80-cent bonus if they made the correct choice. The bonus promised to participants was twice the amount of the fixed payment they would have received for agreeing to take part in the experiment. Thus, participants should be sufficiently incentivized to take the task seriously.

Before making their choices, all participants were provided bogus information about the choices made by the fictitious first 100 participants of the experiment. Panels A and B in Figure 3 show the information received by the later-blue and later-red participants, respectively. All participants were told that each bar in the image represented one of the alleged first 100 participants and the bar color corresponded to his or her choice. For the later-blue participants, 35 of the first 50 bars were red, whereas 35 of the second 50 bars were blue. In contrast, for the later-red participants, 35 of the first 50 bars were blue, whereas 35 of the second 50 bars were red. Participants who made the correct choice (i.e., red dots) were awarded the promised bonus regardless of their conditions.

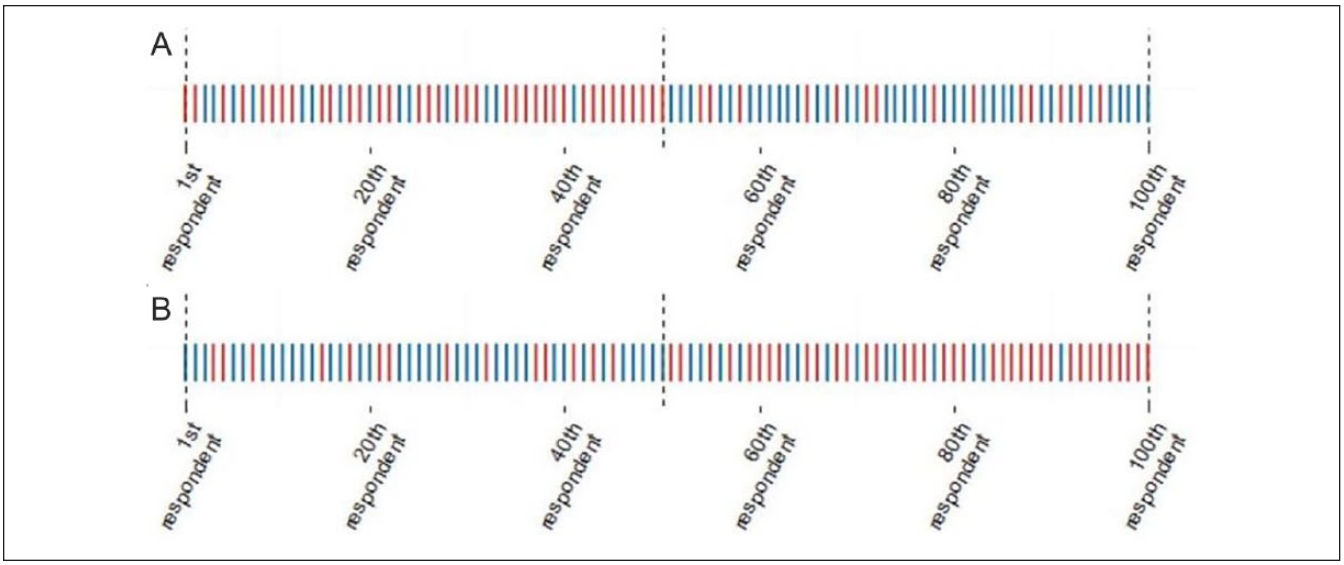

Figure 3. The choices made by the fictitious first 100 participants of Experiment 7.
Note. Panels A and B were shown to later-blue and later-red participants respectively. The color of a given bar represents the choice ostensibly made by previous participants. Color figures are featured online.

Results and Discussion

Given that the bogus information received by participants in both conditions was equally ambiguous (50% choosing red and 50% choosing blue), participants' choices should not have been affected by their conditions. Yet, a higher percentage of later-red participants (69.23%, 95% CI = [55.66%, 80.15%]) chose red than the later-blue participants (45.83%, 95% CI = [32.57%, 59.71%]), $\chi^2(1) = 4.69$, $p = .03$. Consistent with our hypothesis, the majority of the later-red participants (69.23%) believed that red dots outnumbered blue dots, while the majority of the later-blue participants (54.17%) believed in the opposite. These results suggest that even when people are motivated to be as accurate as possible, they might still succumb to progression bias.

In addition, this experiment shows that progression bias does not seem confined to valenced situations where the stakeholders are not neutral about what the truth is. For instance, in Experiment 1, the investigative committee probably hoped to find evidence of the victim harboring malintent so that they could exonerate one of their colleagues. Similarly, in Experiment 3, Mr. Jackson probably hoped that his blood cell count was within the normal range. However, in the current experiment, one would be hard-pressed to identify a reason why the fictitious participants might have had hoped one color outnumbered another.

Experiment 8: Radiation Safety

It follows from the sequential-inquiry schema account for progression bias that once people recognize temporal advancement is actually orthogonal to epistemic advancement, they would be less liable to the undue influence of later inquiries when inter-inquiries differences do not entail quality-disparities. We sought to test this prediction in the current experiment to directly evaluate the merit of the sequential-inquiry schema account.

Past research has shown that joint-evaluation can be utilized to highlight critical features that tend to be overlooked in separate-evaluation (e.g., Hsee, 1996). In the current experiment, we highlighted the lack of necessary association between temporal and epistemic advancement by having participants processing situations where mere temporal advancement occurs jointly with situations where temporal advancement co-occurs with epistemic advancement.

Participants

Two hundred forty-seven MTurk workers (108 males, $M_{age} = 36$) participated in the experiment for monetary compensation.

Materials and Procedure

Participants were randomly assigned to one of four conditions in a 2 (*Priming Type*: joint-evaluation vs. control) × 2 (*Result Sequence*: later-positive vs. later-negative) between-participants design. The experiment consisted of two stages.

In Stage 1, all participants completed a priming task consisting of four numerical estimation problems. At the outset of the priming task, participants were told that "we are interested in your estimation rather than the objectively correct answers."

The four estimation problems for participants in the joint-evaluation condition are reproduced verbatim in the following:

> (1) A soldier returned home from the front line and wanted to see how tall the boy had grown since he left. The solider had the boy stand straight against the wall and marked the position above the top of the boy's head on the wall with a pencil. Then he measured the distance from the floor to the pencil mark with a retractable measuring tape, which read 45 inches. Then he noticed that as he stretched the tape across the distance, he didn't keep the tape straight. So he measured the distance once again while making sure the tape did not sag. The second time, the tape read 41 inches. What do you think was the true height of the boy?
>
> (2) A drug dealer received a Ziploc bag filled with cocaine. He wanted to make sure he was not short changed by his supplier. So he decided to weigh the cocaine with the two scales in his kitchen. He first weighed the bag on one scale, which displayed 23 ounces. Then he put the bag on the other scale, which displayed 19 ounces. How much cocaine do you think this bag really contained?
>
> (3) The coach of a basketball team wanted to measure the vertical leap of the team's newest member (i.e. how high this person could jump off the ground from a standstill). After a 5-min warm-up session, this player jumped once and the coach measured a vertical leap of 34 inches. Then the coach noticed that this player's shoelaces were too loose. So he asked the player to tighten up his shoelaces and jump again. This time the coach's measure showed that he jumped 38 inches. What do you think was the true vertical leap of this player?
>
> (4) A biologist who conducts experiments on lab rats needed to measure the room temperature of a new facility where he kept his lab rats. He turned on the thermometer and placed it close to the left wall of the facility. He got a reading of 63°F. Then he reset the thermometer and moved it close to the right wall of the facility. This time he got a reading of 69°F. What do you think was the true temperature of this facility that houses the lab rats?

In Problems 1 and 3, the second inquiries were accompanied by epistemic improvements. On the contrary, in Problems 2 and 4, the second inquiries were only meaningfully different from the first inquiries in their temporal positions. This setup was intended to highlight the fact that there is no necessary association between temporal advancement and epistemic advancement.

The estimation problems in the control condition were adapted from those in the joint-evaluation condition, with details that might activate sequential-inquiry schema removed. We sought to ensure the cognitive involvement in the control condition would be comparable to the joint-evaluation condition. These problems are reproduced verbatim in the following:

> (1) A soldier returned home from the front line and wanted to see how tall the boy had grown since he left. The solider had the boy stand straight against the wall and marked the position above the top of the boy's head on the wall with a pencil. Then he measured the distance from the floor to the pencil mark with a retractable measuring tape. The soldier himself was 5-foot-11 tall and the boy was 22-month old. Please estimate the height of the boy as measured by the measuring tape?
>
> (2) A drug dealer received a medium-sized Ziploc bag (5-inch by 7-inch) filled with cocaine from his supplier. Wanting to make sure he was not short-changed by his supplier, the drug dealer placed this Ziploc bag on an electronic scale in his kitchen. Please estimate the weight of this Ziploc bag as measured by the kitchen scale?
>
> (3) The coach of a basketball team wanted to know the vertical leap of the team's newest member (i.e., how high this person could jump off the ground from a standstill). After a 5-min warm-up session, this player jumped once and the coach measured how high he jumped. This player was 6-foot-8 tall and weighed 230 pounds. Please estimate how high this player jumped as measured by the coach?
>
> (4) A biologist who conducts experiments on lab rats needed to measure the room temperature of a new facility where he planned to move his lab rats. Unbeknownst to him, the AC had been accidentally turned off for nearly 20 hr when he entered the room in the early morning. He powered on the thermometer and placed it in the center of the facility. The previous night, the average local temperature was 55°F and the facility was on the ground floor. Please estimate the room temperature of this new facility as measured by the biologist in the early morning?

After the priming task, participants proceeded to Stage 2 where they read a different vignette depending on the Result Sequence conditions. Later-positive participants read the following vignette:

> A property developer is considering purchasing a piece of land to build a residential complex. As a precaution, he contacts the local branches of two national organizations involved in protecting the public from the health hazard of radiation, that is, Council on Radiation Protection (CRP) and Health Physics Society (HPS). He requests each organization to send staff to check the radiation level at this site (the lower the better) before he takes the plunge.
>
> The CRP team arrives the next day and brings a CRP-designed Geiger counter to assess the radiation level. Note that a Geiger counters works by counting the number of radioactive particles it detects over a preset time span. According to the readout of the CRP team's Geiger counter, the local radiation level is below the safety level advised by the World Health Organization.
>
> A day later, the HPS team arrives with a Geiger counter designed by the HPS engineers. Unlike the CRP's finding, the readout of

> the HPS team's Geiger counter shows that the local radiation level is above the World Health Organization's advisory safety level.

Later-negative participants read the same vignette except that the readouts of the two Geiger counters were reversed. All participants were asked to rate the extent to which they believed that the piece of vacant land was unsafe in terms of radiation risk for residential development on a 1-to-100 sliding scale where 1 = *definitely safe* and 100 = *definitely unsafe*.

We predicted that participants in the control condition would be more likely to display progression bias in Stage 2 than their counterparts in the joint-evaluation condition because the latter group were supposedly more cognizant of the dissociation between temporal and epistemic advancement thanks to the priming task.

Results and Discussion

We conducted a 2 × 2 ANOVA on participants' ratings of the radiation risk of the vacant land. The analysis revealed a marginally significant two-way interaction between Priming Type (joint-evaluation vs. control) and Result Sequence (later-positive vs. later-negative), $F(1, 243) = 3.49$, $p = .063$, $\eta_p^2 = .014$. In light of our a priori predictions, we proceeded to decompose the interaction with simple effect analyses. Among participants in the control condition during Stage 1, those in the later-positive condition ($M = 62.61$, $SD = 18.17$, 95% CI = [58.00, 67.23]) considered the land to be more unsafe than those in the later-negative condition ($M = 53.57$, $SD = 21.11$, 95% CI = [48.11, 59.02]), $t(120) = 2.54$, $p = .012$, Cohen's $d = 0.46$, thereby showing progression bias. Among participants in the joint-evaluation condition in Stage 1, those in the later-positive condition ($M = 67.89$, $SD = 23.59$, 95% CI = [61.90, 73.88]) considered the land to be just as unsafe as those in the later-negative condition ($M = 68.67$, $SD = 19.37$, 95% CI = [63.79, 73.54]), $t(123) = -0.20$, $p = .84$, Cohen's $d = 0.036$, thereby showing no progression bias. It seemed that participants who were reminded of orthogonality between temporal and epistemic advancement were more discriminating in applying sequential-inquiry schema. Therefore, these results seem to lend direct support for our hypothesis that progression bias arises because people overgeneralize the sequential-inquiry schema.

Experiment 9: Exact Replication

Inspired by the opening anecdote about a psychologist's frustration, the current experiment explored progression bias's relevance for the ongoing replication crisis afflicting our home field of social and personality psychology.

Participants

One hundred fifteen MTurk workers (66 males, $M_{age} = 40$) participated in the experiment for monetary compensation.

Materials and Procedure

Participants were randomly assigned to either the *later-positive* ($n = 57$) or *later-negative* condition ($n = 58$) condition. Later-positive participants read the following vignette:

> A psychology professor wanted to find out if viewing abstract art can make people behave more rationally than viewing impressionist art. The professor paid Amazon to have 100 MTurk workers to complete an online survey designed to answer his question. Unbeknownst to the survey takers, the content of the first task of the survey varied across them. Approximately half of the survey takers viewed several abstract paintings in the first task while the other half viewed several impressionist paintings. Afterwards, all survey takers completed a financial decision-making task.
>
> When the data collection was completed, the professor analyzed the survey takers' responses and found that MTurkers who first viewed abstract paintings did not make more rational decisions than those who first viewed impressionist ones. To be certain, the professor relaunched his survey on MTurk a week later and recruited a new batch of MTurkers to complete it.
>
> When the professor analyzed the responses of the 100 newly recruited survey takers, he found that MTurkers who first viewed abstract paintings made more rational decisions than those who first viewed impressionist ones, contrary to the finding from the first attempt.

The later-negative participants read the same vignette except that the results of the two attempts were reversed. All participants rated how much they believed that abstract art really could promote rational thinking based on the information in the vignette on a 1-to-100 scale, where 1 = *not at all* and 100 = *definitely*.

Results and Discussion

The later-positive participants were more convinced of the professor's hypothesis that abstract art could promote rational thinking ($M = 47.28$, $SD = 30.71$, 95% CI = [39.13, 55.43]) than the later-negative participants ($M = 34.52$, $SD = 30.83$, 95% CI = [26.41, 42.62]), $t(113) = 2.22$, $p = .028$, Cohen's $d = 0.41$. It seemed that when evaluating a viability of a scientific idea, at least laypeople tend to pay disproportionate attention to the finding of a replication study than that of the original study.

Although it remains to be seen whether psychologists are any better at resisting progression bias when placed in similar situations as the participants of the current experiment, we are not optimistic if extant research on how experts are prone to various judgment biases is of any indication (e.g., Fischhoff, Slovic, & Lichtenstein, 1982; Kahneman & Tversky, 1982; Klein, 2005; Slovic, 1999). If we psychologists are just as susceptible to progression bias as common folks, we might want to reconsider our assessment of the

field amid the ongoing replication crisis to ensure that it was not unduly influenced by the more recent replication failures. It is not in the interest of the field to allow our evaluation of a program of research swayed by a typically meaningless factor.

Experiment 10: Cholesterol-Reduction Drug

In the final experiment, we sought to probe the implications of progression bias for ethical practices in business settings. In addition, note that in all the vignette-based experiments so far, the number of inquiries to be integrated was fixed at two. To test the generality of progression bias, the vignette in this experiment includes three inquiries.

Participants

One hundred four MTurk workers (48 males, $M_{age} = 35$) participated in the experiment for monetary compensation.

Materials and Procedure

Participants were randomly assigned to either the *later-positive* ($n = 56$) or *later-negative* condition ($n = 48$). Later-positive participants read the following vignette:

> A pharmaceutical company has recently developed a new chemical compound, which is expected to be more effective at reducing blood cholesterol levels than other drugs currently on the market.
>
> In the first two clinical trials, each with a different group of high-cholesterol patients, the company found that the new chemical was no more effective at cutting down cholesterol level than a commonly used generic drug. The company decided to conduct one more clinical trial, with the same number of patients. However, in this most recent trial, the new chemical was shown to be more effective than the generic drug.
>
> The company decided that the unexpected results of first two trials were probably due to some fluke and went ahead publishing the results of the last clinical trial, proclaiming the potential of this new chemical as a new drug for treating cholesterol-related cardiac diseases.

Later-negative participants read the following vignette instead:

> A pharmaceutical company has recently developed a new chemical compound, which is expected to be more effective at reducing blood cholesterol levels than other drugs currently on the market.
>
> In the first clinical trial with a group of high-cholesterol patients, the company found that the new chemical was more effective at cutting down cholesterol level than a commonly used generic drug. The company decided to conduct two more clinical trials, both with the same number of patients. However, in neither of the more recent two trials was the new chemical shown to be more effective than the generic drug.
>
> The company decided that the unexpected results of last two trials were probably due to some fluke and went ahead publishing the results of the first clinical trial, proclaiming the potential of this new chemical as a new drug for treating cholesterol-related cardiac diseases.

After the vignette, all participants responded to two questions: (a) Based on the results of all three clinical trials, how much they believed that this new chemical was indeed more effective at reducing cholesterol levels than the generic drug, and (b) to what degree they thought the company was justified in attributing the unexpected results of the two failed clinical trials to flukes. Participants indicated their responses to either questions on a 1-to-9 Likert-type scale, where 1 = *not believe at all* or *not at all justified*, respectively, and 9 = *very much believe so* or *extremely justified*, respectively.

Results and Discussion

Consistent with our prediction, later-positive participants were more convinced of the medical efficacy of the company's new chemical ($M = 4.39$, $SD = 2.23$, 95% CI = [3.80, 4.99]) than the later-negative participants ($M = 2.92$, $SD = 1.91$, 95% CI = [2.36, 3.47]), $t(102) = 3.59$, $p < .001$, Cohen's $d = 0.71$. Crucially, we found that later-positive participants were less harsh ($M = 3.46$, $SD = 2.30$, 95% CI = [2.85, 4.08]) toward the company's decision to withhold the null results than later-negative participants ($M = 2.08$, $SD = 1.85$, 95% CI = [1.54, 2.62]), $t(102) = 3.33$, $p = .001$, Cohen's $d = 0.65$.

General Discussion

When seeking out truths, people often conduct multiple inquiries sequentially over a time span. Usually, each successive inquiry is conducted in light of the lessons from its predecessor and therefore should be better positioned to unveil the truth. In these prototypical cases, temporal advancement is accompanied by changes that entail epistemic improvements. However, there are numerous aberrant cases where the differences between a later inquiry and an earlier one do not confer epistemic advantages on the former. For instance, a second reading is unlikely to give a dieter who steps on the scale twice within a minute a better sense of his weight-loss progress. Nonetheless, we consistently found that, even in these aberrant cases, people still seemed to treat temporal advancement as a signal of epistemic advancement. When successive inquiries yielded conflicting results, their global judgments were disproportionately influenced by the results

of the later inquiries, exhibiting a mistake we term progression bias.

Throughout the present report, we discussed at length four plausible accounts for progression bias, which involve sequential-inquiry schema, practice-effect heuristic, dispositional optimism, and novelty, respectively. In a sense, all four accounts are fundamentally the same in that they all can be interpreted as the overgeneralization of an otherwise adaptive implicit belief (or lay theory) to speciously compatible situations where it ceases being applicable. The implicit beliefs involved in these four accounts are, respectively, new inquiries are informed by past ones (sequential-inquiry schema), practice makes perfect (practice-effect heuristic), things generally get better (dispositional optimism), and novel things are more informative (novelty). Therefore, none of the four accounts treat progression bias as rational. Furthermore, these four accounts are not mutually exclusive and could conceivably be operating simultaneously in many situations to tip the scale toward the later inquiries. Nonetheless, the data in the present report as a whole are more in favor of our hypothesis that progression bias mainly stems from the misapplication of sequential-inquiry schema than the other three alternative accounts. When a situation involving a series of inquiries aimed at the same truth is filtered through the lens of sequential-inquiry schema, even inter-inquiry differences that are mere random variation in nature might be viewed as having epistemic implications.

Progression Bias Versus Recency Bias

At first glance, progression bias observed in this research bears a striking resemblance to the various types of recency bias extensively documented in the literature on order effects (e.g., de Bruin, 2005; Garbinsky, Morewedge, & Shiv, 2014; Krosnick, 1999). For instance, participants who were presented with descriptions of three blind-date candidates sequentially tended to rate the last presented candidate the highest (de Bruin & Keren, 2003). As another example, a story was evaluated more positively when it was among the last few stories evaluated than when it was among the first few (O'Connor & Cheema, 2018). How does progression bias fit into this literature? We argue that progression bias we studied in the present research, despite the appearance, differs fundamentally from recency bias studied in order effect literature in at least three ways.

First, progression bias arises when people need to form a global judgment that integrates individual items of information (i.e., findings from the successive inquiries). The key dependent variable is how the individual items should be evaluated *as a whole*. On the contrary, recency bias typically appears in the context where the key dependent variable is the evaluation of individual items *on their own*. In an aforementioned experiment (de Bruin & Keren, 2003), participants gave a rating to each of the three sequentially presented blind-date candidates.

Second, what matters for progression bias is the inherent temporal order of the individual items as we showed in Experiments 5 and 6. In contrast, what matters for recency bias is the presentation order of the individual items. In an experiment mentioned earlier where participants evaluated a series of stories (O'Connor & Cheema, 2018), it was whether one story was read after another rather than one story was written after another that affected participants' ratings. The creation times of these stories were not even mentioned.

Third, the intrinsic limitation of human memory plays at most a marginal role in progression bias but is critical in recency bias. In our experiments, there was neither a protracted intervening period nor a large amount of distracting information separating the successive inquiries. The two to three individual items of information that needed to be integrated can be simultaneously processed in working memory, which is capable of accommodating four chunks of information (Engle, 2002). In contrast, in many cases, recency bias occurs exactly because people could not simultaneously evaluate the individual items in a series due to memory constraints. For instance, Li and Epley (2009) demonstrated that when options in a set were all desirable, the compromised memory for the options presented toward the beginning was the reason why people were more favorable toward options presented toward the end.

Implications

It would not be hard to imagine a wide range of harms that progression bias, when scaled up to real life, can potentially inflict upon individuals, organizations, and societies. A patient may be less diligent about following a health regimen because the nurse on duty in the afternoon undercounted his white blood cells compared with in the morning (Experiment 4). A magazine may end up losing profits to its competitor for the paradoxical reason that it beat its competitor to the punch (Experiment 6). The scientific community might overreact to a replication failure, thereby smothering a promising idea prematurely (Experiment 9).

"Prisoners of the moment" is a term that is experiencing a surge in popularity in sports journalism (e.g., Mike, 2011). It describes a person whose opinions are disproportionately influenced by what is happening just now without giving due consideration to what happened before. For example, when discussing and evaluating an NBA draft prospect, the press and the public tend to overemphasize the last few NCAA tournament games he played. In all likelihood, this overvaluation of March Madness could just be progression bias writ large. Although a college player's performance in March Madness games of the year he declares his eligibility for the NBA draft constitutes only a small part of his full body of work, it is the latest evidence of his potential as a pro.

We also suspect that progression bias might have played a role in the worst environmental disaster in American history. In 2010, 11 workers lost their lives and 17 others were injured

when the drilling rig, Deepwater Horizon, exploded and spilled 206 million gallons of oil into the Gulf of Mexico. The failure to correctly interpret the results of safety tests played a crucial role in the decision to remove drilling mud and surface plugs that would have prevented the explosion. According to one source, two negative pressure tests had been performed on the day of the accident (APPEL News Staff, 2011). Not unlike the later-positive condition in experiments in this present research, the first test advised against drilling but the second test supported it. Would the disaster have been averted, if the results of the two tests had been reversed?

Progression bias might also be implicated in lowballing, a time-tested but morally questionable sales technique (Cialdini, Cacioppo, Bassett, & Miller, 1978; Joule, 1987). In lowballing, a product is initially offered at a lower price than what is always intended, but immediately prior to the finalization of the deal, the price is raised to the intended level. Although the final offer is much less desirable than the initial offer, rarely do consumers back out of the deal once they agree to the initial offer. The commonly accepted explanation for the effectiveness of the lowball technique is people's aversion to reneging on a commitment, which would result in cognitive dissonance (Burger & Petty, 1981). However, if we reconstrue each successive offer in a lowballing situation as the finding from an inquiry conducted by the salesperson to uncover the true value of the target product, we can think of progression bias as a nonmotivational alternative account of this technique's efficacy. Perhaps, people consider the price named in the final offer more reflective of a product's true value than the price named in the initial offer.

Possible Solutions

Given the cost progression bias could inflict on both groups and individuals, an important question that awaits further exploration is how to protect people from the potential harms of overgeneralizing sequential-inquiry schema (Arkes, 1991). Extant research has documented the bias-reducing effect of an increase in either financial incentive or accountability (e.g., Tetlock, 1985). However, raising the stakes of accuracy (Experiment 7)—which presumably encourages System 2 or central-route processing (Evans, 2008; Kahneman & Frederick, 2007; Petty & Cacioppo, 1986)—failed to eliminate the progression bias.

Another approach to debiasing is training. Morewedge and colleagues (2015) showed that exposing participants to one-shot training interventions, such as educational videos, significantly attenuated participants' likelihood of committing a variety of cognitive biases immediately, and up to 3 months, after the intervention. The priming task we used in Experiment 8 may be a potential antidote to the progression bias although further research is needed to assess the durability of the effect.

Limits on Generality

Although the experimental tasks in this research were designed to simulate the types of situations people might encounter in real life, they still deviate from real life on no less than four crucial fronts, which could limit the generality of this research. First and foremost, in real life when people find themselves in ambiguous situations similar to those faced by our participants, they typically can either proactively seek out extra information (e.g., conducting another inquiry) to help adjudicate between the conflicting pieces of evidence or defer making the global judgment until better data become available. Neither option was available to our participants. Therefore, progression bias–related mistakes might not be as common in real life as the data in the present research might suggest. Second, in the present research, people in charge of making global judgments (i.e., the participants) were not at all involved in the administration of the inquiries they were to integrate. However, in many real-life situations, people making the global judgments are also the ones who carry out the inquiries. For instance, radiologists usually base their diagnoses on the imaging tests they conduct themselves. Whether this characteristic of real life would exacerbate or attenuate progression bias is unanswered by the present research. Last but not least, in real life, forming a global judgment that integrates disparate pieces of information is rarely the end goal. Often, what matters are the behaviors that follow from and are supposedly informed by global judgments. As a result, claims about the practical implication of progression bias should be taken with a grain of salt until we know how progression bias would propagate through global judgments to the downstream behaviors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first author received financial support from ShanghaiTech University for the research, authorship, and publication of this article.

Supplemental Material

Supplemental material is available online with this article.

References

APPEL News Staff. (2011). *Academy case study: The deepwater horizon accident lessons for NASA*. Retrieved from https://appel.nasa.gov/2011/05/11/aa_4-4_acs_deepwater_horizon_lessons-html/

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, *110*, 486-498.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*, 207-230.

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

Burger, J. M., & Petty, R. E. (1981). The low-ball compliance technique: Task or person commitment? *Journal of Personality and Social Psychology*, *40*, 492-500.

Cialdini, R. B., Cacioppo, J. T., Bassett, R., & Miller, J. A. (1978). Low-ball procedure for producing compliance: Commitment then cost. *Journal of Personality and Social Psychology*, *36*, 463-476.

Corker, K. (2015, August 7). When Study 1 & Study 2 disagree: Practical recommendations for researchers. Retrieved from https://scienceofpsych.wordpress.com/2015/08/07/when-study-1-and-study-2-disagree-practical-recommendations-for-researchers/

Dahlgreen, W. (2016). *Chinese people are most likely to feel the world is getting better*. Retrieved from https://yougov.co.uk/topics/lifestyle/articles-reports/2016/01/05/chinese-people-are-most-optimistic-world

de Bruin, W. B. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, *118*(3), 245-260.

de Bruin, W. B., & Keren, G. (2003). Order effects in sequentially judged options due to the direction of comparison. *Organizational Behavior and Human Decision Processes*, *92*, 91-101.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19-23. doi:10.1111/1467-8721.00160

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278. doi:10.1146/annurev.psych.59.103006.093629

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*, 668-670. doi:10.1126/science.146.3644.668

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1982). Lay foibles and expert fables in judgments about risk. *American Statistician*, *36*, 240-255.

Fiske, S. T., & Linville, P. W. (1980). What does the schema concept buy us? *Personality and Social Psychology Bulletin*, *6*, 543-557.

Garbinsky, E. N., Morewedge, C. K., & Shiv, B. (2014). Interference of the end: Why recency bias in memory determines when a food is consumed again. *Psychological Science*, *25*, 1466-1474. doi:10.1177/0956797614534268

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451-482. doi:10.1146/annurev-psych-120709-145346

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*, 400-407.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *67*, 247-257. doi:10.1006/obhd.1996.0077

Joule, R. V. (1987). Tobacco deprivation: The foot-in-the-door technique versus the low-ball technique. *European Journal of Social Psychology*, *17*, 361-365.

Kahneman, D., & Frederick, S. (2007). Frames and brains: Elicitation and control of response tendencies. *Trends in Cognitive Sciences*, *11*(2), 45-46. doi:10.1016/j.tics.2006.11.007

Kahneman, D., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Kernis, M. H., Brockner, J., & Frankel, B. S. (1989). Self-esteem and reactions to failure: The mediating role of overgeneralization. *Journal of Personality and Social Psychology*, *57*, 707-714. doi:10.1037/0022-3514.57.4.707

Kleider, H. M., Pezdek, K., Goldinger, S. D., & Kirk, A. (2008). Schema-driven source misattribution errors: Remembering the expected from a witnessed event. *Applied Cognitive Psychology*, *22*, 1-20.

Klein, J. G. (2005). Five pitfalls in decisions about diagnosis and prescribing. *British Medical Journal*, *330*(7494), Article 781.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567.

Kuhn, T. S. (2012). *The structure of scientific revolutions*. University of Chicago Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=3eP5Y_OOuzwC&oi=fnd&pg=PR5&dq=structure+of+scientific+revolutions&ots=xU2SF5iKrM&sig=S9AEhOazNSruaS20VM-mI22-7Ao

Kumaran, D., & Maguire, E. A. (2007). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus*, *17*, 735-748. doi:10.1002/hipo.20326

Li, Y. E., & Epley, N. (2009). When the best appears to be saved for last: Serial position effects on choice. *Journal of Behavioral Decision Making*, *22*, 378-389.

Mather, E. (2013). Novelty, attention, and challenges for developmental psychology. *Frontiers in Psychology*, *4*, 491. doi:10.3389/fpsyg.2013.00491

Mike. (2011). *Prisoners of the moment (repost)*. Retrieved from http://www.ruthlessgolf.com/2011/11/prisoners-of-moment-repost.html

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights From the Behavioral and Brain Sciences*, *2*, 129-140. doi:10.1177/2372732215600886

O'Connor, K., & Cheema, A. (2018). Do evaluations rise with experience? *Psychological Science*, *29*, 779-790.

Peterson, C. (2000). The future of optimism. *American Psychologist*, *55*, 44-55.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion* (pp. 1-24). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4612-4964-1_1

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, *4*, 193-202. doi:10.1038/nrn1052

Roser, M., & Nagdy, M. (2018). Optimism & pessimism. Retrieved from https://ourworldindata.org/optimism-pessimism

Rumelhart, D. E. (1978). *Schemata: The building blocks of cognition*. San Diego: Center for Human Information Processing, University of California.

Scheier, M. F., & Carver, C. S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, *4*, 219-247.

Slovic, P. (1999). Trust, emotion, sex, politics, and science: Surveying the risk assessment battlefield. *Risk Analysis*, *19*, 689-701.

Sternberg, R. J., Roediger, H. L., & Halpern, D. F. (2007). *Critical thinking in psychology*. New York, NY: Cambridge University Press.

Tetlock, P. E. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly*, *48*, 227-236.

Thompson, R. F., & Spencer, W. A. (1966). Habituation: A model phenomenon for the study of neuronal substrates of behavior. *Psychological Review*, *73*, 16-43. doi:10.1037/h0022681

Tomasello, M., & Herron, C. (1988). Down the garden path: Inducing and correcting overgeneralization errors in the foreign language classroom. *Applied Psycholinguistics*, *9*, 237-246. Retrieved from https://www.cambridge.org/core/journals/applied-psycholinguistics/article/down-the-garden-path-inducing-and-correcting-overgeneralization-errors-in-the-foreign-language-classroom/2FF2760DC532294AF7E0E8FE7AD75EC4

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293-315.

Zebrowitz, L. A., & Rhodes, G. (2004). Sensitivity to "bad genes" and the anomalous face overgeneralization effect: Cue validity, cue utilization, and accuracy in judging intelligence and health. *Journal of Nonverbal Behavior*, *28*, 167-185. doi:10.1023/B:JONB.0000039648.30935.1b