Journal of Experimental Social Psychology 96 (2021) 104154

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

# Retrospective and prospective hindsight bias: Replications and extensions of Fischhoff (1975) and Slovic and Fischhoff (1977)☆

Jieying Chen[a,*,1], Lok Ching Kwan (Roxane)[b,1], Lok Yeung Ma (Loren)[b,1], Hiu Yee Choi (HayleyAnne)[b,1], Ying Ching Lo (Lita)[b,1], Shin Yee Au (Sarah)[b,1], Chi Ho Tsang (Toby)[b,1], Bo Ley Cheng[b], Gilad Feldman[b,*]

[a] Department of Business Administration, University of Manitoba, Canada
[b] Department of Psychology, University of Hong Kong, Hong Kong SAR, China

ARTICLE INFO

*Keywords:*
Hindsight bias
Knew-it-all-along effect
Outcome knowledge
Judgment and decision making
Surprise
Confidence
Pre-registered replication

ABSTRACT

Hindsight bias refers to the tendency to perceive an event outcome as more probable after being informed of that outcome. We conducted very close replications of two classic experiments of hindsight bias and a conceptual replication testing hindsight bias regarding the perceived replicability of hindsight bias. In Study 1 ($N = 890$), we replicated Experiment 2 in Fischhoff (1975), and found support for hindsight bias in retrospective judgments ($d_{mean} = 0.60$). In Study 2 ($N = 608$), we replicated Experiment 1 in Slovic and Fischhoff (1977), and found support for hindsight bias in prospective judgments ($d_{mean} = 0.40$). In Study 3 ($N = 520$) we found strong support for hindsight bias regarding perceived likelihood of our replication of hindsight bias ($d = 0.43$–$1.03$). We also included extensions examining surprise, confidence, and task difficulty, yet found mixed evidence with weak to no effects. We concluded support for hindsight bias in both retrospective and prospective judgments, and in evaluations of replication findings, and therefore call for establishing measures to address hindsight bias in valuations of replication work and interpreting research outcomes. All materials, data, and code, were shared on: https://osf.io/nrwpv/.

## 1. Hindsight bias

Hindsight bias refers to the tendency to perceive an event outcome as more probable after being informed of that outcome, resulting in the illusion that the outcome "was known all along" (Fischhoff, 1975; Hawkins & Hastie, 1990; Roese & Vohs, 2012). Examples of hindsight bias include claims that a surprising movie ending was actually predictable, post-election claims that it was obvious who would get elected, students feeling like they knew in advance that an unlikely question was to be on the exam, or financial analysts claiming to have predicted market changes after they happened. Hindsight bias may also affect researchers' interpretations of study findings, leading to an overestimation of their ability to predict the results beforehand and an underestimation of their reliance on the observed outcomes in reconstructing their previous predictions (Fischhoff, 1977).

The earliest empirical investigation that touches upon the idea of hindsight bias that we know of dates back to Forer's (1949) study about students' beliefs about a personality test (see Hoffrage & Pohl, 2003). Students were asked to rate the extent to which the test revealed basic characteristics of their personality, and then recall their ratings after knowing that the feedback received by all students was the same. Although Forer (1949) focused on examining how individuals could be fooled by universal statements about personality (e.g., "At times you are extroverted, affable, sociable, while at other times you are introverted, wary, reserved"), this study uncovered the unexpected finding that feedback may affect memory.

A more formal investigation of hindsight bias came in the mid-1970s, when Fischhoff (1975) published a study that explicitly compared the

☆ This paper has been recommended for acceptance by Professor Michael Kraus.

* Corresponding author.
*E-mail addresses:* jieying.chen@umanitoba.ca (J. Chen), rk1128@hku.hk, rk1128@connect.hku.hk (L.C. Kwan), loren14@connect.hku.hk (L.Y. Ma), hychoi@connect.hku.hk (H.Y. Choi), u3527928@connect.hku.hk (Y.C. Lo), u3519865@connect.hku.hk (S.Y. Au), tbtsang@connect.hku.hk, 13tsangtc1@kgv.hkBo (C.H. Tsang), boleystudies@gmail.com (B.L. Cheng), gfeldman@hku.hk (G. Feldman).
[1] Contributed equally, joint first authors

https://doi.org/10.1016/j.jesp.2021.104154
Received 22 May 2020; Received in revised form 1 April 2021; Accepted 19 April 2021

probability estimates of outcomes before (in foresight) and after (in hindsight) knowing what outcome actually occurred. In this pioneering study, participants were presented with four scenarios and four possible outcomes following each scenario. Then, they were asked to estimate the probabilities of possible outcomes in those scenarios. Some participants were informed of the outcomes of the scenarios, whereas the rest were not. Fischhoff found that participants with outcome knowledge estimated the probability of the informed outcome to be higher than participants who were not given any outcome information, demonstrating hindsight bias. Because this effect held despite the instructions to ignore outcome knowledge, Fischhoff (1975) suggested that individuals were either unaware of their bias, or, if they were aware, they were unable to make judgments in a foresightful state of mind (though Dietvorst and Simonsohn, 2019 suggested an alternative accuracy-based account).

Since the Fischhoff (1975) article was published, hindsight bias has attracted much scholarly attention and led to a sizable body of follow-up research. Several studies investigated whether hindsight bias was "real," or whether it was induced by demand characteristics. For example, Fischhoff (1977) and Wood (1978) found that hindsight bias still held when outcome knowledge was provided as isolated statements, when outcome knowledge was provided with a delay, and when participants were asked to respond as if they were a general college student who might not have known the outcome. These findings alleviated the concern about demand characteristics.

Later studies also differentiated between two main ways to examine hindsight bias (Pohl, 2007). The design used by Fischhoff (1975) is termed the hypothetical design, as participants in the hindsight condition receive feedback about the actual outcome (or, the correct answer), but are asked to answer as if they did not know the outcome. These "as if" answers are then compared with answers by participants in the foresight condition who receive no feedback. The other design is the memory design, in which participants in the hindsight condition first answer some questions, then are informed of the correct answer, and at the end are asked to recall their initial answers (Fischhoff & Beyth, 1975; Wood, 1978). Their recalled answers are then compared with their initial answers.

The hypothetical design and the memory design share many similarities, yet one distinction between them is noteworthy: hindsight bias detected using the memory design is mostly associated with memory distortion and/or the feeling that the known outcome was to happen inevitably, whereas hindsight bias that occurs in the hypothetical design may entail more complex psychological processes (Roese & Vohs, 2012).

Hindsight bias has had significant impact on a wide array of disciplines going beyond psychology, such as economics, management, health science, and law (e.g., Bukszar & Connolly, 1988; Casper, Benedict, & Perry, 1989; Kaplan & Barach, 2002; Thaler, 2016).

## 2. Reasons for hindsight bias: Emotions

Multiple factors were suggested as possible causes for hindsight bias (Blank, Musch, & Pohl, 2007; Hawkins & Hastie, 1990; Roese & Vohs, 2012), including 1) cognitive processes such as memory impairment, biased reconstruction, and sense-making, 2) meta-cognitive processes involving experiences such as surprise, confidence, experienced fluency, ease of reasoning, and 3) social-motivational processes to increase controllability and enhance self-image.

Several models have been proposed to explain hindsight bias. The Reconstruction After Feedback with Take the Best (RAFT; Hoffrage, Hertwig, & Gigerenzer, 2000) model suggested that when a direct recall of the initial answer is not possible, individuals try to reconstruct their initial answer by using relevant cues to reevaluate the question. Both the initial evaluation and the reconstructed evaluation are based on a Take the Best heuristic, where decision is based on the cue that discriminates among choices and has the highest validity. Because feedback transforms the values of elusive cues into discriminating ones and shifts cue values asymmetrically toward the feedback, the reconstructed answer will also be biased toward the feedback, demonstrating hindsight bias. The Selective Activation and Reconstructive Anchoring (SARA; Pohl, Eisenhauer, & Hardt, 2003) model assumes that individuals generate answers, encode feedback, and recall answers based on a probabilistic sampling of associations among external cues and units in the knowledge base. When individuals encode the feedback into their knowledge base, the associations among external cues, feedback, and units that are similar to the feedback are strengthened. This will render units that are more similar to the feedback more likely to be activated in a memory search using those external cues (i.e., selective activation). In addition, after seeing the feedback, individuals may still maintain the feedback in the working memory, or have increased cognitive accessibility to the feedback due to its recent activation. In these cases, feedback may be used as internal retrieval cues, making units similar to the feedback more likely to be retrieved to the working memory (i.e., biased reconstruction). According to SARA, either selective activation or biased reconstruction, or both, can lead to hindsight bias.

In both RAFT and SARA, when encoding feedback, the changes to the knowledge base, cue values, and associations occur automatically. Such knowledge updating is often seen as an adaptive learning process (e.g., Hawkins & Hastie, 1990; Hertwig, Fanselow, & Hoffrage, 2003; Hoffrage et al., 2000; Pohl, Bender, & Lachmann, 2002). However, as Bernstein et al. (2011, p. 389) wrote, "the downside of such automatic knowledge updating is that people tend to forget their original, naive thoughts, views, and predictions."

Other eminent models about the psychological processes underlying hindsight bias include the causal model theory (Blank & Nestler, 2007), Pezzo's (2003) sense-making model, Roese and Vohs' (2012) three-level model, and Sanna and Schwarz's (2006) metacognitive model.

## 3. Role of surprise, overconfidence, and task difficulty

Emotions such as surprise and overconfidence have been suggested as factors in cognitive and metacognitive processes leading to hindsight bias (Bernstein, Aßfalg, Kumar, & Ackerman, 2016). Fischhoff and Beyth (1975, p. 12) argued that "the occurrence of an event increases its reconstructed probability and makes it less surprising than it would have been had the original probability been remembered." They operationalized surprise as "the occurrence of an unlikely event or the nonoccurrence of a likely event" (Fischhoff & Beyth, 1975, p. 12), and found that outcome knowledge reduced surprise (i.e., participants made decreased probability estimates of unlikely events and increased probability estimates of likely events after knowing the outcome). Slovic and Fischhoff (1977, Experiment 3) was the first study that we know of to examine the relationship between subjective surprise feelings and hindsight bias. In this experiment, "hindsight subjects assessed the surprisingness of the reported outcome, and foresight subjects assessed how surprising each of the two possible outcomes would seem were they obtained" (Slovic & Fischhoff, p. 549). They found direct support for the hypothesis that hindsight participants who had outcome knowledge felt less surprised about the outcome than foresight participants who had no outcome knowledge. Later studies investigating the role of surprise in hindsight bias either measured surprise as a subjective feeling (e.g., Hoch & Loewenstein, 1989; Ofir & Mazursky, 1997) or manipulated surprise using expected outcomes or high cognitive loads (e.g., Mazursky & Ofir, 1990; Müller & Stahlberg, 2006).

In addition, some studies found that when experiencing surprise about a highly unusual outcome, individuals may show a reversed hindsight bias, such that their reconstructed probability estimates of the outcome becomes lower than their initial probability estimates (Mazursky & Ofir, 1990; Müller & Stahlberg, 2007; Ofir & Mazursky, 1997). The underlying rationale is that hindsight bias often results from a cognitive failure to become aware of the distorted memory and evidence reconstruction, and to recognize how much oneself has learned from the outcome knowledge prior to the estimation. The feeling of surprise is linked with an awareness that the outcome is different from

what they would have expected given their knowledge of the event. Therefore, when experiencing high levels of surprise, individuals are more likely to conclude that they "never would have known it," estimating the outcome probability to be lower (rather than higher) than the estimates made by individuals without outcome knowledge (Mazursky & Ofir, 1990; Müller & Stahlberg, 2007; Ofir & Mazursky, 1997; Sanna & Schwarz, 2006).

Whereas surprise may help individuals overcome hindsight bias, overconfidence may exacerbate hindsight bias, as it reduces individuals' scrutiny of their own decision-making process and hinders the recognition of the impact of outcome knowledge (Bernstein et al., 2016). Winman, Juslin, and Björkman (1998) found support for a confidence-hindsight mirror effect: tasks that yielded overconfidence led to a hindsight bias, whereas tasks that yielded underconfidence led to a reversed hindsight bias.

The impact of overconfidence and hindsight bias may escalate. For example, physicians may become more overconfident about their judgments of certain physiological indices over time due to accumulated outcome knowledge, which can lead to increasingly stronger hindsight bias (Arkes, 2013). However, studies indicated little to no relationship between physicians' confidence about their judgments of physiological indices and the real accuracy of those judgments (e.g., Dawson et al., 1993; Yang & Thompson, 2010). Thus, without proper caution, the escalation of overconfidence and hindsight bias may lead to undesirable consequences in high-stake decisions.

Other studies investigated the role of task difficulty in hindsight bias (e.g., Harley, Carlsen, & Loftus, 2004), based on the assumption that task difficulty is related to both surprise about the outcome and confidence about the accuracy of one's own judgment (Winman et al., 1998). The arguments are similar to those regarding surprise and confidence.

## 4. Implications of hindsight bias for Science

Hindsight bias holds implications for science, and shows the importance of the ongoing credibility revolution in promoting open science practices (Hom Jr & Van Nuland, 2019; Kerr, 1998; Nosek, Ebersole, DeHaven, & Mellor, 2018; Shrout & Rodgers, 2018; Veldkamp, 2017). First, retrospective hindsight bias suggests that being presented with a study's outcome may lead to overestimating the probability of that outcome. This may result in the skewed perception that this outcome was the expected result and in line with own expectations even when it was not the case. Past research has shown that when evaluating research findings, individuals who had outcome knowledge perceived the research findings to be more obvious and inevitable than individuals who had no outcome knowledge (Wong, 1995). The false belief of having known the outcome all along may lead to Hypothesizing After the Results are Known (HARKing; i.e., presenting a post-hoc hypothesis as if it were an a priori hypothesis; Kerr, 1998), which has been identified as a questionable research practice (QRPs). HARKing makes exploratory analyses seem as if they were confirmatory, thereby leading to an overconfidence in the reported findings and fewer follow-up confirmatory studies, overall increasing rate of false-positive findings in the literature (Bosco, Aguinis, Field, Pierce, & Dalton, 2016; Hom Jr & Van Nuland, 2019; John, Loewenstein, & Prelec, 2012; Shrout & Rodgers, 2018). To fend against hindsight bias, researchers have recommended the endorsement of open-science best practices such as preregistration, Registered Reports, and openly sharing all predictions and decisions throughout the entire research lifecycle (Nosek et al., 2018; van't Veer & Giner-Sorolla, 2016).

Second, prospective hindsight bias may result in overestimating the robustness and the generalizability of an initial finding, believing that replications of a study would result in the same findings, and that replications are therefore of no value and a waste of resources. There are currently immense pressures for novelty in science, discouraging researchers from conducting replications (Nosek, Spies, & Motyl, 2012). Then, even if researchers do conduct a replication study, the combination of hindsight bias and confirmation bias (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) may lead researchers to analyze the data and interpret replication findings in a way that would favor initial findings, or feel pressured to do so by original authors, reviewers, editors, and other gatekeepers in the publication, promotion, and grant systems that perceive original findings as taken for granted or more authoritative. One way of addressing these problems is by encouraging direct close open replications by multiple third-party researchers (Brandt et al., 2014; Nosek et al., 2012; Nosek et al., 2018). Several mass open-science collaboration teams have been formed in the last decade to pursue this direction, such as the Psychological Science Accelerator (Moshontz et al., 2018), Collaborative Replications and Education Project (Wagge et al., 2019), and Many Labs (e.g., Ebersole et al., 2020; Klein et al., 2018).

However, the success of these initiatives depends on slow-to-change publication, granting, and promotion systems that may hinder these efforts. For example, grant authorities may be reluctant to fund, and reviewers and editors may be reluctant to publish, perceiving that this research question has already been addressed and therefore replications hold no contribution. This proposed impact of hindsight bias on the estimation of replication outcome and the evaluation of contribution of replication studies awaits empirical tests. Initial findings regarding journals conducting Registered Reports, publication accepted peer reviewed pre-registrations prior to data collection, both demonstrate these issues and show promise in addressing them (Chambers & Tzavella, 2020; Scheel, Schijen, & Lakens, 2021).

## 5. Current investigation: Two replications, extensions, and a new study

In this research, we conducted a close replication of hindsight bias in retrospective judgment (Study 1), a close replication of hindsight bias in prospective judgment (Study 2), and a study to examine possible hindsight bias regarding replicability of hindsight bias (Study 3).

We aimed to address mixed evidence regarding the magnitude and generalizability of hindsight bias. An early meta-analysis study conducted by Christensen-Szalanski and Willham (1991) on 122 studies on hindsight bias suggested a small effect size of $d = 0.35$, 95% confidence interval (CI) [0.28, 0.41] (sample-size corrected effect size $d = 0.52$, 95% CI [0.43, 0.61]). A more recent meta-analysis study based on 252 independent effect sizes revealed a similar sample-size-corrected effect of $d = 0.39$, 95% CI [0.36, 0.42] (Guilbault, Bryant, Brockway, & Posavac, 2004). In contrast to the two meta-analytical studies, the initial study of hindsight bias by Fischhoff (1975) suggested a much larger effect size ($d = 1.13$) for the supported contrasts between foresight and hindsight. A replication study of Fischhoff and colleagues' classic hindsight bias studies may help examine replicability of the effect using the same stimuli four and a half decades later, to provide an up-to-date estimate of the effect to aid researchers design follow-up studies (Simons, Holcombe, & Spellman, 2014).

We aimed to revisit and examine the replicability of these classic findings, following calls for a credibility revolution following what was coined a "replication/reproducibility crisis" in psychology (e.g., Klein et al., 2018; Open, 2015) and science overall (Camerer et al., 2016; Camerer et al., 2018; Gelman & Loken, 2013; Ioannidis, 2005). Datasets and code for the three studies were shared on: https://osf.io/nrwpv/.

### 5.1. Two pre-registered close replications

We chose Experiment 2 in Fischhoff (1975) as a target for replication for three reasons. First, this article is one of the first rigorous demonstrations of hindsight bias (Fischhoff, 2007; Hoch & Loewenstein, 1989). At the time of writing the article had 3073 citations according to Google Scholar. Second, the study was conducted in the 1970s and employed simplified statistics and reporting. By revisiting these classic methods and stimuli we aimed to refresh and update the methods and reporting

to meet current best practices in psychological science. To our knowledge and based on our communication with the author, this study is the first direct replication of the target experiment.

We chose Slovic and Fischhoff's (1977) Experiment 1 for replication for three key reasons. First, this experiment investigates prospective judgments, in which participants predict the probability of outcomes in future trials. In such judgments, hindsight bias is thought to have occurred if the forecast of the probability in future trials is affected by the outcome knowledge of the initial trial. The article received much attention, with 531 citations according to Google Scholar at the time of writing. Examining prospective judgments is important because hindsight bias may lead to biases in generalized evaluations of research and investigations based on initial, preliminary findings (Slovic & Fischhoff, 1977). By examining both retrospective judgments (in Study 1) and prospective judgments (in Study 2), we aimed to provide a more complete view of how outcome knowledge affects judgments and decision making.

Second, although Davis and Fischhoff's (2014) conducted a replication of the target experiment, we thought it worthwhile to conduct a pre-registered replication by an independent external research team of no direct relationship with the original authors. As suggested by various replication protocols (e.g., KNAW: Royal Dutch Academy of Arts and Sciences, 2018; Simons et al., 2014), independent replications by researchers from a different team can help reduce biases and increase credibility. Our study also enforced a pre-registration which was not included in Davis and Fischhoff (2014) and was conducted on a larger sample ($N = 608$ versus $N = 173$ after filtering the responses from 95 participants who failed the attention checks). Pre-registration is increasingly seen as important in limiting researchers' degrees of freedom and protecting against hindsight fallacy, as it helps reduce the possibility of consciously or unconsciously modifying beliefs about the hypotheses and planned ways of handling the data collection and analysis.

Overall, the two close replications answer calls for more pre-registered direct replication studies and open-science transparent reporting to increase the credibility and trustworthiness of published findings (Gelman & Loken, 2013; Munafò et al., 2017; Nosek & Lakens, 2014). Such efforts are particularly important in light of recent findings of lower-than-expected replicability rates of classic findings by mass pre-registered replications (Camerer et al., 2018; Klein, Hardwicke, et al., 2018; Open, 2015).

Both replication experiments were pre-registered on the Open Science Framework prior to data collection (Study 1: https://osf.io/5bfjg; Study 2: https://osf.io/75h98).

### 5.2. Extensions: Surprise and overconfidence

In addition, we added several extensions. Although the role of surprise and (over)confidence in hindsight bias seem widely accepted, our knowledge about their effects is in fact limited. First, the relationship between receiving outcome knowledge and surprise about the outcome needs further clarification. Some studies found that participants with outcome knowledge were less surprised by the outcome compared to those without outcome knowledge (e.g., Slovic & Fischhoff, 1977), whereas other studies found surprise as a moderator of hindsight bias (e.g., Ofir & Mazursky, 1997). Second, there are multiple ways of manipulating and measuring surprise (e.g., high/low probability of the outcome, warning/no warning about a stimulus, congruence/incongruence with outcome expectation) (see Ash, 2009; Nestler & Egloff, 2009; Ofir & Mazursky, 1997; Pezzo, 2003; Slovic & Fischhoff, 1977), yet these are often disjointed. For example, Pezzo (2003) manipulated surprise by outcome feedback that was either congruent or incongruent with participants' expectation, yet found that "regardless of whether outcomes were generally congruent or incongruent, people who found them to be still surprising after 5 minutes of thought showed less hindsight bias" (p. 430). Third, theoretical arguments in past research suggested that surprise and confidence may mediate and/or moderate the relationship between hindsight (vs. foresight) and probability estimates, yet past studies seldom explicitly and systematically tested these mechanisms.

We therefore proposed extensions regarding the roles of surprise and confidence. In Study 1, we tested the mediating and moderating roles of surprise. In Study 2, we tested the mediating and moderating roles of surprise, overconfidence, and task difficulty.

### 5.3. New study: Hindsight bias over replicability of hindsight bias

The purpose of the third study was to examine hindsight bias regarding the perceived replicability of hindsight bias. In our other replication work, we are often faced with reviewers who argued that our replication findings were not surprising, regardless of whether they were successful or not, and claiming that our replications added nothing new. Study 3 aimed to show the importance and generalizability of hindsight studies to directly address these issues by testing whether, ironically, hindsight bias replications may themselves be subject to hindsight bias.

In this study, we asked participants to contemplate the study design of Fischhoff's (1975) Experiment 2 and to then estimate the probabilities of a successful replication and of a failed replication. If hindsight bias holds, then participants who were informed of the outcome of the replication study would estimate the probability to be higher than participants who did not know the outcome and participants who were informed of the opposite outcome.

This study was pre-registered on the Open Science Framework prior to data collection (Study 3: https://osf.io/qyznw).

## 6. Study 1: Replicating Experiment 2 of Fischhoff (1975)

### 6.1. Target experiment and hypotheses

#### 6.1.1. Replication: Retrospective hindsight bias

In Experiment 2 of Fischhoff (1975), 172 students from an introductory statistics class in an Israeli university participated in the study (details available in the Supplementary Materials). Participants first read a passage describing an event, and were then asked to estimate the probabilities of four possible outcomes for the event. Participants were randomly assigned to two types of conditions: those in the *Before condition* did not have any outcome knowledge (i.e., they did not know which of the four outcomes actually occurred), whereas those from the *After conditions* were given the outcome knowledge but were asked to estimate as if they had not known the outcome. Because for each event, there were four possible outcomes, there were four After conditions, with each condition stating that one of the presented outcomes had actually occurred. Despite being asked to ignore their knowledge of the outcome, participants in the After conditions estimated a higher probability for the outcome to which they were told has occurred, demonstrating hindsight bias.

We made the following prediction for the replication study of Experiment 2 of Fischhoff (1975):

*H1: Probability estimates (hindsight bias). Compared with participants in the Before condition, participants in the After conditions estimate a higher probability of the outcome that they knew had occurred.*

#### 6.1.2. Extension: Surprise

We proposed extension hypotheses regarding the processes leading to hindsight bias. Feelings of surprise signal the difficulty of generating alternatives to the outcome, increase the need to scrutinize the cognitive process, and deepen the extent of sense making after receiving outcome knowledge (Bernstein et al., 2016; Pezzo, 2003; Sanna & Schwarz, 2006). Our literature review suggested that surprise could play one or both of two roles in hindsight bias. The first role is an indicator or an accompanying outcome of hindsight bias. An implicit and untested inference of this line of reasoning is that surprise is an intermediate

outcome in the cognitive processes leading to hindsight bias. For example, Slovic and Fischhoff (1977) suggested that hindsight bias occurred when outcome knowledge led individiauls to feel less surprised and biased their probability estimates toward the known outcome. The second role is a required condition that shapes the magnitude of hindsight bias, or a moderator of hindsight bias. For example, Sanna and Schwarz (2006) argued that hindsight bias occurs when individuals feel the outcome is unsurprising, and it could reverse when individuals feel the outcome is surprising (i.e., the "I never would have known it" effect or the "backfire effect"; Hawkins & Hastie, 1990; Hoch & Loewenstein, 1989). Some models considered both roles of surprise simultaneously. For example, Pezzo's (2003) sense-making model suggested that a surprising outcome is required to trigger sense-making activities (surprise as a moderator); while the person might experience some initial surprise (surprise as a mediator), successful sense-making activities lead to hindsight bias and reduce end-state surprise feelings (surprise as an accompanying outcome).

We therefore tested three effects of surprise: as an outcome of experimental condition, as a mediator of the effect of experimental condition on probability estimates, and as a moderator of the effect of experimental condition on probability estimates.[2] In order to test these effects, we asked participants to report their feelings of surprise about the outcome. We proposed that:

*H2: Surprise ratings (extension).*

*H2a: Compared with participants in the Before condition, participants in the After conditions report lower levels of surprise regarding the outcome for which they knew had occurred.*

*H2b: Surprise mediates the relationship between outcome knowledge and probability estimates. (exploratory).*

*H2c: Surprise moderates the relationship between outcome knowledge and probability estimates, such that hindsight bias is stronger in the low-surprise group than in the high-surprise group. (exploratory).*

### 6.2. Method

#### 6.2.1. Power analysis

The planned sample size for the replication study was calculated based on an effect size of $d = 1.13$, 95% CI [0.44, 1.82] for a single before-after contrast, estimated from the target experiment (see Supplementary Materials for details). We conducted a power analysis using G-Power (Faul, Erdfelder, Buchner, & Lang, 2009). In order to achieve a statistical power of 95% with an alpha of 0.05 (two-tailed), a sample size of 46 per comparison would be required. Because the study adopted a between-subject design (4 events with 4 possible outcomes each), we approximated a total sample size of 46 * 4 * 4 = 736. In consultation with the original author and the editor, we removed the stimuli and results relating to Events C and D. We therefore updated this analysis posthoc to indicate a total required sample size of 368.

#### 6.2.2. Participants

A total of 442 American participants were recruited from Amazon Mechanical Turk online through CloudResearch (Litman, Robinson, & Abberbock, 2017) (245 females, 196 males, 1 undisclosed, $M_{age} = 39.78$, $SD_{age} = 11.46$, see Supplementary Materials for details about sample characteristics; descriptives in this section were updated to reflect the exclusion of data collection for Events C and D, explained below).

#### 6.2.3. Procedure and materials

The materials used in this replication study were obtained from the author of the target experiment (see Supplementary Materials). There were four events: Event A, the British-Gurka struggle; Event B, the near-riot in Atlanta; Events C: Mrs. Dewar in therapy; and Event D: George in therapy. We note that in consultation with the original author and the editor we removed the descriptions of the stimuli of Events C and D, and related findings. We jointly strongly believe that these stimuli should no longer be used in future research.

Events A and B were each described in a passage ranged from 185 to 235 words in length, followed by four possible outcomes. For example, Event A described a war between the British and the Gurkas in South Asia in 1814. The four possible outcomes were: (1) British resulted in victory; (2) Gurka resulted in victory; (3) The two sides reached a military stalemate, but were unable to come to a peace settlement; (4) The two sides reached a military stalemate and came to a peace settlement.

This study used a between-subject design. Participants were randomly assigned to one of five experimental conditions: one Before condition and four After conditions (each associated with one informed outcome). Each participant was presented with one of the two events used in the target experiment. That is, participants were exposed to one of the 5 (condition) x 2 (event) possibilities. Participants in the Before condition read the assigned passage alone, whereas participants in the After conditions read the assigned passage followed by a sentence which provided the outcome knowledge (e.g., Outcome: British resulted in victory).

Participants were then asked a comprehension question, "*To make sure you read and understood the scenario, please answer the following comprehension question: What was the outcome of the event?*". In order to proceed to the next stage of the experiment, participants in the Before condition had to choose "The case did not indicate the outcome," whereas participants in the After conditions had to choose the informed outcome.

*6.2.3.1. Probability estimates.* Participants were asked to provide probability estimates for each of the four possible outcomes of the event. For the Before condition, the question read, "In light of the information appearing in the passage, please estimate the probability of occurrence of each of the four possible outcomes listed below. There are no right or wrong answers, answer based on your intuition. (The probabilities should sum to 100%)". For the After conditions, in addition to the sentences above, participants also read "Answer as if you do not know the outcome, estimating the case at that time before outcomes were known."

*6.2.3.2. Surprise ratings.* Following the probability estimates, participants were asked to rate their levels of surprise (i.e., "How surprised would you be if the outcome was that the (outcome)?") on a 7-point Likert scale (1 = *Not surprised at all*, 7 = *Very surprised*). Participants in the Before condition were asked to rate their surprise levels regarding all four possible outcomes; participants in After conditions were only asked to rate their surprise levels regarding the informed outcome.

#### 6.2.4. Replication evaluation: Very close replication

Our replication study is a very close replication based on the criteria proposed in LeBel, Berger, Campbell, and Loving (2017) and LeBel, McCarthy, Earp, Elson, and Vanpaemel (2018). According to LeBel and colleagues' taxonomy, a very close replication shares the same independent variable (IV) operationalization, dependent variable (DV) operationalization, IV stimuli, and DV stimuli with the original study; only the procedural details, physical setting, and contextual variables (e.g., linguistic or cultural adaptations) differ from the original study. Similarly, Brandt et al. (2014, p. 218) wrote that "close replications refer to those replications that are based on methods and procedures as close as possible to the original study … ideally the only differences between the two are the inevitable ones (e.g., different participants…)." In Study 1, the IV operationalization, DV operationalization, IV stimuli, and DV stimuli were all the same as those used in the original study, with a few necessary adjustments to improve on the design or to accommodate

[2] A variable can be both a mediator and a moderator of a relationship (James & Brett, 1984; Judd, Kenny & McClelland, 2001; Karazsia & Berlin, 2018). Such relationships have been tested in previous studies (e.g., Connor-Smith & Compas, 2002; Wei, Mallinckrodt, Russell & Abraham, 2004; Zhou, Wang, Chen & Shi, 2012)

**Table 1**
Study 1: Classification of the replication, based on LeBel et al. (2018).

| Design facet | Replication | Details of deviation |
|---|---|---|
| IV operationalization | Same | |
| DV operationalization | Same | |
| IV stimuli | Same | • Changed the word "Negro" into "African American" in the passage of Event A |
| DV stimuli | Same | • Added surprise measure after the replication. |
| Procedural details | Similar | • Used a larger sample size: Original study: 172; Replication study: 890<br>• Added one comprehension question for each scenario.<br>• Added funnel questions at the end of the study. |
| Physical settings | Different | • Changed from offline data collection (participants were students from Hebrew University and the University of the Negev) to online data collection (participants were recruited from CloudResearch). |
| Contextual variables | Similar | |
| Replication classification | Very close replication | |

*Note.* IV = Independent variable, DV = dependent variable.

contextual requirements. See Table 1 for a summary of classification, necessary adjustments, and theoretical extensions.

### 6.3. Results

#### 6.3.1. Replication: Probability estimates

We summarized the descriptives of the probability estimates in Table 2. Violin plots of the probability estimates are available in Supplementary Materials. The numbers of interest are the probability estimate of an outcome in the Before condition, and probability estimate of that same outcome in the After condition in which this outcome was informed to have occurred (numbers marked in bold).

Because there are two events with four outcomes each, we conducted 8 sets of Mann-Whitney *U* tests. As shown in Table 3, in 7 of the 8 sets of comparison (except Event A-Outcome 2), the mean probability estimates in the After condition were higher than those in the Before condition. The results remained largely the same when we adjusted the *p* values using the Benjamini and Hochberg (1995) false discovery rate control method.

Historically, the correct outcomes of Events A and B were Outcome 1, yet the mean probability estimates of these two outcomes in the Before condition were not higher than chance (21.40% and 7.46%, respectively). Specifically, the probability estimate for Outcome 1 (British

**Table 2**
Study 1: Means and standard deviations of probability estimates.

| Experimental Condition | Sample Size | Outcome Informed | Outcome Evaluated | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Outcome 1 | | Outcome 2 | | Outcome 3 | | Outcome 4 | |
| | | | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| Event A: British-Gurka struggle | | | | | | | | | | |
| Before | 43 | None | **21.40** | 18.17 | **38.61** | 26.60 | **23.49** | 19.93 | **16.51** | 15.53 |
| After | 45 | Outcome 1 | **45.51** | 28.59 | 21.18 | 19.45 | 19.69 | 16.25 | 13.62 | 11.46 |
| | 42 | Outcome 2 | 26.05 | 20.35 | **43.62** | 23.62 | 18.48 | 18.66 | 11.86 | 9.52 |
| | 44 | Outcome 3 | 21.93 | 17.13 | 23.18 | 16.14 | **31.59** | 19.61 | 23.30 | 14.10 |
| | 43 | Outcome 4 | 25.49 | 17.84 | 28.40 | 22.72 | 18.72 | 15.97 | **27.40** | 23.98 |
| Event B: Near riot in Atlanta | | | | | | | | | | |
| Before | 46 | None | **7.46** | 9.25 | **25.91** | 23.88 | **12.91** | 18.43 | **53.72** | 26.66 |
| After | 46 | Outcome 1 | **25.44** | 23.11 | 22.63 | 17.58 | 22.28 | 21.88 | 29.65 | 18.76 |
| | 44 | Outcome 2 | 11.61 | 12.50 | **50.02** | 29.13 | 9.52 | 10.34 | 28.84 | 22.18 |
| | 44 | Outcome 3 | 15.23 | 13.64 | 17.50 | 12.60 | **29.77** | 28.53 | 37.50 | 24.53 |
| | 45 | Outcome 4 | 9.87 | 12.18 | 12.98 | 12.24 | 11.20 | 16.82 | **65.96** | 27.76 |

*Note:* The bolded numbers indicate the key sets of comparison of interest (i.e., the Before and After probability estimates of the same outcome). The foresight ratings of all four outcomes came from the same participants in the foresight condition. The hindsight ratings of the four outcomes came from participants in the four hindsight conditions, respectively. Following a discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

**Table 3**
Study 1: Mann-Whitney U tests of probability estimates difference between before and after conditions.

| After - Before | *Mean Difference (Rank)* | | | | | | | 95% CI for $\phi$ | | | 95% CI for *d* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *U* | *z* | *p* | $p_{adjusted}$ | *r* | $\phi$ | LL | UL | *d* | LL | UL |
| Event A Outcome 1 | 23.0 | 462 | 4.24 | <0.001 | <0.001 | 0.45 | 0.76 | 0.65 | 0.84 | 1.00 | 0.53 | 1.46 |
| Event A Outcome 2 | 5.8 | 780 | 1.09 | 0.277 | 0.277 | 0.12 | 0.57 | 0.45 | 0.68 | 0.20 | −0.23 | 0.63 |
| Event A Outcome 3 | 11.5 | 695 | 2.15 | 0.032 | 0.043 | 0.23 | 0.63 | 0.51 | 0.74 | 0.41 | −0.02 | 0.84 |
| Event A Outcome 4 | 14.0 | 624.5 | 2.62 | 0.009 | 0.014 | 0.28 | 0.66 | 0.54 | 0.76 | 0.54 | 0.10 | 0.97 |
| Event B Outcome 1 | 26.7 | 444 | 4.87 | <0.001 | <0.001 | 0.51 | 0.79 | 0.68 | 0.87 | 1.02 | 0.56 | 1.48 |
| Event B Outcome 2 | 24.6 | 459.5 | 4.48 | <0.001 | <0.001 | 0.47 | 0.77 | 0.66 | 0.85 | 0.91 | 0.45 | 1.36 |
| Event B Outcome 3 | 20.9 | 543 | 3.82 | <0.001 | <0.001 | 0.40 | 0.73 | 0.62 | 0.82 | 0.71 | 0.26 | 1.14 |
| Event B Outcome 4 | 11.3 | 778.5 | 2.05 | 0.041 | 0.047 | 0.21 | 0.62 | 0.50 | 0.73 | 0.45 | 0.03 | 0.87 |

*Note*. We calculated three effect sizes of the Mann-Whitney *U* tests, which are *r* (the correlation between being in the hindsight condition and winning in the rank comparison with the other condition, see Fritz, Morris, & Richler, 2012), $\phi$ (the probability that a score in the hindsight condition was higher than that in the foresight condition, see Fay & Malinovsky, 2018), and Cohen's *d* (the standard difference in the mean ranking between the hindsight condition and the foresight condition, assuming that the rankings follow a normal distribution, see Cohen, 1988). *p* values were adjusted using the Benjamini and Hochberg (1995) false discovery rate control method. Following a discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

**Table 4**
Study 1 Extension: Means and standard deviations of surprise ratings.

| Experimental Condition | Outcome Evaluated | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Outcome 1 | | | Outcome 2 | | | Outcome 3 | | | Outcome 4 | | |
| | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* |
| Event A: British-Gurka struggle | | | | | | | | | | | | |
| Before | 43 | 4.35 | 2.14 | 43 | 3.95 | 2.16 | 43 | 3.42 | 1.76 | 43 | 4.53 | 1.84 |
| After | 45 | 3.20 | 2.00 | 42 | 4.10 | 1.88 | 44 | 3.41 | 1.76 | 43 | 4.60 | 1.55 |
| Event B: Near-riot in Atlanta | | | | | | | | | | | | |
| Before | 46 | 5.89 | 1.55 | 46 | 2.78 | 1.55 | 46 | 5.46 | 1.57 | 46 | 1.96 | 1.38 |
| After | 46 | 5.17 | 1.70 | 44 | 2.91 | 1.65 | 44 | 5.36 | 1.94 | 45 | 1.91 | 1.44 |

*Note.* The foresight ratings of all four outcomes came from the same participants in the foresight condition. The hindsight ratings of the four outcomes came from participants in the four hindsight conditions, respectively. Hindsight participants only rated their surprise over the outcome which they knew had occurred. Following a discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

resulted in victory) in Event A (Before condition) was not significantly different from chance (one-sample *t*-test: $t = -1.30$, $df = 42$, $p = .200$, $d = -0.20$). The probability estimate for Outcome 1 (dispersion and no outbreak of violence) in Event B (Before condition) was the lowest among those for all four outcomes, and it was significantly smaller than chance (one-sample t-test: $t = -12.87$, $df = 45$, $p = .000$, $d = -1.90$). These suggest that the participants did not have much knowledge about the historical background of these two events, relieving the concern that prior knowledge gained before participating in this study impacted participants' reactions to these two experimental stimuli. Importantly, as Event B is the only event that is linked to the American history, the findings address the concern that using an American sample (versus the Israeli sample used in the original study) reduced the task difficulty of this question or impacted the magnitude of hindsight bias.

Because Mann-Whitney *U* tests are nonparametric, we calculated three effect sizes: (1) *r*, the correlation between experimental group membership and whether the rank is higher or lower than the other group (see Fritz et al., 2012), (2) $\phi$, the probabilistic index reflecting the likelihood that the score in one group is smaller than or equal to that of the other group, estimated using the receiver operating characteristic curve under the proportional odds assumption (see Fay & Malinovsky, 2018), and (3) Cohen's *d*, the standard difference between the mean rankings of the two groups, assuming that the rankings in the two groups follow a normal distribution (Cohen, 1988).

As shown in Table 3, the correlations *r*s between being in the hindsight condition and winning in the rank comparison with the other condition were all positive. The sizes of correlations were mostly medium to large (Cohen, 1988). The effect sizes $\phi$s, reflecting the probability that a score in the hindsight condition was higher than that in the foresight condition, did not include 0.50 in all but one set of comparison (i.e., Event A-Outcome 2). However, when we calculated the Cohen's *d*s under the assumption of a normal distribution of the rankings, two comparisons had confidence intervals that overlapped with the null (i.e., Event A-Outcome 2, Event A-Outcome 3). The Cohen's *d* effects were mostly medium to large.

#### 6.3.2. Robustness checks: Alternative tests and exclusion criteria

To examine the robustness of the findings, we conducted additional analyses on the probability estimates (see Supplementary Materials). Results of Student's independent samples *t*-tests of probability estimates were largely consistent with the results of the Mann-Whitney *U* tests. When we analyzed the data with only participants who met a set of pre-registered criteria (i.e., understood the English used in the study, was serious in the study, and did not correctly guess the purpose of the study), the results regarding the probability estimates remained mostly the same. We concluded robust support for Hypothesis 1.

#### 6.3.3. Extension: Surprise ratings

We detailed the descriptives of the surprise ratings in Table 4. Violin plots of the surprise ratings are available in Supplementary Materials.

Similar to previous analyses with probability estimates, we conducted 8 sets of Mann-Whitney *U* tests to compare the differences in surprise ratings between the Before condition and the After conditions. As shown in Table 5, a total of two sets of comparisons were significant, based on *p* value and the confidence interval of $\phi$. Specifically, for Event A Outcome 1 and Event B Outcome 1, surprise ratings in the After condition were significantly lower than those in the Before condition, and the effect sizes were small to medium. The results of the other three sets of comparison (Event C-Outcome 2, Event C-Outcome 4, Event D-Outcome 2) were in the opposite direction of our prediction, with the surprise ratings in the After condition being higher than those in the Before condition (small to medium effect sizes). When we adjusted the *p* values using the Benjamini and Hochberg (1995) false discovery rate control method, none of the Mann-Whitney *U* tests remained significant. Results of Student's independent samples t-tests of surprise ratings (see Supplementary Materials) were largely consistent with the results of the Mann-Whitney *U* tests. Overall, the results provided little to no support for Hypothesis 2(a) regarding surprise ratings.

We found no support for exploratory Hypotheses 2 that surprise acted as a mediator of the relationship between outcome knowledge and probability estimates. We found mixed support for exploratory

**Table 5**
Study 1: Extension: Mann-Whitney U tests of differences in surprise between Before and After conditions.

| After - Before | *Mean Difference (Rank)* | | | | | | | 95% CI for $\phi$ | | | 95% CI for *d* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *U* | *z* | *p* | $p_{adjusted}$ | *r* | $\phi$ | Lower | Upper | *d* | Lower | Upper |
| Event A Outcome 1 | −13.76 | 665 | −2.56 | 0.011 | 0.044 | −0.27 | 0.34 | 0.24 | 0.46 | −0.56 | −0.99 | −0.12 |
| Event A Outcome 2 | 1.67 | 867.5 | 0.32 | 0.752 | 0.897 | 0.03 | 0.52 | 0.40 | 0.64 | 0.07 | −0.36 | 0.50 |
| Event A Outcome 3 | −0.69 | 931 | −0.13 | 0.897 | 0.897 | −0.01 | 0.49 | 0.38 | 0.61 | −0.01 | −0.43 | 0.42 |
| Event A Outcome 4 | 1.86 | 884.5 | 0.35 | 0.725 | 0.897 | 0.04 | 0.52 | 0.40 | 0.64 | 0.04 | −0.38 | 0.46 |
| Event B Outcome 1 | −13.80 | 740.5 | −2.57 | 0.010 | 0.044 | −0.27 | 0.35 | 0.25 | 0.46 | −0.44 | −0.86 | −0.02 |
| Event B Outcome 2 | 1.40 | 980.5 | 0.26 | 0.795 | 0.897 | 0.03 | 0.52 | 0.40 | 0.63 | 0.08 | −0.34 | 0.49 |
| Event B Outcome 3 | 1.91 | 969 | 0.36 | 0.719 | 0.897 | 0.04 | 0.52 | 0.41 | 0.63 | −0.05 | −0.47 | 0.36 |
| Event B Outcome 4 | −1.03 | 1011.5 | −0.21 | 0.833 | 0.897 | −0.02 | 0.49 | 0.39 | 0.59 | −0.03 | −0.44 | 0.38 |

*Note. p* values were adjusted using the Benjamini and Hochberg (1995) false discovery rate control method. Following a discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

**Table 6**
Study 1: Comparison of results of the original study and the replication study.

| | Cohen's *d* [95% CI] | *p*-value | Note |
|---|---|---|---|
| Fischhoff (1975) | 1.13 [0.44, 1.82] | <0.001 | |
| Replication | | | |
| Event A Outcome 1 | 1.00 [0.53, 1.46] | <0.001 | Signal – consistent |
| Event A Outcome 2 | 0.20 [−0.23, 0.63] | 0.277 | No signal – inconsistent, smaller |
| Event A Outcome 3 | 0.41 [−0.02, 0.84] | 0.032 | No signal – inconsistent, smaller |
| Event A Outcome 4 | 0.54 [0.10, 0.97] | 0.009 | Signal – inconsistent, smaller |
| Event B Outcome 1 | 1.02 [0.56, 1.48] | <0.001 | Signal – consistent |
| Event B Outcome 2 | 0.91 [0.45, 1.36] | <0.001 | Signal – consistent |
| Event B Outcome 3 | 0.71 [0.26, 1.14] | <0.001 | Signal – consistent |
| Event B Outcome 4 | 0.45 [0.03, 0.87] | 0.041 | Signal – inconsistent, smaller |

*Note*: Following a discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.
According to LeBel et al. (2019), there is a signal if the confidence interval of the replication effect size excludes zero, and the replication result is considered consistent with the original study if the confidence interval of the replication effect size includes the effect size of the original study.

Hypothesis 2c that surprise acted as a moderator, such that the relationship between outcome knowledge and probability estimates was stronger when surprise was lower rather than higher. However, in our original analysis when all four events were included, we did not find support for the moderating effect of surprise. While we have decided to remove results related to Events C and D, which is a deliberate deviation from the preregistration, we caution our readers about the conflicting findings of the moderating effect of surprise in Study 1 when different events were included in the analysis. We provided all related details and analyses in the Supplementary Materials.

### 6.4. Discussion

We aimed to replicate Fischhoff (1975)'s Experiment 2, a classic study of hindsight bias. Following the original study, we hypothesized that participants provided with outcome knowledge would estimate a greater probability for the outcome which they knew had occurred, compared to participants without outcome knowledge. This hypothesis was supported in 7 of the 8 sets of comparison of probability estimates, and the effect sizes were mostly medium to large. Once participants were informed of the outcome, they perceived the outcome to be more probable, even if they were asked to ignore the outcome, demonstrating hindsight bias. These findings therefore support the idea that participants were either unaware of or unable to resist the influence of outcome knowledge.

#### 6.4.1. Evaluation of replication findings: Mostly successful replication

In Table 6 we compared the results of the target experiment and the replication study using the criteria described in LeBel, Vanpaemel, Cheung, and Campbell (2019). All the 8 sets of comparison of probability estimates were in the same direction as in the original study. The replication effects were medium to large, though slightly smaller than those found in the original study. In 4 of the 8 sets of probability estimates comparisons, the confidence intervals of the effect sizes (Cohen's *d*s) of the replication study included $d = 1.13$, which is the effect size estimated from the target experiment. In Fig. 1 we provided a forest plot

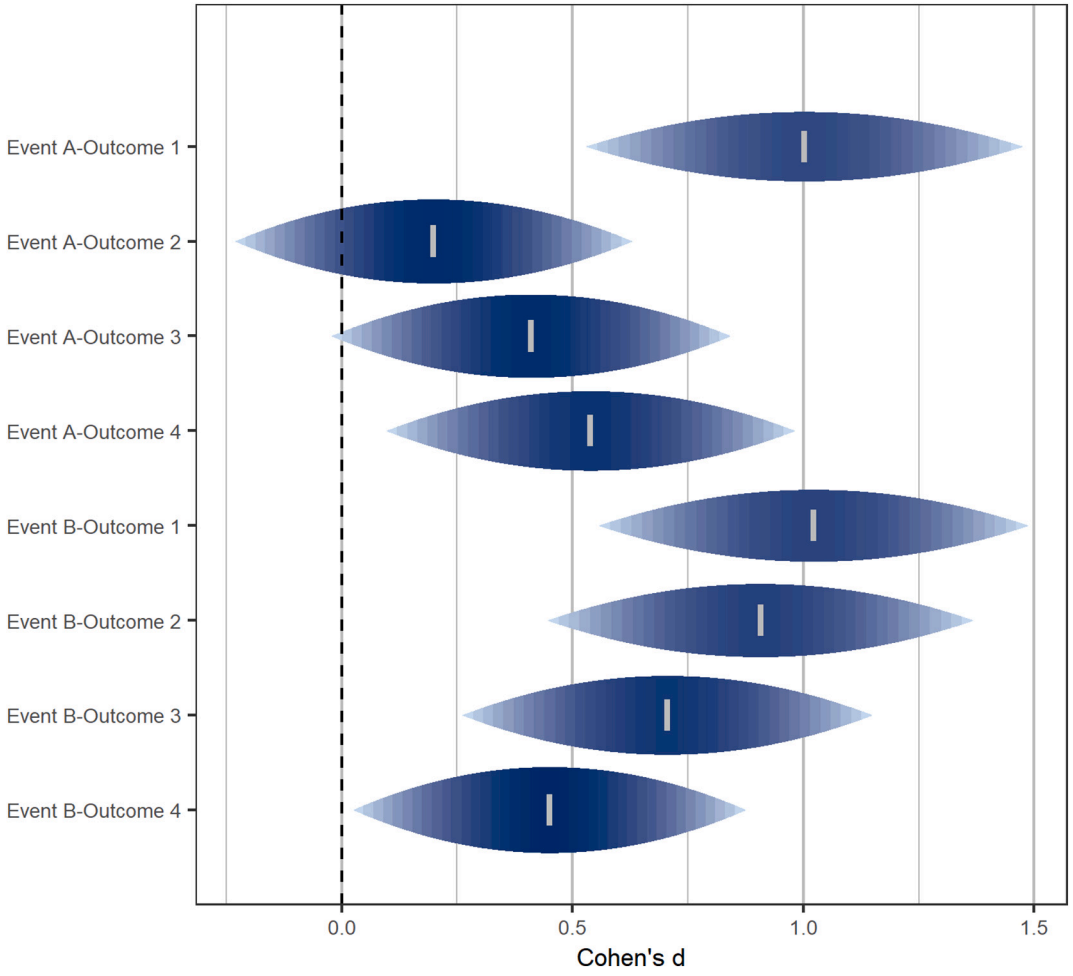



**Fig. 1.** Study 1: forest plot for probability estimates.

of the probability estimates contrasts. Overall, we conclude this replication of hindsight bias as successful.

#### 6.4.2. Extension: Surprise ratings

Beyond the replication, we extended the experiment by investigating an intuitive yet understudied dependent variable, the level of surprise associated with the known outcome. Judging from null hypothesis significance testing (NHST), effect sizes, and confidence intervals, 2 of the 8 sets of surprise ratings comparisons were significant in the predicted direction.

Contrary to our expectations, we found no support for surprise as a mediator in the relationship between outcome knowledge and probability estimates. Additional analyses showed that surprise ratings and probability estimates were indeed negatively correlated, both in the Before condition and in the After conditions (see Supplementary Materials). These results suggest that the negative correlation between surprise ratings and probability estimates may be caused by factors other than hindsight bias. Also, we found inconclusive findings for the exploratory hypothesis that surprise acted as a moderator of the relationship between outcome knowledge and probability estimates.

## 7. Study 2: Replicating experiment 1 of Slovic and Fischhoff (1977)

### 7.1. Target experiment and hypotheses

#### 7.1.1. Replication: Prospective hindsight bias

In Experiment 1 of Slovic and Fischhoff (1977), 184 American participants were recruited via university newspaper. All participants read four vignettes about scientific research. For each vignette, participants in the *foresight condition* read that two outcomes were possible in the first trial, whereas participants in the *hindsight condition* read that the first trial had been conducted and one of the two outcomes had occurred. They were then asked why they thought the outcome(s) might occur, and then predicted the probability that the previously observed outcome would repeat in future research trials. The results suggested a sense of inevitability of the disclosed outcome among hindsight participants: their predicted probabilities of the previously observed outcome to repeat were higher than those of participants in the foresight condition ($d = 0.36$). Davis and Fischhoff (2014) replicated this experiment, which produced similar effects (overall effect: 0.27–0.33, $d = 0.20$ to 0.44) that the disclosed outcome of the initial trial was perceived to be more likely to occur in future trials in hindsight than in foresight.

We extended the original design and tested exploratory analyses regarding the mechanisms underlying hindsight bias, using a different set of materials and decisions (i.e., prospective judgments). In addition to surprise, we asked participants to report their levels of confidence about the accuracy of their own judgments. To better understand if the nature of the task would have an impact on hindsight bias, we also measured participants' overall levels of perceived difficulty of the prediction task.

We followed Experiment 1 in Slovic and Fischhoff (1977) to predict that hindsight bias would be observed in prospective judgments. Individuals often use past information to form judgments about the future (Aarts, Verplanken, & Van Knippenberg, 1998; Ouellette & Wood, 1998). If individuals' beliefs about past events changed due to outcome knowledge, then those changed beliefs may trigger hindsight bias when people use them to make prospective judgments. In addition, knowing the outcome of the initial trial may increase the perceived inevitability of the outcome, which will increase the expectation that the outcome will repeatedly occur in the future. Therefore, we predicted:

*H3: Participants in the hindsight condition estimate a greater probability that the outcome will continue to occur in future trials, compared with participants in the foresight condition.*

### 7.2. Extension: Surprise, confidence, and task difficulty

For the extension hypotheses, we first examined the effects of surprise and confidence. By surprise, we refer to individuals' feelings of surprise if a particular outcome would occur in future trials (Slovic & Fischhoff, 1977). By confidence, we refer to individuals' feelings of confidence about the accuracy of their own judgments (Granhag, Strömwall, & Allwood, 2000). We chose to study these two factors because these have been suggested as mechanisms that affect hindsight bias: beliefs about events' objective likelihoods, and beliefs about one's own prediction ability subjectively (Roese & Vohs, 2012).

As in Study 1, we hypothesized that surprise ratings are lower among participants in the hindsight condition than those in the foresight condition. We also tested the hypothesis that surprise mediates or moderates the relationship between hindsight condition and probability estimates as in Study 1.

*H4: Surprise ratings (extension).*

*(H4a) Participants in the hindsight conditions report lower levels of surprise regarding the outcome for which they knew had initially occurred compared with participants in the foresight condition.*

*(H4b) Surprise mediates the relationship between the hindsight condition and probability estimates. (exploratory)*

*(H4c) Surprise moderates the relationship between hindsight condition and probability estimates, such that hindsight bias is stronger in the low-surprise group than in the high-surprise group. (exploratory)*

Like surprise, past research has also theorized and examined multiple roles that confidence can play in hindsight bias. For example, overconfidence is often proposed as a consequence of outcome knowledge (Davis & Fischhoff, 2014; Slovic, Lichtenstein, & Fischhoff, 1988). Other studies examined the moderating role of confidence in hindsight bias. For example, Arkes, Wortmann, Saville, and Harkness (1981) found that a procedure to reduce overconfidence by asking for reasons for each possible outcome reduced hindsight bias. Also, Werth and Strack (2003) found that the magnitude of hindsight bias was contingent on the feeling of confidence, which served as a signal of whether the individual would have known the answer or not. They found that participants who experienced higher confidence showed greater hindsight bias than participants who experienced lower confidence.

Therefore, we hypothesized that participants in the hindsight condition will report greater confidence about the accuracy of their estimation than participants in the foresight condition. Furthermore, like surprise, we examined whether confidence mediates or moderates the relationship between hindsight condition and probability estimates.

*H5: Confidence ratings (extension).*

*(H5a) In prospective judgments, compared with participants in the foresight condition, participants in the hindsight conditions report higher levels of confidence about the accuracy of their judgments.*

*(H5b) Confidence mediates the relationship between hindsight condition and probability estimates. (exploratory)*

*(H5c) Confidence moderates the relationship between hindsight condition and probability estimates, such that hindsight bias is stronger in the high-confidence group than in the low-confidence group. (exploratory)*

To examine the effect of the characteristics of the task, we also measured the extent to which participants perceived the task to be difficult. We expected that participants in the hindsight condition will report lower levels of task difficulty than participants in the foresight condition. This is because the foresight condition could dilute participants' attention by asking them to consider two outcomes simultaneously, whereas the hindsight condition could cue participants to ignore the outcome that did not occur in the initial trial (Slovic & Fischhoff, 1977). Lower levels of perceived task difficulty, in turn, may contribute to hindsight bias, as the subjective difficulty to generate alternative outcomes can be taken as an indication that those outcomes are implausible (Harley et al., 2004; Roese & Vohs, 2012; Sanna & Schwarz, 2006). We therefore tested the following:

*H6: Task difficulty (exploratory extension).*

**Table 7**
Study 2: Questions asked in the virgin rat scenario.

| Foresight condition | Hindsight outcome A condition | Hindsight outcome B condition |
|---|---|---|
| 1. Try and estimate, what are the probabilities of the following outcomes (these probabilities should total 100%)<br>Virgin rat **will exhibit** maternal behavior: _____<br>Virgin rat **will NOT exhibit** maternal behavior: _____<br>Total: _____<br>2. If the virgin rat **does exhibit** maternal behavior, what is the probability that in a replication of this experiment with 10 additional virgin female rats (these probabilities should total 100%)<br>a. All will exhibit maternal behavior?: _____<br>b. Some will exhibit maternal behavior?: _____<br>c. None will exhibit maternal behavior?: _____<br>Total: _____<br>*3. If the virgin rat **does exhibit** maternal behavior, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised*<br>4. If the virgin rat **does NOT exhibit** maternal behavior, what is the probability that in a replication of this experiment with 10 additional virgin female rats (these probabilities should total 100%)<br>a. All will exhibit maternal behavior?: _____<br>b. Some will exhibit maternal behavior?: _____<br>c. None will exhibit maternal behavior?: _____<br>Total: _____<br>*5. If the virgin rat **does NOT exhibit** maternal behavior, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised*<br>*6. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the Virgin Rat experiment? 0 = Extremely not confident … 6 = Extremely confident* | Outcome: The initial virgin rat **exhibited** maternal behavior in the first trial.<br>1. What is the probability that in a replication of this experiment with 10 additional virgin female rats (these probabilities should total 100%)<br>a. All will exhibit maternal behavior?: _____<br>b. Some will exhibit maternal behavior?: _____<br>c. None will exhibit maternal behavior?: _____<br>Total: _____<br>*2. Do you think the finding that the virgin rat **exhibited** maternal behavior is surprising? 1 = Not surprising at all … 5 = Extremely surprising*<br>*3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the Virgin Rat experiment? 0 = Extremely not confident … 6 = Extremely confident* | Outcome: The initial virgin rat **did NOT exhibit** maternal behavior in the first trial.<br>1. What is the probability that in a replication of this experiment with 10 additional virgin female rats (these probabilities should total 100%)<br>a. All will exhibit maternal behavior?: _____<br>b. Some will exhibit maternal behavior?: _____<br>c. None will exhibit maternal behavior?: _____<br>Total: _____<br>*2. Do you think the finding that the virgin rat **did not exhibit** maternal behavior is surprising? 1 = Not surprising at all … 5 = Extremely surprising*<br>*3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the Virgin Rat experiment? 0 = Extremely not confident … 6 = Extremely confident* |
| **For all three conditions, after reading all four scenarios**<br>*How difficult was it to make estimations of outcomes probabilities? 1 = Extremely easy … 7 = Extremely difficult* | | |

*Note.* Questions italicized in the table are the extension questions; they were not italicized in the Qualtrics survey.

**Table 8**
Study 2: Classification of the Replication, based on LeBel et al. (2018)

| Design facet | Replication | Details of deviation |
|---|---|---|
| IV operationalization | Same | |
| DV operationalization | Same | |
| IV stimuli | Same | • Changed outcome B in the Y-Test scenario from "*Places in Area B*" to "*Places in Area C*," so that outcome A and outcome B were symmetric. |
| DV stimuli | Similar | • Removed reasons for why the outcome had occurred.<br>• Added surprise, confidence, and task difficulty measures. |
| Procedural details | Similar | • Used a larger sample size: Original study: 184 (sample size per group varied from 24 to 37); Replication study: 604 (197 hindsight, 204 foresight outcome A, 203 foresight outcome B)<br>• Added one comprehension question for each scenario.<br>• Added funnel questions at the end of the study. |
| Physical settings | Different | • Changed from offline data collection (participants were recruited via a student newspaper at the University of Oregon) to online data collection (participants were recruited from CloudResearch). |
| Contextual variables | Different | |
| Replication classification | Very close replication | |

*Note.* IV = Independent variable, DV = dependent variable.

*(H6a) In prospective judgments, compared with participants in the foresight condition, participants in the hindsight condition report lower levels of task difficulty.*

*(H6b) Task difficulty mediates the relationship between hindsight condition and probability estimates.*

*(H6c) Task difficulty moderates the relationship between hindsight condition and probability estimates, such that hindsight bias is stronger among those who perceive the task to be easy than among those who perceive the task to be difficult.*

### 7.3. Method

#### 7.3.1. Power analysis

The planned sample size for the replication study was estimated from the target experiment (see Supplementary Materials for details). We estimated the effect sizes based on *p* values, because they were the only statistics available from the target experiment. The *p* values of pairwise comparisons ranged from 0.001, 0.01, to 0.05. We chose $p = .05$, which lead to $d = 0.36$, 95% CI [0.00, 0.72]. We conducted a power analysis using G-Power (Faul et al., 2009). In order to achieve a statistical power of 95% with alpha of 0.05 (one-tailed), a sample size of at least 168 people would be required for each condition, totaling a sample size of 504 for three conditions: foresight, hindsight outcome A, hindsight outcome B. In anticipation of unexpected situations such as careless responses and to make sure that our study would be over-powered, we planned to recruit about ten more participants per comparison.

#### 7.3.2. Participants

A total of 604 American participants were recruited online through CloudResearch (300 females, 302 males, 2 undisclosed, $M_{age} = 38.5$, $SD_{age} = 12.00$, see Supplementary Materials for details about sample characteristics). We did not allow participants who took part in Study 1 to take part in Study 2.

#### 7.3.3. Procedure and materials

The study used a between-subject design. Participants were randomly assigned to one of three conditions. In the *foresight* condition, participants were not presented with any outcomes of an initial trial. In the *hindsight* conditions, because there were two possible outcomes for each scientific trial scenario, half of the participants read that outcome A had occurred in the initial trial (hindsight outcome A condition), and the other half read that outcome B had occurred in the initial trial (hindsight outcome B condition). All participants read all four scenarios: virgin rat, hurricane seeding, gosling imprinting, and Y test, shown in a random order.

The descriptions of the four scenarios were adapted from Slovic and Fischhoff's (1977) Experiment 1 on hindsight bias (see Supplementary Materials for full materials). We use the virgin rat scenario to illustrate the materials and the question format:

*Virgin Rat.*

Several researchers intend to perform the following experiment:

They will inject blood from a mother rat into a virgin rat immediately after the mother rat has given birth. After the injection, the virgin rat will be placed in a cage with the newly born baby rats, after removal of the actual mother.

The possible outcomes were:

(a) the virgin rat exhibited maternal behavior or.

(b) the virgin rat failed to exhibit maternal behavior.

Following each scenario, participants were required to correctly answer comprehension questions before proceeding to the next stage of the study. For the virgin rat scenario, the comprehension question was, "*Which rat will be placed in a cage with the newly born baby?*" The correct answer was "*Virgin rat with mother rat blood injection.*"

Then, participants were asked questions measuring probability estimates (of the initial trial for foresight condition, and of the future trials for both foresight and hindsight conditions), followed by our extension questions measuring surprise and confidence. We present the questions for the virgin rat scenario in Table 7.

*7.3.3.1. Probability estimates of future trials.* Participants were asked to estimate the probability that the outcome would occur in "all," "some," and "none" (or "A," "B," and "C" for the Y-test scenario) of future trials. The percentages of the three items ("all," "some," and "none") needed to add up to 100%. Participants in foresight condition were asked to rate the probabilities of two possible outcomes; participants in hindsight conditions were only asked to rate the outcome which they knew had occurred in the initial trial.

*7.3.3.2. Extension: Surprise ratings.* Following the probability estimates, participants were asked to rate their levels of surprise regarding the outcome(s) (i.e., "Do you think the (outcome) is surprising?") on a 5-point Likert scale (1 = *not surprising at all*, 5 = *extremely surprising*). Participants in the foresight condition were asked to rate the levels of surprise regarding two possible outcomes; participants in the hindsight conditions were only asked to rate the outcome which they were knew had occurred in the initial trial.

*7.3.3.3. Confidence ratings.* For each scenario, participants were asked to rate their confidence (i.e., "How confident are you about the accuracy of your predictions on the probability of the future outcomes of the (scenario)?") on a 7-point Likert scale (0 = *extremely not confident*, 6 = *extremely confident*).

*7.3.3.4. Task difficulty.* After reading all four scenarios, participants

**Table 9**
Study 2: Mean Probabilities in Future Trials (in percentage %).

| Initial result and kind of replication | Foresight | | | Hindsight | | |
|---|---|---|---|---|---|---|
| | *N* | *Mean* | *SD* | *N* | *Mean* | *SD* |
| Virgin rat experiment | | | | | | |
| Outcome A: Shows maternal behavior | | | | | | |
| a. All show maternal behavior** | | **29.16** | 28.09 | | **38.42** | 29.19 |
| b. Some show maternal behavior | 197 | 34.57 | 26.04 | 204 | 36.58 | 25.37 |
| c. None show maternal behavior*** | | 36.27 | 31.44 | | 25.00 | 26.04 |
| Outcome B: Fails to show maternal behavior | | | | | | |
| a. All show maternal behavior | | 17.73 | 23.68 | | 13.89 | 21.81 |
| b. Some show maternal behavior | 197 | 28.08 | 23.90 | 203s | 25.90 | 23.56 |
| c. None show maternal behavior | | **54.20** | 32.83 | | **60.21** | 33.18 |
| Hurricane seeding experiment | | | | | | |
| Outcome A: Intensity increases | | | | | | |
| a. All increase | | **47.74** | 30.13 | | **49.35** | 28.73 |
| b. Some increase | 197 | 33.80 | 24.37 | 204 | 34.99 | 24.98 |
| c. None increase | | 18.45 | 20.60 | | 15.66 | 18.59 |
| Outcome B: Intensity weakens | | | | | | |
| a. All weaken | | **29.59** | 25.52 | | **34.00** | 26.39 |
| b. Some weaken** | 197 | 34.51 | 23.60 | 203 | 41.24 | 25.47 |
| c. None weaken*** | | 35.91 | 30.19 | | 24.77 | 24.50 |
| Gosling imprinting experiment | | | | | | |
| Outcome A: Approaches duck | | | | | | |
| a. All approach duck* | | **39.14** | 27.63 | | **45.26** | 30.62 |
| b. Some approach duck | 197 | 38.50 | 25.96 | 204 | 39.63 | 27.93 |
| c. None approach duck*** | | 22.36 | 24.58 | | 15.10 | 17.73 |
| Outcome B: Approaches goose | | | | | | |
| a. All approach goose** | | **38.10** | 30.38 | | **46.39** | 33.13 |
| b. Some approach goose | 197 | 38.98 | 27.09 | 203 | 36.42 | 27.95 |
| c. None approach goose* | | 22.92 | 24.71 | | 17.19 | 21.90 |
| Y-test experiment | | | | | | |
| Outcome A: Places dot in Area A | | | | | | |
| a. Places in Area A | | **59.62** | 23.92 | | **61.96** | 22.66 |
| b. Places in Area B | 197 | 13.90 | 14.67 | 204 | 15.80 | 17.53 |
| c. Places in Area C* | | 26.48 | 17.98 | | 22.24 | 16.21 |
| Outcome B: Places dot in Area C | | | | | | |
| a. Places in Area A | | 51.54 | 24.18 | | 47.52 | 23.36 |
| b. Places in Area B | 197 | 14.68 | 15.04 | 203 | 13.76 | 14.84 |
| c. Places in Area C* | | **33.78** | 21.56 | | **38.73** | 22.70 |

*Note.* Options and numbers marked in bold represent the kind of replication that was reported to have occurred in the initial trial (hindsight) or could possibly occur in the initial trial (foresight). The foresight ratings of both outcome A and outcome B came from the same participants in the foresight condition. The hindsight ratings came from participants in the hindsight outcome A condition or the hindsight outcome B condition, respectively. *$p < .05$, **$p < .01$, ***$p < .001$.

were required to rate the difficulty of the prediction task (i.e., "*How difficult was it to make estimations of outcomes probabilities?*") on a 7-point Likert scale (1 = *extremely easy*, 7 = *extremely difficult*).

### 7.3.4. Replication evaluation: Very close replication

Our replication study is a very close replication based on the criteria proposed in LeBel et al. (2017) and LeBel et al. (2018). Our IV operationalization and DV operationalization were the same as those used in the original study. For IV stimuli, we made the necessary adjustment to change outcome B in the Y-Test scenario from "*Places in Area B*" to

**Table 10**
Study 2: Independent Samples Student's T-Tests of Probability Estimates between Foresight and Hindsight (Outcome A/B) Conditions.

| Hindsight vs. Foresight | *Mean Difference* | *t* | *df* | *p* | $p_{adjusted}$ | Cohen's *d* | Cohen's *d* 95% CI | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| Virgin rat experiment | | | | | | | | |
| Outcome A: Shows maternal behavior | | | | | | | | |
| **a. All show maternal behavior**** | 9.26 | 3.24 | 399 | 0.001 | 0.006 | 0.32 | 0.12 | 0.52 |
| b. Some show maternal behavior | 2.01 | 0.78 | 399 | 0.434 | 0.521 | 0.08 | -0.12 | 0.27 |
| c. None show maternal behavior*** [a] | -11.27 | -3.92 | 399 | <0.001 | <0.001 | -0.39 | -0.59 | -0.19 |
| Outcome B: Fails to show maternal behavior | | | | | | | | |
| a. All show maternal behavior | -3.83 | -1.69 | 398 | 0.093 | 0.159 | -0.17 | -0.37 | 0.03 |
| b. Some show maternal behavior | -2.18 | -0.92 | 398 | 0.359 | 0.453 | -0.09 | -0.29 | 0.10 |
| **c .None show maternal behavior** | 6.01 | 1.82 | 398 | 0.069 | 0.151 | 0.18 | -0.02 | 0.38 |
| Hurricane seeding experiment | | | | | | | | |
| Outcome A: Intensity increases | | | | | | | | |
| **a. All increases** | 1.61 | 0.55 | 399 | 0.584 | 0.637 | 0.05 | -0.14 | 0.25 |
| b. Some increases | 1.18 | 0.48 | 399 | 0.632 | 0.659 | 0.05 | -0.15 | 0.24 |
| c. None increases | -2.79 | -1.43 | 399 | 0.155 | 0.248 | -0.14 | -0.34 | 0.05 |
| Outcome B: Intensity weakens | | | | | | | | |
| **a. All weaken** | 4.41 | 1.70 | 398 | 0.090 | 0.159 | 0.17 | -0.03 | 0.37 |
| b. Some weaken** | 6.73 | 2.74 | 398 | 0.006 | 0.029 | 0.27 | 0.08 | 0.47 |
| c. None weaken*** [a] | -11.14 | -4.06 | 398 | <0.001 | <0.001 | -0.41 | -0.61 | -0.21 |
| Gosling imprinting experiment | | | | | | | | |
| Outcome A: Approaches duck | | | | | | | | |
| **a. All approach duck*** | 6.12 | 2.10 | 399 | 0.036 | 0.086 | 0.21 | 0.01 | 0.41 |
| b. Some approach duck | 1.13 | 0.42 | 399 | 0.674 | 0.674 | 0.04 | -0.15 | 0.24 |
| c. None approach duck*** [a] | -7.26 | -3.40 | 399 | 0.001 | 0.006 | -0.34 | -0.54 | -0.14 |
| Outcome B: Approaches goose | | | | | | | | |
| **a. All approach goose**[a]** | 8.29 | 2.61 | 398 | 0.009 | 0.036 | 0.26 | 0.06 | 0.46 |
| b. Some approach goose | -2.56 | -0.93 | 398 | 0.353 | 0.453 | -0.09 | -0.29 | 0.10 |
| c. None approach goose* | -5.73 | -2.46 | 398 | 0.014 | 0.042 | -0.25 | -0.44 | -0.05 |
| Y-test experiment | | | | | | | | |
| Outcome A: Places dot in Area A | | | | | | | | |
| **a. Places in Area A** | 2.34 | 1.00 | 399 | 0.316 | 0.446 | 0.10 | -.10 | 0.30 |
| b. Places in Area B [a] | 1.90 | 1.18 | 399 | 0.240 | 0.360 | 0.12 | -0.08 | 0.31 |
| c. Places in Area C* | -4.24 | -2.48 | 399 | 0.013 | 0.042 | -0.25 | -0.45 | -0.05 |
| Outcome B: Places dot in Area C | | | | | | | | |
| a. Places in Area A | -4.02 | -1.69 | 398 | 0.091 | 0.159 | -0.17 | -0.37 | 0.03 |
| b. Places in Area B | -0.93 | -0.62 | 398 | 0.535 | 0.611 | -0.06 | -0.26 | 0.13 |
| **c. Places in Area C*** | 4.95 | 2.23 | 398 | 0.026 | 0.069 | 0.22 | 0.03 | 0.42 |

*Note.* Bolded options indicate the pairs of comparisons of interest. [a] Levene's test was significant. *$p < .05$, **$p < .01$, ***$p < .001$. *p* values were adjusted using the Benjamini and Hochberg (1995) false discovery rate control method.

"*Places in Area C*," so that outcome A and outcome B were symmetric. For DV stimuli, we removed the request for writing down the reasons for why the outcome had occurred, in order to reduce the time required for the experiment in an online setting where participants might have shorter focus than when they were in a physical laboratory. These adjustments were necessary and did not fundamentally change the stimuli used in the replication study. We therefore consider this replication a very close replication of the original study. See Table 8 for a summary of classification, necessary adjustments, and theoretical extensions.

### 7.4. Results

#### 7.4.1. Probability estimates

We summarized the descriptive statistics of probability estimates in Table 9. Violin plots of the probability estimates are available in Supplementary Materials. As there were four scenarios (virgin rat, hurricane seeding, gosling imprinting, Y-test), two possible outcomes (A or B) for the initial trial, and three possible outcomes of future trials (all, some, none for the first three scenarios; A, B, C for the Y-test scenario), we conducted 24 sets of independent samples Student's *t*-tests.

These eight key sets of comparisons are bolded in Tables 9 and 10. For the virgin rat, hurricane seeding, and gosling imprinting scenarios, among the three options (i.e., all, some, and none repetition), we were particularly interested in the probability estimates for repetition in all future trials. For the Y-test scenario with only one future trial, we were interested in the probability estimate of the dot being placed in the same area as in the initial trial.

As shown in Table 10, in four of the eight comparisons, the probability estimates in the hindsight condition were higher than those in the foresight condition, demonstrating hindsight bias. In the other four sets of comparison, the differences in the probability estimates between the hindsight condition and the foresight condition were weaker.

Overall, the results provide moderate support for Hypothesis 3. The effects in all eight sets of comparisons were in the direction of participants in the hindsight condition providing higher estimates than those in the foresight condition, although there were variations depending on the scenario and the outcome.

#### 7.4.2. Extension: Surprise ratings

We summarized the descriptives of surprise ratings in Table 11, and the violin plots are available in the Supplementary Materials. Similar to previous analyses for probability estimates, we conducted eight sets of independent samples Student's *t*-tests to compare the surprise ratings in the foresight and hindsight conditions.

As shown in Table 12, three of the eight sets of comparison of surprise ratings were in support of hindsight bias: hurricane seeding-outcome B, gosling imprinting-outcome B, and Y-test-outcome B. Overall, the results provide some support for Hypothesis 4(a) regarding surprise ratings.

#### 7.4.3. Extension: Confidence ratings

As shown in Table 12, only one of the eight sets of comparison were in support of difference in the confidence ratings between the foresight condition and the hindsight condition: virgin rat scenario-Outcome B. The results for the virgin rat-Outcome A were contrary to our expectation. All other confidence ratings comparison sets had much weaker effects. We concluded results provide no support for Hypothesis 5(a) regarding confidence ratings.

#### 7.4.4. Task difficulty

We conducted an independent samples Student's t-test to examine the difference in the perceived task difficulty. Participants in the hindsight outcome A condition ($M = 4.41$, $S{\cdot}D = 1.61$) reported lower levels of task difficulty than participants in the foresight condition ($M = 4.98$, $S{\cdot}D = 1.43$), $t(399) = -3.79$, $p < .001$, $d = -0.38$, 95% CI [−0.58, −0.18]. Similarly, participants in the hindsight outcome B condition ($M = 4.40$, $S{\cdot}D = 1.51$) reported lower levels of task difficulty than participants in the foresight condition ($M = 4.98$, $S{\cdot}D = 1.43$), $t(398) = -3.98$, $p < .001$, $d = -0.40$, 95% CI [−0.60, −0.20]. Overall, we conclude strong support for Hypothesis 6(a) that participants in the hindsight conditions perceived the task to be less difficult than participants in the foresight condition.

#### 7.4.5. Robustness checks: Alternative tests and exclusion criteria

To examine the robustness of the findings, we conducted additional analyses (see Supplementary Materials for details). First, we tested the Hypotheses 3, 4(a), 5(a), and 6(a) using Mann-Whitney *U* tests, and the results were highly similar to those obtained using Student's independent samples *t*-tests. Second, when we analyzed the data with only participants who met a set of pre-registered exclusion criteria (i.e., self-reported English proficiency and seriousness, and guessing study purpose), we found little to no differences.

#### 7.4.6. Mediation and moderation analyses

We tested the mediation and the moderation hypotheses (see Supplementary Materials for details). Surprise partially mediated the relationship between hindsight (vs. foresight) and probability estimates, supporting H4(b), and confidence moderated the relationship between hindsight (vs. foresight) and probability estimates, supporting H5(c). We found no support for the mediating effects of confidence in H5(b) or task difficulty in H6(b), and no support for the moderating effects of surprise in H4(c) or task difficulty in H6(c).

**Table 11**
Study 2: Means and Standard Deviations of Surprise Ratings and Confidence Ratings.

| Scenario | Outcome A | | | | Outcome B | | | |
|---|---|---|---|---|---|---|---|---|
| | Foresight | | Hindsight | | Foresight | | Hindsight | |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Surprise | | | | | | | | |
| Virgin rat | 3.13 | 1.40 | 2.93 | 1.25 | 1.75 | 1.05 | 1.57 | 0.95 |
| Hurricane seeding | 2.03 | 1.14 | 2.13 | 1.19 | 3.01 | 1.26 | 2.67 | 1.16 |
| Goose imprinting | 2.20 | 1.21 | 2.08 | 1.10 | 2.16 | 1.14 | 1.90 | 1.13 |
| Y-test | 1.81 | 1.06 | 1.66 | 0.95 | 2.46 | 1.17 | 2.14 | 1.01 |
| Confidence | | | | | | | | |
| Virgin rat | 3.61 | 1.56 | 3.17 | 1.58 | 3.61 | 1.56 | 3.91 | 1.5 |
| Hurricane seeding | 3.27 | 1.68 | 3.39 | 1.61 | 3.27 | 1.68 | 3.25 | 1.45 |
| Goose imprinting | 3.41 | 1.62 | 3.49 | 1.53 | 3.41 | 1.62 | 3.67 | 1.48 |
| Y-test | 3.52 | 1.47 | 3.63 | 1.47 | 3.52 | 1.47 | 3.34 | 1.41 |

*Note*. Surprise ratings: 1 = *not surprising at all*, 5 = *extremely surprising*. Confidence ratings: 0 = *extremely not confident*, 6 = *extremely confidence*. The foresight ratings of both outcome A and outcome B came from the same participants in the foresight condition. The hindsight ratings came from participants in the hindsight outcome A condition or the hindsight outcome B condition, respectively. Hindsight participants only rated their surprise over the outcome which they knew had occurred in the initial trial.

**Table 12**
Study 2: Independent samples student's T-tests of surprise and confidence ratings between foresight and hindsight conditions.

| Hindsight vs. Foresight | *t* | *df* | *p* | $p_{adjusted}$ | *d* | 95% CI of *d* | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper |
| Surprise | | | | | | | |
| **Outcome A** | | | | | | | |
| a. Virgin rat | -1.48[a] | 399 | .140 | .187 | -.15 | -.35 | .05 |
| b. Hurricane seeding | .88 | 399 | .382 | .382 | .09 | -.11 | .29 |
| c. Gosling imprinting | -.67 | 399 | .320 | .366 | -.10 | -.30 | .10 |
| d. Y-test | -1.54 | 399 | .124 | .187 | -.15 | -.29 | -.01 |
| **Outcome B** | | | | | | | |
| a. Virgin rat | -1.79 | 398 | .074 | .148 | -.18 | -.38 | .02 |
| b. Hurricane seeding** | -2.82 | 398 | .005 | .020 | -.28 | -.48 | -.08 |
| c. Gosling imprinting* | -2.30 | 398 | .022 | .059 | -.23 | -.43 | -.03 |
| d. Y-test** | -2.92[a] | 398 | .004 | .020 | -.29 | -.49 | -.09 |
| Confidence | | | | | | | |
| **Outcome A** | | | | | | | |
| a. Virgin rat** | -2.79 | 399 | .006 | .048 | -.28 | -.48 | -.08 |
| b. Hurricane seeding | .75 | 99 | .454 | .605 | .07 | -.13 | .27 |
| c. Gosling imprinting | .50 | 399 | .616 | .704 | .05 | -.15 | .25 |
| d. Y-test | .78 | 399 | .436 | .605 | .08 | -.12 | .28 |
| **Outcome B** | | | | | | | |
| a. Virgin rat* | 1.98 | 398 | .049 | .196 | .20 | .002 | .40 |
| b. Hurricane seeding | -.14[a] | 398 | .885 | .885 | -.01 | -.21 | .19 |
| c. Gosling imprinting | 1.70 | 398 | .091 | .243 | .17 | -.03 | .37 |
| d. Y-test | -1.20 | 398 | .232 | .464 | -.12 | -.32 | .08 |

*Note*. Levene's test was significant. * $p < .05$, ** $p < .01$, *** $p < .001$. *p* values were adjusted using the Benjamini and Hochberg (1995) false discovery rate control method.

### 7.5. Discussion

We aimed to replicate Slovic and Fischhoff's (1977) Experiment 1, a study of hindsight bias in prospective judgments. In line with the findings in the original study, we found support for our predictions in four of the eight sets of comparison. Overall, our findings provide moderate support for hindsight bias in prospective judgments.

#### 7.5.1. Replication: Mostly successful

We compared the results of the target experiment and the replication study based on the criteria described in LeBel et al. (2019). As summarized in Table 13 and Fig. 2, in four of the eight sets of probability estimates comparison, we found signals for successful replication. The effect sizes observed in the replication study were similar to those of the target experiment for one outcome, smaller for two outcomes, and larger for one outcome. Overall, we conclude this a mostly successful replication.

## 8. Study 3: Predictions on the replicability of Fischhoff (1975)

### 8.1. Design and procedure

In this study, we asked participants to predict the replicability of Experiment 2 of Fischhoff (1975) and expected hindsight bias over the replicability of hindsight bias.

All participants first read a brief introduction to the main findings of Experiment 2 of Fischhoff (1975). To ease participants' understanding, we 1) removed "Experiment 2" and simply used "Fischhoff (1975)" in this introduction, and 2) focused only on the results about probability estimates in Fischhoff (1975). Participants were then randomly assigned to one of three conditions: Foresight, Hindsight Outcome Success, and Hindsight Outcome Fail. Those in the Foresight condition were told that a group of researchers intended to conduct a replication of Fischhoff (1975), and there were two possible outcomes: successful replication or failed replication. In addition, those in the Hindsight Outcome Success condition were told that the outcome of the replication was successful; those in the Hindsight Outcome Fail condition were told that the outcome of the replication was a failed replication. All participants were asked to write down the reasons for a successful replication and the reasons for a failed replication. They then provided probability estimates of successful and failed replications. They also answered questions about surprise, confidence, and task difficulty.

### 8.2. Hypotheses

Because Study 2 replicated the finding that people tend to use the results of past findings to predict future research outcomes, we expected that:

*H7: Participants in the Foresight condition will predict the probability of a successful replication to be higher than chance (50%).*

In addition, as suggested by previous research on hindsight bias, outcome knowledge might bias probability estimates toward the known outcome. If participants' probability estimates are influenced by knowledge about the replication outcome, then those who were informed of a successful replication would perceive a successful replication to be more probable than those who did not have outcome knowledge, whereas those who were informed of a failed replication would perceive a successful replication to be less probable than those who did not have outcome knowledge. Such hindsight bias may occur through cognitive processes such as memory impairment, biased reconstruction, sense-making, and meta-cognitive experiences, as well as social-motivational processes to increase perceived controllability and enhance self-image (Blank et al., 2007). For example, information about a successful replication may impact the person's memory by

**Table 13**
Study 2: Comparison of Results in the Original Study and the Replication Study.

| Scenario | p-value original | Original effect: Cohen's *d*[a] | *p*-value replication | Replication effect: Cohen's *d* [95% CI] | Replication summary |
|---|---|---|---|---|---|
| Slovic & Fischhoff, 1977 | < 0.05 | 0.36 [0, 0.72] | | | |
| Present Study | | | | | |
| Virgin Rat A | < 0.05 | 0.36 | 0.001 | 0.32 [0.12, 0.52] | Signal – consistent |
| Virgin Rat B | > 0.05 | 0 | 0.069 | 0.18 [−0.02, 0.38] | No signal – consistent |
| Hurricane Seeding A | < 0.001 | 0.61 | 0.584 | 0.05 [−0.14, 0.25] | No signal – inconsistent |
| Hurricane Seeding B | < 0.05 | 0.36 | 0.090 | 0.17 [−0.03, 0.37] | No signal – consistent |
| Gosling Imprinting A | < 0.001 | 0.61 | 0.036 | 0.21 [0.01, 0.41] | Signal – inconsistent, smaller |
| Gosling Imprinting B | > 0.05 | 0 | 0.009 | 0.26 [0.06, 0.46] | Signal – inconsistent, larger |
| Y-Test A | < 0.001 | 0.61 | 0.316 | 0.10 [−0.10, 0.30] | No signal – inconsistent |
| Y-Test B | < 0.001 | 0.61 | 0.026 | 0.22 [0.03, 0.42] | Signal – inconsistent, smaller |

*Note*: a. Estimated using largest possible *p*-values (e.g., 0.001 if $p < .001$; 0.05 if $p < .05$; 0.99 if $p > .05$; see the power analysis in the Supplementary Materials for details).

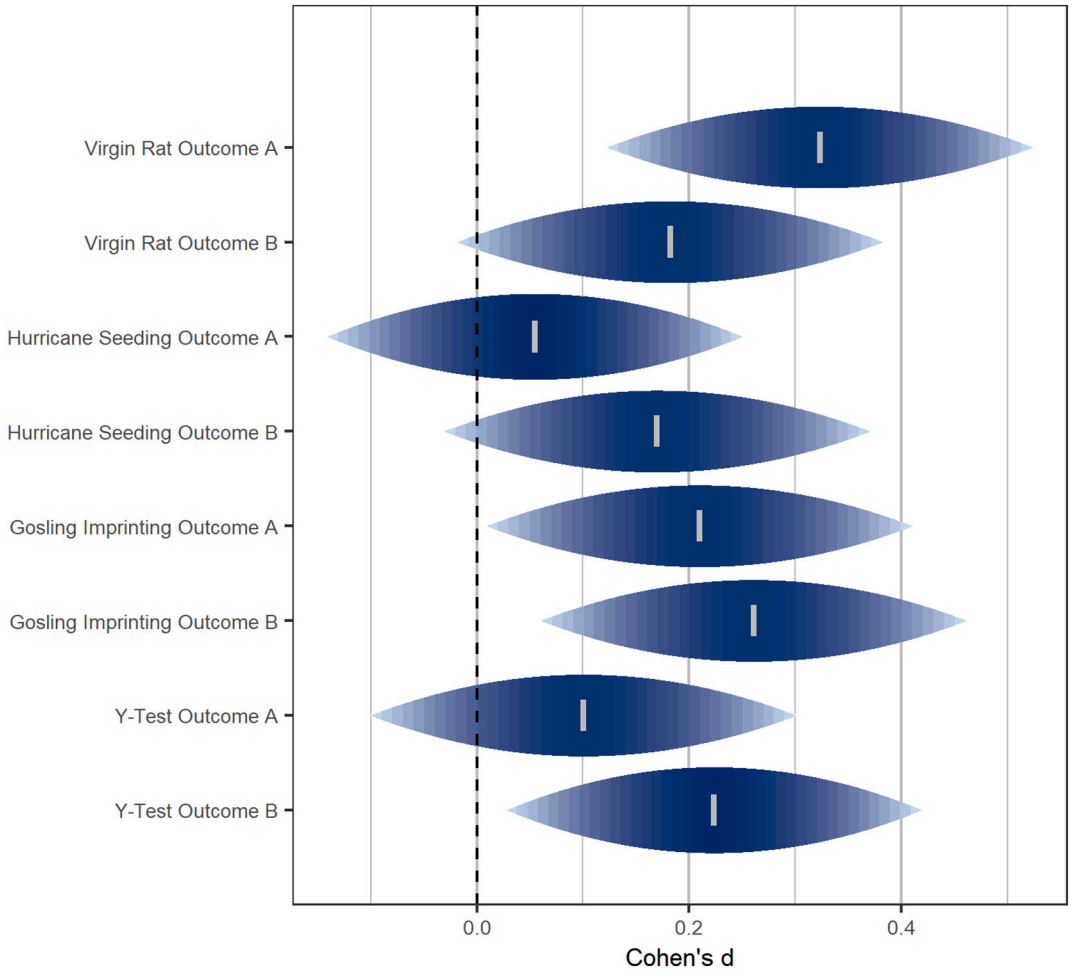



**Fig. 2.** Study 2: Forest Plot of the Effect Size of Probability Estimates.

strengthening the association between relevant cues (e.g., the type of study to be replicated and the research question) and the outcome of a successful replication, or overwriting old knowledge with the newly informed knowledge unconsciously. (e.g., Blank & Nestler, 2007; Hoffrage et al., 2000; Pohl et al., 2003).

Hence, presenting evidence regarding hindsight bias will result in participants in the Hindsight Outcome Success condition predicting the highest probability for successful replication, followed by participants in the Foresight condition, and lastly participants in the Hindsight Outcome Fail condition.

Therefore:

*H8: Participants in the Hindsight Outcome Success condition estimate the probability of a successful replication to be higher than that estimated by participants in the Hindsight Outcome Fail condition.*

*H9: Participants in the Hindsight conditions estimate a greater probability for the informed outcome of replication, compared with participants in the Foresight condition.*

### 8.3. Method

#### 8.3.1. Power analysis

The planned sample size for the replication study was calculated based on pretests indicating an effect size of $d = 0.4$ (see supplementary for details), with power of 95% with alpha of 0.05 (two-tailed) requiring a sample size of 164 people for each condition, totaling a sample size of 492. We collected slightly more responses to address the possibility of unexpected exclusions.

**Table 14**
Study 3: Mean Estimations of Outcomes of a Replication of Fischhoff (1975) (in percentage %).

| | Foresight ($n = 154$) | | Hindsight Outcome Success: Successful Replication ($n = 178$) | | Hindsight Outcome Fail: Failed Replication ($n = 188$) | |
|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| Estimated probabilities | | | | | | |
| a. Successful replication | 65.36 [a] | 18.08 | 73.07 [b] | 17.46 | 52.22 [c] | 22.62 |
| b. Failed replication | 34.64 [a] | 18.08 | 26.93 [b] | 17.46 | 47.78 [c] | 22.62 |
| Surprise | | | | | | |
| a. Successful replication | 2.22 [a] | 1.28 | 2.16 [a] | 1.24 | 2.42 [a] | 1.26 |
| b. Failed replication | 3.06 [a] | 1.13 | 3.38 [b] | 1.12 | 2.89 [a,c] | 1.14 |
| Confidence | 3.99 [a] | 1.29 | 4.18 [a] | 1.30 | 3.64 [b] | 1.39 |
| Task difficulty | 3.98 [a] | 1.66 | 3.89 [a] | 1.73 | 4.19 [a] | 1.58 |

*Note.* $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Means with different superscripts (a, b, c) were significantly different from each other.

#### 8.3.2. Participants

A total of 520 American participants were recruited online through CloudResearch (228 females, 289 males, 3 undisclosed, $M_{age} = 38.96$, $SD_{age} = 12.18$, see Supplementary Materials for details about sample characteristics).

**Table 15**
Study 3: Independent Samples Student's T-Tests of Estimations of Outcomes of a Replication of Fischhoff (1975).

| Hindsight vs. Foresight | *Mean Difference* | *t* | *df* | *p* | Cohen's *d* | 95% CI of Cohen's *d* | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper |
| Estimated probabilities of successful replication | | | | | | | |
| Hindsight Outcome Success vs. Foresight | 7.71 | 3.95 | 330 | <0.001 | 0.43 | 0.21 | 0.65 |
| Hindsight Outcome Fail vs. Foresight | −13.15 | −5.84 | 340 | <0.001 | −0.64 | −0.86 | −0.41 |
| Hindsight Outcome Success vs. Hindsight Outcome Fail | 20.85 | 9.84 [a] | 364 | <0.001 | 1.03 | 0.80 | 1.26 |
| Surprise about successful replication | | | | | | | |
| Hindsight Outcome Success vs. Foresight | −0.06 | −0.42 | 330 | 0.677 | −0.05 | −0.27 | 0.17 |
| Hindsight Outcome Fail vs. Foresight | 0.20 | 1.45 | 340 | 0.149 | 0.16 | −0.05 | 0.37 |
| Hindsight Outcome Success vs. Hindsight Outcome Fail | −0.26 | −1.97 | 364 | 0.050 | −0.21 | −0.42 | 0.00 |
| Surprise about failed replication | | | | | | | |
| Hindsight Outcome Success vs. Foresight | 0.32 | 2.56 | 330 | 0.011 | 0.28 | 0.06 | 0.50 |
| Hindsight Outcome Fail vs. Foresight | −0.16 | −1.33 | 340 | 0.184 | −0.14 | −0.35 | 0.07 |
| Hindsight Outcome Success vs. Hindsight Outcome Fail | 0.48 | 4.07 | 364 | <0.001 | 0.43 | 0.22 | 0.64 |
| Confidence | | | | | | | |
| Hindsight Outcome Success vs. Foresight | 0.19 | 1.31 | 330 | 0.192 | 0.14 | −0.08 | 0.36 |
| Hindsight Outcome Fail vs. Foresight | −0.35 | −2.40 | 340 | 0.017 | −0.26 | −0.47 | −0.04 |
| Hindsight Outcome Success vs. Hindsight Outcome Fail | 0.54 | 3.80 | 364 | <0.001 | 0.40 | 0.19 | 0.61 |
| Task difficulty | | | | | | | |
| Hindsight Outcome Success vs. Foresight | −0.09 | −0.50 | 330 | 0.620 | −0.05 | −0.27 | 0.17 |
| Hindsight Outcome Fail vs. Foresight | 0.21 | 1.17 | 340 | 0.243 | 0.13 | −0.08 | 0.34 |
| Hindsight Outcome Success vs. Hindsight Outcome Fail | −0.30 | −1.73 | 364 | 0.085 | −0.18 | −0.39 | 0.03 |

*Note.* a. Levene's test was nonsignificant for all comparisons.

#### 8.3.3. Procedure and materials

The study used a between-subject design. Participants were randomly assigned to one of three conditions. In the *Foresight* condition, participants did not receive any knowledge about the actual outcome of the replication study. In the hindsight conditions, because there were two possible outcomes for each scientific trial scenario, half of the participants read that the replication was successful (*Hindsight Outcome Success condition*), and the other half read that replication failed (*Hindsight Outcome Fail condition*). Following the information, participants were required to correctly answer two comprehension questions before proceeding to the next stage of the study. Participants then responded to two open-ended questions asking the reasons for successful or failed replications.

#### 8.3.4. Probability estimates of replication outcomes

Participants were then asked to provide probability estimates for both Outcome A (the hindsight bias effect will be successfully replicated) and Outcome B (the hindsight bias effect will fail to replicate). In the *Foresight* condition, the instructions were: "In light of the information appearing in the paragraphs provided, please estimate the probabilities of occurrence of the two possible outcomes in the replication study. There are no right or wrong answers, answer based on your intuition. (The probabilities should sum to 100%)." In the *Hindsight* conditions, the instructions contained an additional sentence: "Answer as if you do not know the outcome, estimating the probabilities at that time before the replication study was launched."

#### 8.3.5. Surprise, confidence, and task difficulty ratings: exploratory

We added exploratory measures of surprise, confidence, and task difficulty. Exploratory hypotheses and findings are reported in the supplementary.

Participants were asked to rate their surprise about both Outcome A and Outcome B, confidence about the accuracy of their estimation, and perceived task difficulty. Measures of surprise, confidence, and task difficulty were similar or identical to those used in Study 2.

### 8.4. Results

We summarized the descriptive statistics of probability estimates, surprise, confidence, and task difficulty in Table 14. Violin plots of these variables are available in Supplementary Materials.

We conducted a one-sample *t*-test to test H7. We found that the probability estimates for a successful replication ($Mean_{Prob} = 65.36\%$, $S.D._{Prob} = 18.08\%$) were higher than chance (50%), $t(153) = 10.55$, $p < .001$, $d = 0.85$. We concluded support for H7.

We conducted independent samples *t*-tests to test H8 and H9. As shown in Table 15, participants who were informed of Outcome Success estimated a successful replication to be *more* probable than participants who were informed of Outcome Fail, $t(364) = 9.84$, $p < .001$, Cohen's $d = 1.03$, 95% CI [0.80, 1.26]. In addition, participants who were informed of Outcome Success estimated a successful replication to be *more* probable than participants who did not know the outcome, $t(330) = 3.95$, $p < .001$, Cohen's $d = 0.43$, 95% CI [0.21, 0.65]. In contrast, participants who were informed of Outcome Fail estimated a successful replication to be *less* probable than participants who did not know the outcome, $t(340) = -5.84$, $p < .001$, Cohen's $d = -0.64$, 95% CI [−0.86, −0.41]. The results therefore provided strong support for H8 and H9.

### 8.5. Robustness checks

To examine the robustness of the findings, we conducted additional analyses (see Supplementary Materials for details). When we analyzed the data with only participants who met a set of pre-registered exclusion criteria (i.e., self-reported English proficiency and seriousness, and guessing study purpose), we found little to no differences between the results with the full sample and the results after exclusion.

### 8.6. Exploratory extensions

We found some support for the mediating role and the moderating role of surprise over the alternative outcome for the relationship between Hindsight Outcome Success condition and probability estimates of Outcome A. However, there was no support for any other hypothesized mediating or moderating effects, and we concluded weak to no support for the mediating or moderating effects. Hypotheses, analyses, and results are provided in the supplementary.

### 8.7. Discussion

We found strong support of hindsight bias for the replicability of hindsight bias. First, being presented with an outcome of Fischhoff's (1975) original study, participants' probability estimates of a successful replication were higher than chance. Second, participants' probability estimates of a certain outcome were higher when they knew the

outcome than when they did not know the outcome.

## 9. General Discussion

We conducted very close replications of Experiment 2 in Fischhoff (1975) and Experiment 1 in Slovic and Fischhoff (1977), and found support for hindsight bias in both retrospective and prospective judgments. In retrospective judgments (Study 1: replication of Fischhoff, 1975), participants were asked to predict the probability of an outcome in a past event. Compared to participants who had no knowledge about the actual outcome of the event, participants who knew the actual outcome estimated the probability of the actual outcome to be higher, even if they were asked to estimate as if they did not know the actual outcome. In prospective judgments (Study 2: replication of Slovic & Fischhoff, 1977), participants were told that researchers had conducted an initial trial of an experiment, and would conduct either one or multiple trials of the same kind in the future. The participants' job was to predict the outcome of those future trials. Compared to participants who had no knowledge of the actual outcome of the initial trial, participants who knew the actual outcome of the initial trial predicted the probability of the actual outcome in future trials to be higher.

Building on these two replication studies, we added a third study to examine hindsight bias in estimating the replicability of hindsight bias. Our findings suggest that estimates of replication outcomes were heavily influenced by outcome knowledge. Overall, participants predicted a successful replication for Fischhoff (1975). The probability estimates of a successful replication were highest among those who were informed of a successful replication, moderate among those who were not informed of an outcome, and lowest among those who were informed of a failed replication. Our findings suggest that probability estimations regarding research and replication outcomes were affected by hindsight bias.

### 9.1. Replications: comparison with original findings

In our two replication studies, results were mostly in line with the original findings with some minor deviations. We concluded these replications as mostly successful despite these deviations for two reasons. First, study materials were designed almost half a century ago, and some participants may have been more knowledgeable about some of these stimuli than participants in the 1970s. For example, in the Y-test scenario of Study 2, a 4-year-old child was asked to determine the relative position of a dot to the letter Y when viewed from the back of the easel, like in a left-right mirror image. Back in 1970s, people might not necessarily know the more likely choice of the child. However, today, following wider dissemination of findings in developmental and cognitive psychology, more people may have had the insight that mirror-image confusions are prevalent among children, because the abilities that are required to make the correct choice, such as spatial cognition (Colby, 2009) and theory of mind (Wellman & Liu, 2004), are not well-developed among 4-year olds (Gregory, Landau, & McCloskey, 2011). In the target experiment, the average probability of outcome A (“places in area A", showing a lack of spatial cognition and theory of mind) in the foresight condition was 0.29. However, in the replication study, the number was much higher (0.60), possibly indicating a shift of knowledge regarding this phenomenon over the decades. Similarly, in the hurricane seeding scenario in Study 2, the average probability of outcome A (“All increase”) was 0.29 in the target experiment, and 0.48 in our replication study. When participants hold certain knowledge prior to taking part in the study, their probability estimates may be less influenced by the study's manipulation of outcome knowledge (of the initial trial), weakening hindsight bias. Given these changes, we consider our findings an impressive demonstration of the generalizability and relevancy of the effect.

Second, for Study 2, while the target experiment asked the participants to write down why they thought the outcome would happen, we did not include this question in the replication study. When asked to provide explanations of an outcome, the person would have to temporarily assume that outcome is true, and then assess its plausibility. Such cognitive processes can lead the person to perceive the outcome to be more plausible, persuasive, or even inevitable (Koehler, 1991). It is therefore possible that writing down the reasons for the outcome reinforces participants' belief that the outcome is true, which in turn intensifies hindsight bias. In our replication study we had to make adjustments to remove the step of providing explanations and this may have led to the observed effect size to be smaller than the case when participants were asked to provide explanations.. We note, however, that this explanation does not clarify the weaker effects in Study 1. It could be that the effect size of hindsight bias is larger for retrospective judgments, and smaller for prospective judgments. This possibility awaits further investigation.

### 9.2. Extensions

We added several extensions. In Study 1, we found no support for the mediating effect of surprise in the relationship between hindsight condition and probability estimates, and inconclusive results for the moderating effect of surprise on the relationship between hindsight condition and probability estimates. In Study 2, we found some support for surprise, but not for confidence, as a mediator of the relationship between hindsight condition and probability estimates. In addition, we found support for confidence, but not for surprise, as a moderator of the relationship between hindsight condition and probability estimates. Hindsight bias was evident when confidence about one's own judgments was high, but it was reversed when confidence was low. In Study 3, we found weak to no support for the mediating role and the moderating role of surprise. Other than that, there was no support for the mediating or the moderating effects of surprise, confidence, and task difficulty.

Given these mixed findings, we are hesitant to offer any conclusions regarding surprise and confidence. Past findings regarding the effect of surprise were not unequivocal. Although many articles argued that hindsight bias could be caused by a lack of scrutiny and consideration of alternatives associated with a lack of surprise feelings (Sanna & Schwarz, 2006; Slovic & Fischhoff, 1977), other research noted that a certain level of surprise is required for hindsight bias to occur—after all, if the person already had the knowledge (thus would not feel surprised), then his/her estimation of the probability shall not be affected by the outcome knowledge provided by the researcher (Pezzo, 2003). In testing the robustness of hindsight bias, some research found that hindsight bias persisted even when the materials and outcome knowledge were difficult or unexpected by the participants (e.g., Ash, 2009; Fischhoff, 1977; Hoch & Loewenstein, 1989; Roese & Olson, 1996; Wood, 1978), suggesting that surprise did not necessarily hinder hindsight bias. Furthermore, Schkade and Kilbourne (1991) found that hindsight bias was larger when outcomes were inconsistent with expectations than when they were consistent. The authors reasoned that this could be because the process of assimilating the outcome knowledge into what was already known was immediate and at least partially automatic. Thus, the more different and surprising the outcome knowledge was from prior knowledge, the larger the hindsight bias; the more familiar the outcome knowledge was from prior knowledge, the less likely that a cognitive reconstruction leading to hindsight bias will occur. More research is needed to clarify these varying theoretical arguments and mixed findings about the role of surprise in hindsight bias.

Previous studies have linked hindsight bias to confidence, yet there are studies that failed to detect such associations. Ross (2012) found that the effect of outcome knowledge on probability estimates and that on confidence are disconnected. In addition, Schatz (2019) failed to find support for the relationship between receiving outcome knowledge and confidence across ten studies. These and our findings suggest more research is needed to understand role of confidence in hindsight bias, yet it is possible that these links have been overestimated.

In addition, studies in the literature tend to consider surprise and

confidence as two sides of the same coin, based on an assumption that feelings of surprise may reduce a person's confidence about a judgment. However, we found no indication for such an association. Future studies may aim to differentiate and contrast surprise and confidence in hindsight bias.

We found no support for the mediating effect or moderating effect of subjective task difficulty in the relationship between hindsight condition and probability estimates. Although participants in the hindsight condition perceived the task to be easier, this decreased perceived difficulty did not seem to predict probability estimates. Task difficulty was negatively associated with confidence about one's own judgments, and weakly positively associated with surprise of the outcome. Similar to surprise, the literature also showed discrepancies in whether hindsight bias is larger in more difficult or less difficult tasks (see for example Arkes et al., 1981; Harley et al., 2004). More research is needed to address these discrepancies and clarify the role of task difficulty in hindsight bias.

### 9.3. Take-aways for Science: Endorsement of Open Science practices

In the introduction we discussed direct and important implications of hindsight bias for science. Beyond our successful replications of classic hindsight bias studies, we also successfully demonstrated the application of hindsight bias regarding our very own replication of hindsight bias.

We were asked by the editor and reviewers to discuss our views on possible ways to address hindsight bias in the scientific process. First, there is the issue of raising awareness to hindsight bias pitfalls. To be able to overcome this bias, there needs to be some awareness that the problem exists, and some scholars in the open-science community have been trying to raise awareness to the impact of cognitive biases and study these systematically using meta research (e.g., Bishop, 2019, 2020a, 2020b). Second, pre-registrations - if done appropriately - seem like a promising direction against researchers fooling themselves by making a public commitment regarding their hypotheses, design, procedures, and data analysis plans (Nosek et al., 2018; Shrout & Rodgers, 2018; van't Veer & Giner-Sorolla, 2016). These may at the very least address the issues of unintended memory reconstruction and HARKing, since researchers can easily go back to their pre-registrations and examine their findings against their prior plans. These may also partly serve to ensure others of the researchers' open transparent research process, and demonstrate researchers' public commitment to overcoming their own biases.

Third, Registered Reports publication format (Chambers & Tzavella, 2020; Simons et al., 2014) and results-blind review (Button, Bal, Clark, & Shipley, 2016) can reduce hindsight bias in the publication review process by addressing outcome driven interpretations and the pressures on authors to adhere to a certain outcome. Determining whether to accept or reject a replication study prior to data collection also helps address outcome bias (Baron & Hershey, 1988; Savani & King, 2015), where a failed replication (i.e., a bad outcome) leads to perceiving the study or the replicators as lower quality compared to a successful replication (i.e., a good outcome). Endorsement of Replication Registered Reports as an integral part of the scientific process, with directions like the Pottery Barn rule (if you publish it, you commit to publishing replications of it; Edlund, Cuccolo, Irgens, Wagge, & Zlokovich, 2020; Srivastava, 2012) and a commitment to publishing all well-executed replications (e.g., Chambers, 2018) may help overcome inherent biases against replications as being more predictable and of lower value (Zwaan, Etz, Lucas, & Donnellan, 2018).

Lastly, and most important, systematically documenting and openly sharing everything about the research life-cycle, from initial idea and research question, through process, design, and decisions, to materials, data, and code, with public commitment and openness toward third party open peer review, can greatly reduce human biases introduced in the scientific process and encourage collaboration and sharing. This is the essence of open science.

### 9.4. Limitations and future research

In all three studies, we used the hypothetical design to test hindsight bias ("answer as if you did not know the outcome"). However, this design makes it difficult to examine psychological processes underlying hindsight bias. We therefore encourage future studies to 1) replicate further studies about hindsight bias which had a stronger focus on the underlying psychological processes, and 2) extend our findings in Study 3 using other designs, such as memory recall (Pohl, 2007), and multinomial processing trees (Bernstein et al., 2011; Groß & Bayen, 2015; Hell, Gigerenzer, Gauggel, Mall, & Müller, 1988).

We conducted all studies using an American sample, and future studies may aim to extend our efforts to also examine samples from other diverse cultures.

We discussed possible implications of hindsight bias for science, yet these were inferred rather than directly tested. We believe that this is a promising and much needed area of research. Future research may aim to directly examine whether and to what extent hindsight bias influences researchers' decisions to embark on replications and reviewers' and editors' decisions to publish a replication study. If such a bias is found, it would be imperative to further examine the impact of our above suggested solutions and other potential remedies to overcome this bias.

This replication presented us with a special challenge, regarding some of the events included in the original stimuli of Fischhoff (1975). Events C and D used in the original were from a classic clinical psychology book by Ellis from the 1960s. The original authors reflected on the use of these stimuli and noted that the scenarios described patients "in terms that fit now–antiquated mores and theories" (Fischhoff, 2007, p. 11; also see interview in Klein, Hegarty, & Fischhoff, 2017). In correspondence with the original author and the editor we felt it needed to include a warning note that that these stimuli should no longer be used in follow-up research. We removed the reporting of these materials and analyses of these events from the manuscript and the supplementary.

## 10. Conclusion

We conducted two close replication studies and one novel study to investigate hindsight bias. In Study 1, we found support for hindsight bias as in Experiment 2 of Fischhoff (1975). Participants were more likely to estimate the probability of an outcome to be higher when they knew that the outcome actually occurred. In Study 2, we found some support for hindsight bias as in Experiment 1 of Slovic and Fischhoff (1977). When informed of the outcome of an initial trial, participants were more likely to predict this same outcome to repeatedly occur in future trials. In Study 3, we found support for hindsight bias over the replicability of hindsight bias. We found mixed weak to no support for the mediating and moderating roles of surprise, confidence, and task difficulty. We conclude that after almost five decades since the original studies were published, we found consistent evidence for hindsight bias.

## Financial disclosure/funding

This research was supported by the European Association for Social Psychology seedcorn grant.

## Authorship declaration

Gilad led the reported replication effort with the team listed below. Gilad supervised each step of the project, conducted the pre-registration, and ran data collection. Jieying followed up on initial work by the other coauthors to verify analyses and conclusions, added advanced tables and plots, designed, ran, and analyzed the third study, and completed the manuscript submission draft. Jieying and Gilad jointly finalized the manuscript for submission.

Lok Ching (Roxane) Kwan, Lok Yeung (Loren) Ma, Hiu Yee (HayleyAnne) Choi, Ying Ching (Lita) Lo, Shin Yee (Sarah) Au, and Chi Ho

(Toby) Tsang conducted the two replication studies as part of university coursework. They conducted an initial analysis of the paper, designed the replication, initiated the extensions, wrote the pre-registrations, conducted initial data analyses, and wrote initial replication reports.

Bo Ley Cheng guided and assisted the replication effort.

## Contributor roles taxonomy

In the table below, employ CRediT (Contributor Roles Taxonomy) to identify the contribution and roles played by the contributors in the current replication effort. Please refer to the url (https://www.casrai.org/credit.html) on details and definitions of each of the roles listed below.

| Role | Jieying Chen | Gilad Feldman | Lok Ching Kwan, Lok Yeung (Loren) Ma, Hiu Yee (HayleyAnne) Choi, Ying Ching (Lita) Lo, Shin Yee (Sarah) Au, and Chi Ho (Toby) Tsang | Bo Ley Cheng |
|---|---|---|---|---|
| Conceptualization | X | X | | |
| Pre-registrations | X | X | X | |
| Data curation | | X | | |
| Formal analysis | X | X | X | |
| Funding acquisition | | X | | |
| Investigation | X | X | X | |
| Methodology | X | X | X | |
| Pre-registration peer review / verification | X | X | X | X |
| Data analysis peer review / verification | X | | X | |
| Project administration | | X | | X |
| Resources | | X | | |
| Software | X | X | X | |
| Supervision | | X | | X |
| Validation | X | X | | |
| Visualization | X | | | |
| Writing-original draft | X | X | | |
| Writing-review and editing | X | X | | |

## Declaration of Competing Interest

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

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2021.104154.

## References


Aarts, H., Verplanken, B., & Van Knippenberg, A. (1998). Predicting behavior from actions in the past: Repeated decision making or a matter of habit? *Journal of Applied Social Psychology, 28*, 1355–1374.

Arkes, H. R. (2013). The consequences of the hindsight bias in medical decision making. *Current Directions in Psychological Science, 22*, 356–360.

Arkes, H. R., Wortmann, R. L., Saville, P. D., & Harkness, A. R. (1981). Hindsight bias among physicians weighing the likelihood of diagnoses. *Journal of Applied Psychology, 66*, 252–254.

Ash, I. K. (2009). Surprise, memory, and retrospective judgment making: Testing cognitive reconstruction theories of the hindsight bias effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 916–933.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54*, 569–579.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, 57*, 289–300.

Bernstein, D., Aßfalg, A., Kumar, R., & Ackerman, R. (2016). *Looking backward and forward on hindsight bias. Handbook of Metamemory (pp. 289–304)*. Oxford, UK: Oxford University Press.

Bernstein, D. M., Erdfelder, E., Meltzoff, A. N., Peria, W., & Loftus, G. R. (2011). Hindsight bias from 3 to 95 years of age. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 378–391.

Bishop, D. (2019). Fixing the replication crisis: The need to understand human psychology. *APS Observer, 32*(10).

Bishop, D. (2020a). How scientists can stop fooling themselves over statistics. *Nature, 584*(7819), 9.

Bishop, D. (2020b). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th sir Frederic Bartlett lecture. *Quarterly Journal of Experimental Psychology, 73*(1), 1–19.

Blank, H., Musch, J., & Pohl, R. F. (2007). Hindsight bias: On being wise after the event. *Social Cognition, 25*, 1–9.

Blank, H., & Nestler, S. (2007). Cognitive process models of hindsight bias. *Social Cognition, 25*, 132–146.

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69*, 709–750.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224.

Bukszar, E., & Connolly, T. (1988). Hindsight bias and strategic choice: Some problems in learning from experience. *Academy of Management Journal, 31*, 628–641.

Button, K. S., Bal, L., Clark, A., & Shipley, T. (2016). Preventing the ends from justifying the means: Withholding results to address publication bias in peer-review. *BMC Psychology, 4*(1), 1–7.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., … Altmejd, A. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour, 2*, 637–644.

Casper, J. D., Benedict, K., & Perry, J. L. (1989). Juror decision making, attitudes, and the hindsight bias. *Law and Human Behavior, 13*, 291–310.

Chambers, C. D. (2018). Reproducibility meets accountability: Introducing the replications initiative at Royal Society Open Science. In *Royal Society Open Science*. Retrieved from https://royalsociety.org/blog/2018/10/reproducibility-meets-accountability/.

Chambers, C. D., & Tzavella, L. (2020). *Registered Reports: Past, Present and Future*. https://doi.org/10.31222/osf.io/43298.

Christensen-Szalanski, J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes, 48*, 147–168.

Cohen, J. E. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Colby, C. L. (2009). Spatial Cognition. *Encyclopedia of Neuroscience*, 165–171.

Davis, A. L., & Fischhoff, B. (2014). Communicating uncertain experimental evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 261–274.

Dawson, N. V., Connors, A. F., Jr., Speroff, T., Kemka, A., Shaw, P., & Arkes, H. R. (1993). Hemodynamic assessment in the critically ill: Is physician confidence warranted? *Medical Decision Making, 13*, 258–266.

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D. J., Buttrick, N. R., Chartier, C. R., … Szecsi, P. (2020). Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science, 3*(3), 309–331.

Edlund, J., Cuccolo, K., Irgens, M. S., Wagge, J. R., & Zlokovich, M. S. (2020). *Saving Science Through Replication Studies*. https://doi.org/10.31234/osf.io/efypc.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.

Fay, M. P., & Malinovsky, Y. (2018). Confidence intervals of the Mann-Whitney parameter that are compatible with the Wilcoxon-Mann-Whitney test. *Statistics in Medicine, 37*, 3991–4006.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 104*, 288–299.

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 349–358.

Fischhoff, B. (2007). An early history of hindsight research. *Social Cognition, 25*(1), 10–13.

Fischhoff, B., & Beyth, R. (1975). I knew it would happen: Remembered probabilities of once—Future things. *Organizational Behavior and Human Performance, 13*, 1–16.

Forer, B. R. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology, 44*, 118–123.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*, 2–18.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem even when there is no "fishing expectation" or "p-hacking" and the research hypothesis was posited ahead of time. Retrieved from *https://osf.io/n3axs*.

Granhag, P. A., Strömwall, L. A., & Allwood, C. M. (2000). Effects of reiteration, hindsight bias, and memory on realism in eyewitness confidence. *Applied Cognitive Psychology, 14*, 397–420.

Gregory, E., Landau, B., & McCloskey, M. (2011). Representation of object orientation in children: Evidence from mirror-image confusions. *Visual Cognition, 19*, 1035–1062.

Groß, J., & Bayen, U. J. (2015). Adult age differences in hindsight bias: The role of recall ability. *Psychology and Aging, 30*, 253–258.
Guilbault, R. L., Bryant, F. B., Brockway, J. H., & Posavac, E. J. (2004). A meta-analysis of research on hindsight bias. *Basic and Applied Social Psychology, 26*, 103–117.
Harley, E. M., Carlsen, K. A., & Loftus, G. R. (2004). The“ saw-it-all-along” effect: Demonstrations of visual hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 960–968.
Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107*, 311–327.
Hell, W., Gigerenzer, G., Gauggel, S., Mall, M., & Müller, M. (1988). Hindsight bias: An interaction of automatic and motivational factors? *Memory & Cognition, 16*, 533–538.
Hertwig, R., Fanselow, C., & Hoffrage, U. (2003). Hindsight bias: How knowledge and heuristics affect our reconstruction of the past. *Memory, 11*, 357–377.
Hoch, S. J., & Loewenstein, G. F. (1989). Outcome feedback: Hindsight and information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 605–619.
Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 566–581.
Hoffrage, U., & Pohl, R. (2003). Research on hindsight bias: A rich past, a productive present, and a challenging future. *Memory, 11*, 329–335.
Hom, H. L., Jr., & Van Nuland, A. L. (2019). Evaluating scientific research: Belief, hindsight bias, ethics, and research evaluation. *Applied Cognitive Psychology, 33*, 675–681.
Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2* (8), Article e124.
John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532.
Kaplan, H., & Barach, P. (2002). Incident reporting: Science or protoscience? Ten years later. *BMJ Quality & Safety, 11*, 144–145.
Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.
Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., … Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology, 4*, 1–15.
Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., … Sowden, W. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490.
KNAW: Royal Dutch Academy of Arts and Sciences. (2018). Replication studies: Improving reproducibility in the empirical sciences. Amsterdam, Netherlands Retrieved from *https://knaw.nl/en/news/publications/replication-studies*.
Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110*(3), 499–519.
LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology, 11*, 254–261.
LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1*, 389–402.
LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology, 3*. MP.2018.843.
Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. Com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433–442.
Mazursky, D., & Ofir, C. (1990). “I could never have expected it to happen”: The reversal of the hindsight bias. *Organizational Behavior and Human Decision Processes, 46*, 20–33.
Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science, 1*(4), 501–515.
Müller, P. A., & Stahlberg, D. (2006). *Surprise as information: Metacognitive influences on hindsight bias*. Unpublished manuscript. Germany: University of Mannheim.
Müller, P. A., & Stahlberg, D. (2007). The role of surprise in hindsight bias: A metacognitive model of reduced and reversed hindsight bias. *Social Cognition, 25*, 165–184.
Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., … Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 1–9.
Nestler, S., & Egloff, B. (2009). Increased or reversed? The effect of surprise on hindsight bias depends on the hindsight component. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1539–1544.
Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600–2606.
Nosek, B. A., & Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology, 45*, 137–141.
Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631.
Ofir, C., & Mazursky, D. (1997). Does a surprising outcome reinforce or reverse the hindsight bias? *Organizational Behavior and Human Decision Processes, 6*, 51–57.
Open, S. C. (2015). Psychology. Estimating the reproducibility of psychological science. *Science, 349*(6251). aac4716.
Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin, 124*, 54–74.
Pezzo, M. (2003). Surprise, defence, or making sense: What removes hindsight bias? *Memory, 11*, 421–441.
Pohl, R., Eisenhauer, M., & Hardt, O. (2003). SARA: A cognitive process model to simulate the anchoring effect and hindsight bias. *Memory, 11*, 337–356.
Pohl, R. F. (2007). Ways to assess hindsight bias. *Social Cognition, 2*, 14–31.
Pohl, R. F., Bender, M., & Lachmann, G. (2002). Hindsight bias around the world. *Experimental Psychology, 49*, 270–282.
Roese, N. J., & Olson, J. M. (1996). Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology, 32* (3), 197–227.
Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science, 7*, 411–426.
Ross, M. (2012). *The hindsight bias: Judgment task differentiation*. doctoral dissertation. Old dominion university.
Sanna, L. J., & Schwarz, N. (2006). Metacognitive experiences and human judgment: The case of hindsight bias and its debiasing. *Current Directions in Psychological Science, 15*, 172–176.
Savani, K., & King, D. (2015). Perceiving outcomes as determined by external forces: The role of event construal in attenuating the outcome bias. *Organizational Behavior and Human Decision Processes, 130*, 136–146.
Schatz, D. A. (2019). *Boundaries of the hindsight bias*. Doctoral dissertation. Berkeley: University of California.
Scheel, A. M., Schijen, M., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. In *Advances in Methods and Practices in Psychological Science*.
Schkade, D. A., & Kilbourne, L. M. (1991). Expectation-outcome consistency and hindsight bias. *Organizational Behavior and Human Decision Processes, 49*, 105–123.
Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*, 487–510.
Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science, 9*, 552–555.
Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 544–551.
Slovic, P., Lichtenstein, S., & Fischhoff, B. (1988). Decision-making. In R. C. Atkinson, et al. (Eds.), *Learning and cognition: Vol. 2. Steven's handbook of experimental psychology* (pp. 673–738). New York, NY: Wiley.
Srivastava, S. (2012). A Pottery Barn rule for scientific journals. Retreived from: https://hardsci.wordpress.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/.
van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2–12.
Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review, 106*, 1577–1600.
Veldkamp, C. (2017). *The human fallibility of scientists: Dealing with error and bias in academic research*. doctoral dissertation. Tilburg University.
Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632–638.
Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing research with undergraduate students via replication work: The collaborative replications and education project. *Frontiers in Psychology, 10*, 247.
Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*, 523–541.
Werth, L., & Strack, F. (2003). An inferential approach to the knew-it-all-along phenomenon. *Memory, 11*(4–5), 411–419.
Winman, A., Juslin, P., & Björkman, M. (1998). The confidence–hindsight mirror effect in judgment: An accuracy-assessment model for the knew-it-all-along phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 415.
Wong, L. Y. S. (1995). Research on teaching: Process-product research findings and the feelings of obviousness. *Journal of Educational Psychology, 87*(3), 504.
Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 345–353.
Yang, H., & Thompson, C. (2010). Nurses' risk assessment judgements: A confidence calibration study. *Journal of Advanced Nursing, 66*, 2751–2760.
Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*.

**Jieying Chen** is an assistant professor at the Department of Business Administration, University of Manitoba. Her research focuses on judgment and decision-making, cross-cultural interactions, strategic human resource management, and mindfulness.

**Lok Ching (Roxane) Kwan**, **Lok Yeung (Loren) Ma, Hiu Yee (HayleyAnne) Choi, Ying Ching (Lita) Lo, Shin Yee (Sarah) Au, and Chi Ho (Toby) Tsang** were students at the University of Hong Kong during the academic year 2018-9.

**Bo Ley Cheng** was the teaching assistant at the University of Hong Kong psychology department during the academic year 2018–9.

**Gilad Feldman** is an assistant professor with the University of Hong Kong psychology department. His research focuses on judgment and decision-making.

# Hindsight Bias: Replication and extension Supplementary

## Project process outline

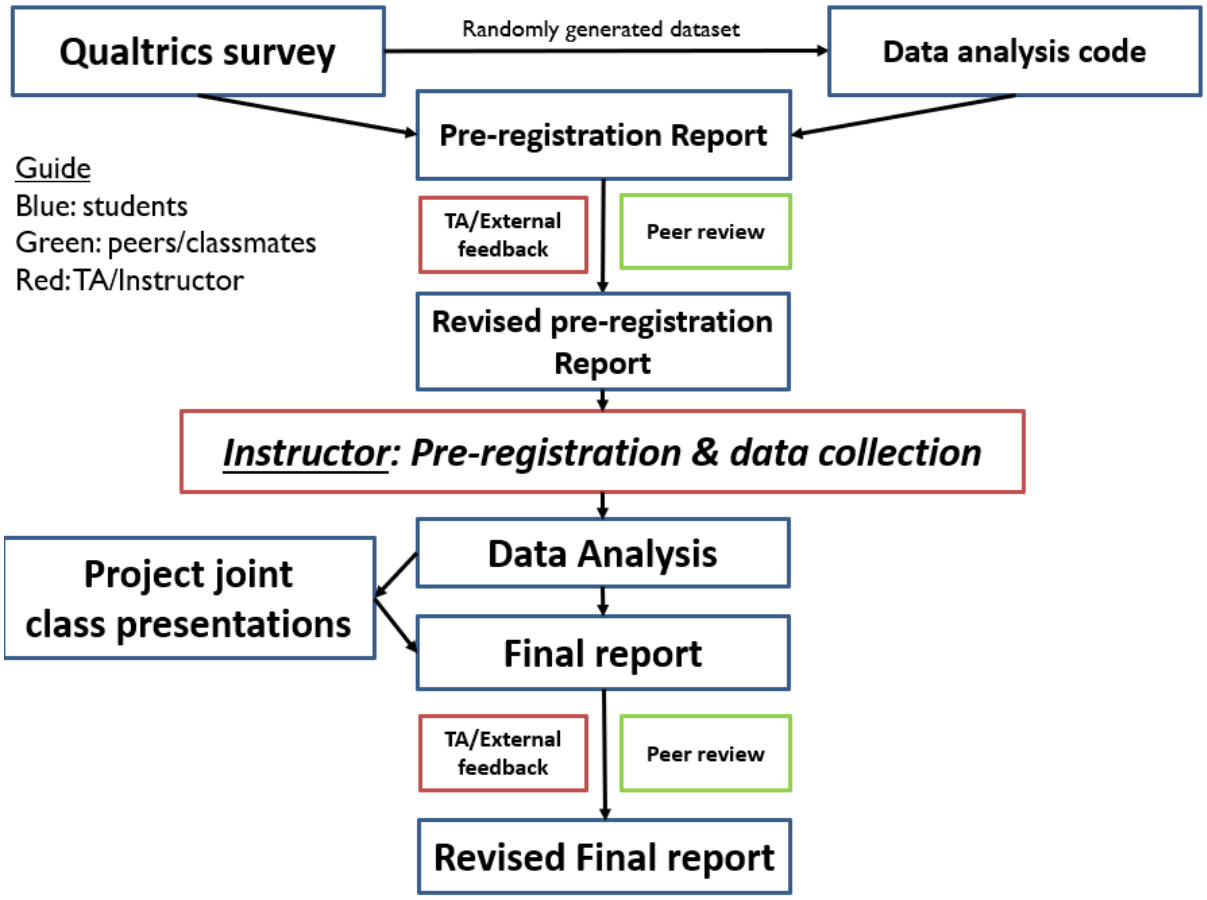
The current replication is part of the mass pre-registered replication project, with the aim of revisiting well-known research findings in the area of social psychology and judgment and decision making (JDM) and examining the reproducibility and replicability of these findings.

The project outline is shown in Figure S1. For each of the replication projects, researchers completed full pre-registrations, data analysis, and APA style submission ready reports. Authors independently reproduced the materials and designed the replication experiment, with a separate pre-registration document. The researchers then peer-reviewed one another to try and arrive at the best possible design. Then, the lead and corresponding authors reviewed the integrated work and the last corresponding author made final adjustments and conducted the pre-registration and data collection.

The OSF page of the project contains one Qualtrics survey design used for the data collection, and pre-registration documents submitted by each of the researchers. In the manuscript, unless otherwise noted, we followed the most conservative of the pre-registrations.

Figure S1

*Project Process Outline*



# Study 1

## Study 1: Transparency report

### PREREGISTRATION SECTION

(1) Prior to analyzing the complete data set, a time-stamped preregistration was posted in an independent, third-party registry for the data analysis plan. **Yes**

(2) The manuscript includes a URL to all preregistrations that concern the present study. **Yes**

(3) The study was preregistered… **before any data were collected**

**The preregistration fully describes…**

(4) all inclusion and exclusion criteria for participation (e.g., English speakers who achieved a certain cutoff score in a language test). **Yes**

(5) all procedures for assigning participants to conditions. **Yes**

(6) all procedures for randomizing stimulus materials. **Yes**

(7) any procedures for ensuring that participants, experimenters, and data-analysts were kept naive (blinded) to potentially biasing information. **Yes**

(8) a rationale for the sample size used (e.g., an a priori power analysis). **Yes**

(9) the measures of interest (e.g., friendliness). **Yes**

(10) all operationalizations for the measures of interest (e.g., a questionnaire measuring friendliness). **Yes**

(11) the data preprocessing plans (e.g., transformed, cleaned, normalized, smoothed). **Yes**

(12) how missing data (e.g., dropouts) were planned to be handled. **Yes**

(13) the intended statistical analysis for each research question (this may require, for example, information about the sidedness of the tests, inference criteria, corrections for multiple testing, model selection criteria, prior distributions etc.). **Yes**

**Comments about your Preregistration**

No comments.

---

## METHODS SECTION

**The manuscript fully describes…**

(14) the rationale for the sample size used (e.g., an a priori power analysis). **Yes**

(15) how participants were recruited. **Yes**

(16) how participants were selected (e.g., eligibility criteria). **Yes**

(17) what compensation was offered for participation. **No**

(18) how participant dropout was handled (e.g., replaced, omitted, etc.). **Yes**

(19) how participants were assigned to conditions. **Yes**

(20) how stimulus materials were randomized. **Yes**

(21) whether (and, if so, how) participants, experimenters, and data-analysts were kept naive to potentially biasing information. **NA**

(22) the study design, procedures, and materials to allow independent replication. **Yes**

(23) the measures of interest (e.g., friendliness). **Yes**

(24) all operationalizations for the measures of interest (e.g., a questionnaire measuring friendliness). **Yes**

(25) any changes to the preregistration (such as changes in eligibility criteria, group membership cutoffs, or experimental procedures)? **Yes**

**Comments about your Methods section**

No comments.

---

## RESULTS AND DISCUSSION SECTION

**The manuscript…**

(26) distinguishes explicitly between “confirmatory” (i.e., prespecified) and “exploratory” (i.e., not prespecified) analyses. **Yes**

(27) describes how violations of statistical assumptions were handled. **Yes**

(28) justifies all statistical choices (e.g., including or excluding covariates; applying or not applying transformations; use of multi-level models vs. ANOVA). **Yes**

(29) reports the sample size for each cell of the design. **Yes**

(30) reports how incomplete or missing data were handled. **Yes**

(31) presents protocols for data preprocessing (e.g., cleaning, discarding of cases and items, normalizing, smoothing, artifact correction). **Yes**

**Comments about your Results and Discussion**

No comments.

---

## DATA, CODE, AND MATERIALS AVAILABILITY SECTION

**The following have been made publicly available…**

(32) the (processed) data, on which the analyses of the manuscript were based. **Yes**

(33) all code and software (that is not copyright protected). **Yes**

(34) all instructions, stimuli, and test materials (that are not copyright protected). **Yes**

(35) Are the data properly archived (i.e., would a graduate student with relevant background knowledge be able to identify each variable and reproduce the analysis)? **Yes**

(36) The manuscript includes a statement concerning the availability and location of all research items, including data, materials, and code relevant to the study. **Yes**

**Comments about your Data, Code, and Materials**

No comments.

## Study 1: Relationship between Experiments 1 and 2 of Fischhoff (1975)

While we chose to replication Experiment 2 of Fischhoff (1975), we would like to draw our readers' attention to the association between Experiment 1 and Experiment 2 in Fischhoff (1975). Experiment 1 used a similar design and identical scenarios compared to Experiment 2, with the exception that the instructions in Experiment 1 did not ask the participants to ignore the outcome knowledge when providing estimation. Because the instruction to ignore (i.e., respond as you would have had you not known the outcome) is a critical condition for establishing hindsight bias, we chose to replicate Experiment 2 but not Experiment 1 of Fischhoff (1975). Importantly, in Experiment 2 of Fischhoff (1975), the data for the Before condition were obtained from Experiment 1. The total sample for Experiment 2 ($n$ = 172) included 92 participants in the Before condition who responded to one of the four events in Experiment 1 and 80 participants for in the After (ignore) conditions who responded to all four events.

## Study 1: Power analysis

### Original sample size and p-values

As described earlier, the total sample for Experiment 2 ($n$ = 172) included 92 participants in the Before condition who responded to one of the four events and 80 participants for in the After (ignore) conditions who responded to all four events. Based on the information provided in Table 3 of the original article, the average sample size per condition was (17+20+15+18+18+39+17+21+20+20+19+19+19+15+20+17+17+18+20+18)/20=19.35 ≈ 20 (rounded).

We used the $p$ value of .001 as an estimate of the $p$ value for the probability estimates in the Mann-Whitney U test. This is based on the original result that the Before-After (ignore) difference of probability estimates (9.2%, $p$ value unreported) in Experiment 2 of Fischhoff (1975) was not significantly different from the Before-After difference (10.8%, $p < .001$) in Experiment 1 of Fischhoff (1975) ($p > .10$, Mann-Whitney $U$ test).

### Estimation of effect size in the original study

The effect size of the original experiment was estimated using the following procedure in Mavis (Hamilton, Aydin, & Mizumoto, 2016).

- Estimated number of participants of treatment group = 20
- Estimated number of participants of comparison group = 20
- Estimated $p$-value in the original study = 0.001
- **Cohen's $d$ = 1.13, 95% CI [0.44 , 1.82]**

### Calculation of the minimum sample size required

We estimated the minimum sample size required using the effect size of $d$ = 1.13, power of .95, and significance level of .05 (two-tailed) in G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009). The minimum sample size required for each comparison was 23 + 23 = 46. Because there are 4 comparisons for each event, and 4 events in total, the total required sample size was 46 * 4 * 4 = 736.

Table S1

*Study 1: Sample Size Calculation*

| **t tests**- Means: Wilcoxon-Mann-Whitney test (two groups) | | |
|---|---|---|
| **Options:** | A.R.E. method | |
| **Analysis**: | A priori: Compute required sample size | |
| **Input**: | Tail(s) | = Two |
| | Parent distribution | = Normal |
| | Effect size d | = 1.13 |
| | α err prob | = 0.05 |
| | Power (1-βerr prob) | = 0.95 |
| | Allocation ratio N2/N1 | = 1 |
| **Output**: | Noncentrality parameter δ | = 3.7446657 |
| | Critical t | = 2.0181861 |
| | Df | = 41.9267643 |
| | Sample size group 1 | = 23 |
| | Sample size group 2 | = 23 |
| | Total sample size | = 46 |
| | Actual power | = 0.9552109 |

## Study 1: Changes made after the pre-registration

1. Exploratory hypotheses: While we proposed to test the exploratory hypotheses using basic statistical methods such Mann-Whitney *U* tests and correlations in the pre-registration, we added more sophisticated analyses of mediating and moderating effects in the manuscript.
2. In consultation with the original lead author and the editor, we removed the descriptions of the stimuli of Events C and D, as the stimuli contain problematic descriptions of the person undergoing therapy. We also removed the results of Events C and D. We jointly strongly believe that Events C and D should no longer be used in future research.

## Study 1: Materials and scales used in the replication + extension experiment

**Event A**

**British-Gurka struggle**

For some years after the arrival of Hastings as governor-general of India, the consolidation of British power involved serious war. The first of these wars took place on the northern frontier of Bengal where the British were faced by the plundering raids of the Gurkas of Nepal. Attempts had been made to stop the raids by an exchange of lands, but the Gurkas would not give up their claims to country under British control, and Hastings decided to deal with them once and for all. The campaign began in November, 1814. It was not glorious. The Gurkas were only some 12,000 strong; but they were brave fighters, fighting in territory well-suited to their raiding tactics. The older British commanders were used to war in the plains where the enemy ran away from a resolute attack. In the mountains of Nepal it was not easy even to find the enemy. The troops and transport animals suffered from the extremes of heat and cold, and the officers learned caution only after sharp revers. Major-General Sir D. Octerlony was the one commander to escape from these minor defeats.

Outcome: (Not provided) / British resulted in victory. / Gurka resulted in victory. / The two sides reached a military stalemate, but were unable to come to a peace settlement. / Outcome: The two sides reached a military stalemate and came to a peace settlement.

(from The Age of Reform by E.L. Woodward, Oxford, 1938, pp. 393-394)

**Event B**

**Atlanta**

On Saturday, June 17, 1967, the same type of minor arrest that had initiated the Cincinnati race riot took place in Atlanta. On the 18th, an African American youth was superficially wounded by a police officer in a scuffle following his refusal to stop after short-circuiting a burglar alarm in the Dixie Hills Shopping Center. A decision was made by Dixie Hills residents (all black) to organize committees and hold a protest meeting the night of the second incident. Approximately 250 people were present at the meeting. When a number of African American leaders urged the submission of a petition of grievances through legal channels, the response was lukewarm. When Stokely Carmichael (leader of the militant black power organization SNCC) took the podium, the response was tumultuous. The press reported him as saying, "It's not a question of law and order. We are concerned with the liberation of black people. We have to build a revolution." As the people present at the meeting poured into the street, they were joined by others. The crowd soon numbered 1,000. From alleys and rooftops, rocks and bottles were thrown at the nine police officers on the scene. Windows of the police cars were broken. Firecrackers exploded in the darkness. The police believed that they had been fired on. Reinforced by approximately 60 to 70 officers, the police began firing over the heads of the crowd.

Outcome: (Not provided) / The crowd dispersed after that and there was no outbreak of violence. / The crowd dispersed after that and there were outbreaks of violence in several other places in town./ The crowd refused to disperse after that and there were no further actions that led to any outbreak of violence. / The crowd refused to disperse and there was an outbreak of violence.

(from National Advisory Commission on Civil Disorders Report, Bantam, 1968, pp. 28-30.)

**Event C**

**Mrs. Dewar in therapy**

[*NOTE: in consultation with the original lead author and the editor we removed the descriptions of the stimuli of Event C, as the stimuli contain problematic descriptions of the person undergoing therapy. We jointly strongly believe that this set of stimuli should no longer be used in future research.*]

Outcome: (Not provided) / Mrs. Dewar terminated the therapy, and experienced no improvement in her condition. / Mrs. Dewar terminated the therapy, and experienced improvement in her condition. / Mrs. Dewar continued the therapy, and experienced no improvement in her condition. / Mrs. Dewar continued the therapy, and experienced improvement in her condition.

(from A. Ellis, Psychosexual and marital problems. in L.A. Berg & L.A. Pennington. An Introduction to Clinical Psychology. Ronald Press, 1966, p. 262-3)

**Event D**

**George in therapy**

[*NOTE: in consultation with the original lead author and the editor we removed the descriptions of the stimuli of Event D, as the stimuli contain problematic descriptions of the person undergoing therapy. We jointly strongly believe that this set of stimuli should no longer be used in future research.*]

Outcome: (Not provided) / George continued the therapy, and showed no improvement. / George continued the therapy, and improvement was shown. / George terminated the therapy, and experienced no improvement. / George terminated the therapy, and showed improvement.

(from A. Ellis, Psychosexual and marital problems. in L.A. Berg & L.A. Pennington. An Introduction to Clinical Psychology. Ronald Press, 1966, p. 264)

## Study 1: Sample characteristics

Most of the participants (*n* = 442, 99.10%) were born in the United States, and the others were born in countries such as Jamaica, Japan, and Nigeria. When asked about their family's social class, 14 participants self-identified as lower class (3.17%), 103 as working class (23.30%), 78 as lower middle class (17.65%), 216 as middle class (48.87%), 29 as upper middle class (6.56%), and two as upper class (0.45%).

## Study 1: Additional analyses

### Study 1: Results with the sample after applying exclusion criteria

In the preregistrations, we proposed to analyze the data using a sample after applying a set of pre-specified exclusion criteria:

- All participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale)
- Participants who were not serious about filling in the survey (self-report < 4, on a 1-5 scale).
- Participants who correctly guessed the hypothesis of this study in the funneling section. We screened all participants' answers to the purpose of the study, and removed all cases that mentioned hindsight bias or how outcome knowledge could affect probability estimates or surprise.

The results are summarized in Table S2, and are highly similar to those obtained using the full sample. Readers interested in reproducing the analyses can visit our OSF webpage, look for files Event A.omv, Event B.omv, and use the filter function in JAMOVI to conduct the analyses. JAMOVI is a free, open-source software.

Table S2

*Study 1: Mann-Whitney U Tests of Probability Estimates Difference and Surprising Ratings Difference between Before and After (Ignore) Conditions (After Applying Exclusion Criteria, n = 848)*

| | **Probability Estimates** | | | **Surprise Ratings** | | |
|---|---|---|---|---|---|---|
| | ***U*** | ***p*** | ***Cohen's d*** | ***U*** | ***p*** | ***Cohen's d*** |
| Event A Outcome 1 | 462 | < .001 | 1.002 | 665 | .011 | -.056 |
| Event A Outcome 2 | 730.5 | 0.236 | 0.225 | 836 | .828 | .047 |
| Event A Outcome 3 | 611 | 0.033 | 0.429 | 837 | .992 | .024 |
| Event A Outcome 4 | 546.5 | 0.015 | 0.508 | 766 | .776 | .019 |
| Event B Outcome 1 | 429 | < .001 | 1.013 | 696 | .008 | -.461 |
| Event B Outcome 2 | 417 | < .001 | 0.966 | 971.5 | .886 | .063 |
| Event B Outcome 3 | 543 | < .001 | 0.705 | 969 | .722 | -.053 |
| Event B Outcome 4 | 768 | 0.048 | 0.440 | 998 | .902 | -.017 |

*Note*. Following discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

## Study 1: Violin plots of probability estimates

Figure S2a-d

*Violin Plots of Probability Estimates in Before and After Conditions (Event A)*

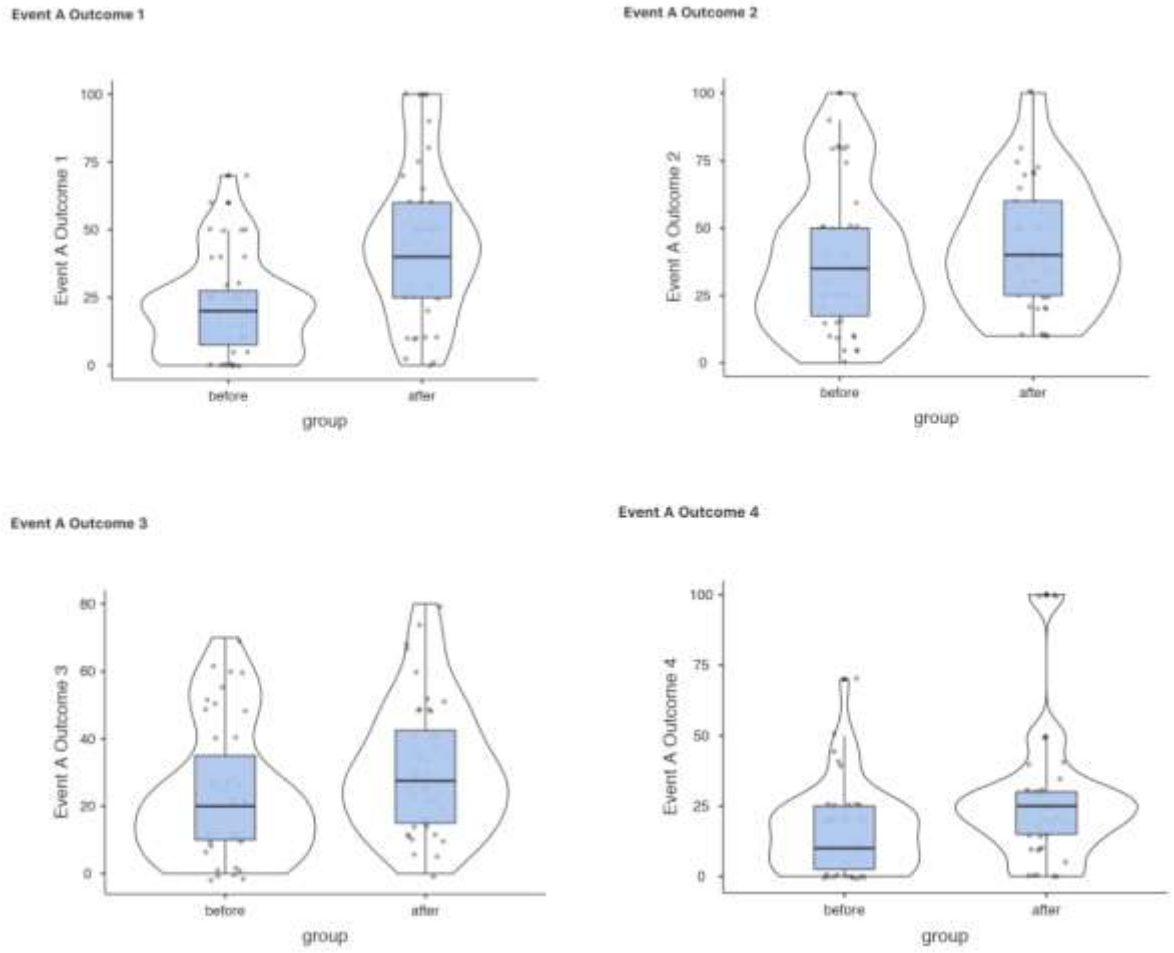
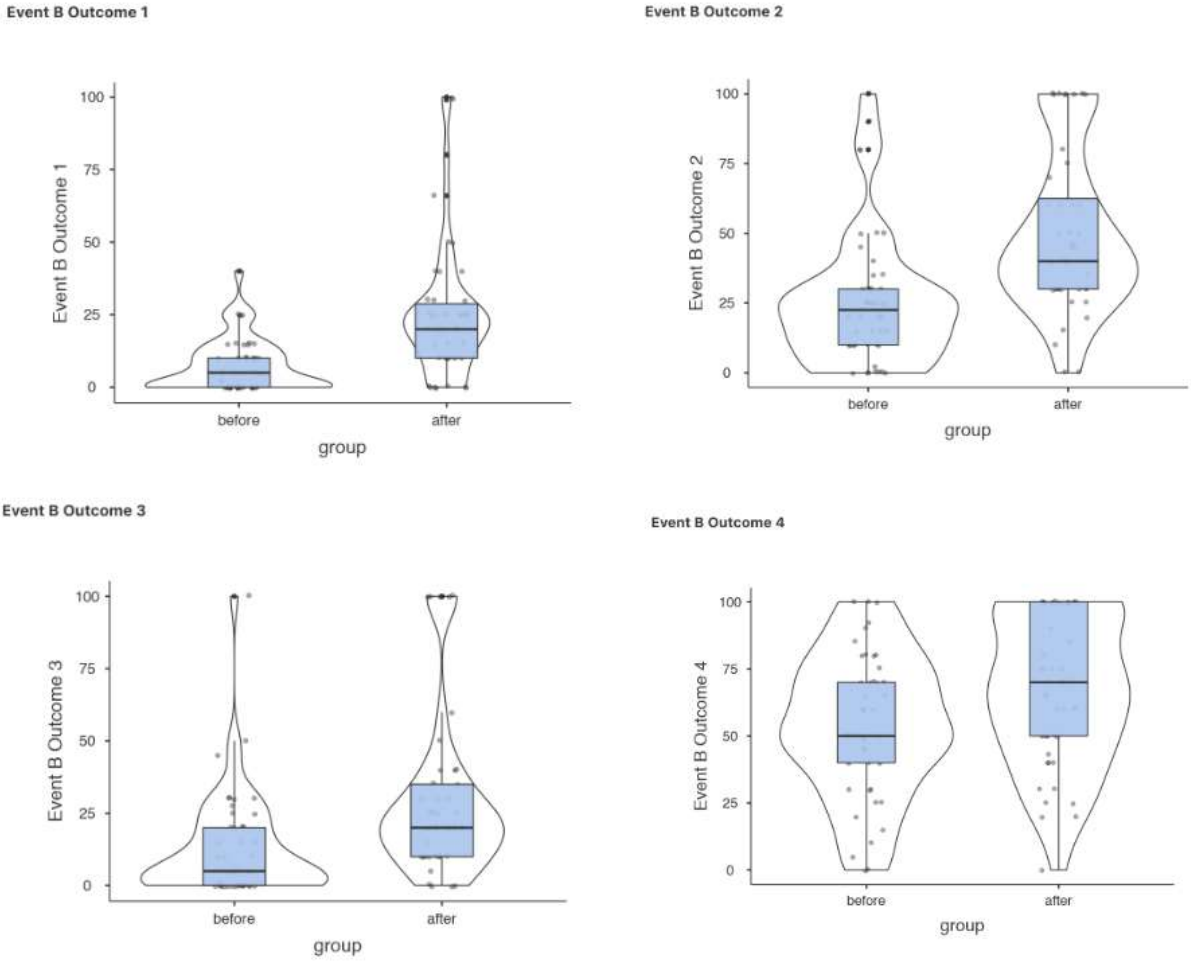


Figure S3a-d

*Violin Plots of Probability Estimates in Before and After Conditions (Event B)*



## Study 1: Violin plots of surprise ratings

Figure S4a-d

*Violin Plots for Surprise between Before and After Condition (Event A)*

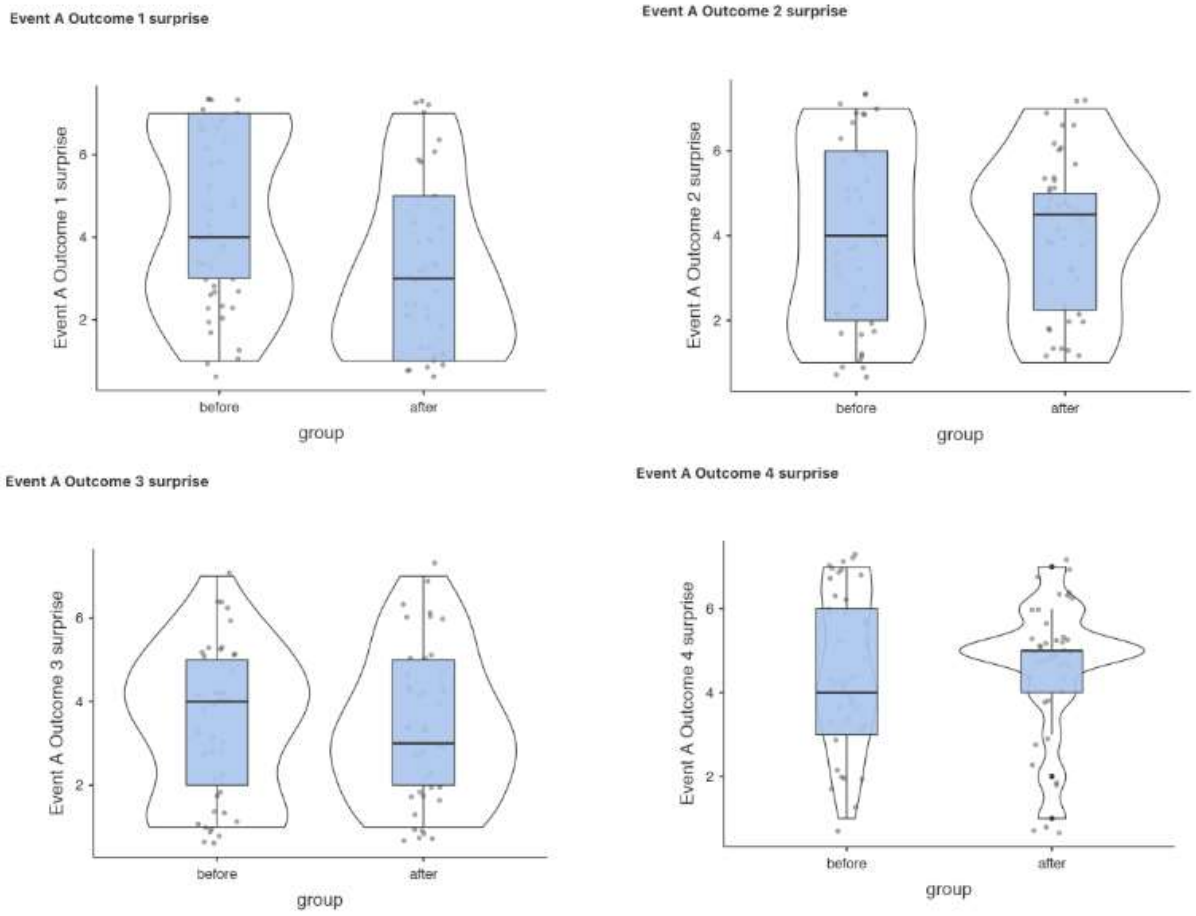
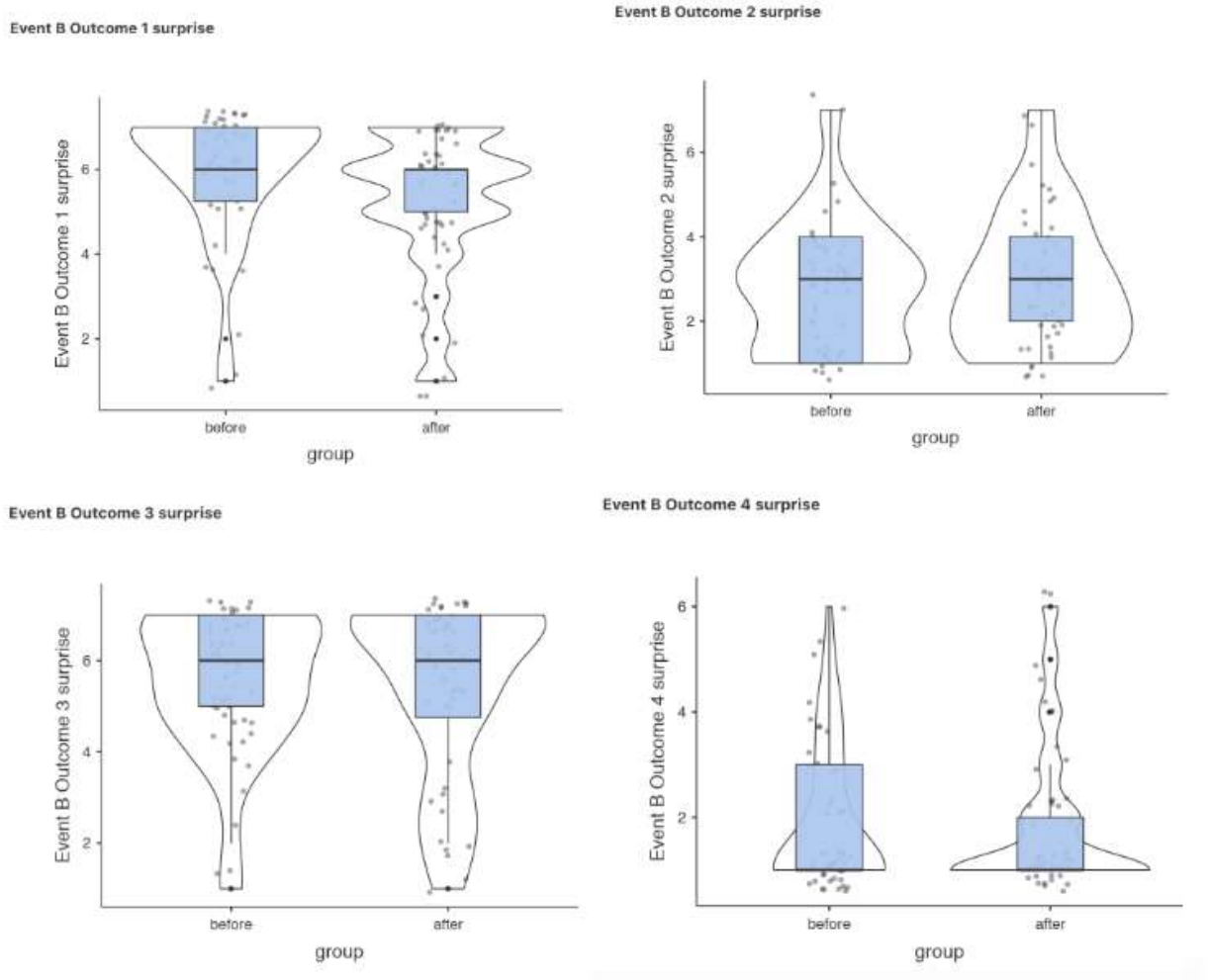


Figure S5a-d

*Violin Plots for Surprise between Before and After Condition (Event B)*



## Study 1: Codes for calculating confidence intervals of $\phi$ of probability estimates

We used the R package asht (Fay, 2017) and the following codes to calculate the confidence intervals of the effect size $\phi$, which reflects the probability that a score in the hindsight condition was higher than that in the foresight condition.

```
library(readxl)
f1975data <- read_excel("20189PSYCINCFischhoff1975_r input.xlsx", na = "#NULL!")
library(asht)
# wmwTest defaults are: two-sided, 95% confidence interval under the proportional odds assumption with continuity correction (Fay & Malinovsky, 2010; Newcombe, 2006)

# Event A Outcome 1
xa1 <- f1975data$f1975_EventA_bfor_pr_1
xa1 <- xa1[!is.na(xa1)]
ya1 <- f1975data$f1975_EventA_aft1_pr_1
ya1 <- ya1[!is.na(ya1)]
wmwTest(xa1,ya1)
# Event A Outcome 2
xa2 <- f1975data$f1975_EventA_bfor_pr_2
xa2 <- xa2[!is.na(xa2)]
ya2 <- f1975data$f1975_EventA_aft2_pr_2
ya2 <- ya2[!is.na(ya2)]
wmwTest(xa2,ya2)
```

```
# Event A Outcome 3
xa3 <- f1975data$f1975_EventA_bfor_pr_3
xa3 <- xa3[!is.na(xa3)]
ya3 <- f1975data$f1975_EventA_aft3_pr_3
ya3 <- ya3[!is.na(ya3)]
wmwTest(xa3,ya3)
# Event A Outcome 4
xa4 <- f1975data$f1975_EventA_bfor_pr_4
xa4 <- xa4[!is.na(xa4)]
ya4 <- f1975data$f1975_EventA_aft4_pr_4
ya4 <- ya4[!is.na(ya4)]
wmwTest(xa4,ya4)

# Event B Outcome 1
xa1 <- f1975data$f1975_EventB_bfor_pr_1
xa1 <- xa1[!is.na(xa1)]
ya1 <- f1975data$f1975_EventB_aft1_pr_1
ya1 <- ya1[!is.na(ya1)]
wmwTest(xa1,ya1)
# Event B Outcome 2
xa2 <- f1975data$f1975_EventB_bfor_pr_2
xa2 <- xa2[!is.na(xa2)]
ya2 <- f1975data$f1975_EventB_aft2_pr_2
```

```
ya2 <- ya2[!is.na(ya2)]

wmwTest(xa2,ya2)

# Event B Outcome 3

xa3 <- f1975data$f1975_EventB_bfor_pr_3

xa3 <- xa3[!is.na(xa3)]

ya3 <- f1975data$f1975_EventB_aft3_pr_3

ya3 <- ya3[!is.na(ya3)]

wmwTest(xa3,ya3)

# Event B Outcome 4

xa4 <- f1975data$f1975_EventB_bfor_pr_4

xa4 <- xa4[!is.na(xa4)]

ya4 <- f1975data$f1975_EventB_aft4_pr_4

ya4 <- ya4[!is.na(ya4)]

wmwTest(xa4,ya4)
```

## Study 1: Codes for calculating confidence intervals of Cohen's *d*

We used the R package psych (Revelle, 2019) and the following codes to calculate the confidence intervals of the effect size Cohen's *d*.

```
library(psych)

# Probability Estimates

cohen.d.ci(d = 1.002, n1 = 43, n2 = 45, alpha = .05)

cohen.d.ci(d = 0.199, n1 = 43, n2 = 42, alpha = .05)

cohen.d.ci(d = 0.41, n1 = 43, n2 = 44, alpha = .05)

cohen.d.ci(d = 0.539, n1 = 43, n2 = 43, alpha = .05)

cohen.d.ci(d = 1.022, n1 = 46, n2 = 46, alpha = .05)

cohen.d.ci(d = 0.907, n1 = 46, n2 = 44, alpha = .05)

cohen.d.ci(d = 0.705, n1 = 46, n2 = 44, alpha = .05)

cohen.d.ci(d = 0.45, n1 = 46, n2 = 45, alpha = .05)


# Surprise Ratings

cohen.d.ci(d = -0.556, n1 = 43, n2 = 45, alpha = .05)

cohen.d.ci(d = 0.07, n1 = 43, n2 = 42, alpha = .05)

cohen.d.ci(d = -0.005, n1 = 43, n2 = 44, alpha = .05)

cohen.d.ci(d = 0.041, n1 = 43, n2 = 43, alpha = .05)

cohen.d.ci(d = -0.44, n1 = 46, n2 = 46, alpha = .05)

cohen.d.ci(d = 0.079, n1 = 46, n2 = 44, alpha = .05)

cohen.d.ci(d = -0.053, n1 = 46, n2 = 44, alpha = .05)

cohen.d.ci(d = -0.0322, n1 = 46, n2 = 45, alpha = .05)
```

## Study 1: Student's independent samples t-tests of probability estimates

The results of Student's independent samples are mostly consistent with those by independent Mann-Whitney U tests (see Table S3). Exceptions is Event A Outcome 3, for which the p values of the Mann-Whitney U tests were significant, but the p values of Student's independent samples t-tests were marginally significant.

Table S3

*Study 1: Student's Independent Samples T-Test of Probability Estimates Difference between Before and After (ignore) Conditions*

| | | *t* | df | *p* | Mean difference | SE difference | Cohen's *d* |
|---|---|---|---|---|---|---|---|
| Event A Outcome 1 | Student's t | 4.70 [a] | 86.00 | < .001 | -24.12 | 5.13 | 1.00 |
| Event A Outcome 2 | Student's t | 0.92 | 83.00 | 0.361 | -5.01 | 5.46 | 0.20 |
| Event A Outcome 3 | Student's t | 1.91 | 85.00 | 0.059 | -8.10 | 4.24 | 0.41 |
| Event A Outcome 4 | Student's t | 2.50 | 84.00 | 0.014 | -10.88 | 4.36 | 0.54 |
| Event B Outcome 1 | Student's t | 4.90 [a] | 90.00 | < .001 | -17.98 | 3.67 | 1.02 |
| Event B Outcome 2 | Student's t | 4.30 | 88.00 | < .001 | -24.11 | 5.60 | 0.91 |
| Event B Outcome 3 | Student's t | 3.34 [a] | 88.00 | 0.001 | -16.86 | 5.04 | 0.71 |
| Event B Outcome 4 | Student's t | 2.15 | 89.00 | 0.035 | -12.24 | 5.71 | 0.45 |

[a] Levene's test is significant ($p < .05$), suggesting a violation of the assumption of equal variances.

## Study 1: Student's independent samples t-tests of surprise ratings

Although surprise ratings were on a 7-point Likert scale, we report results of Mann-Whitney U tests rather than that of independent t-tests for two reasons: first, the Mann-Whitney U test was the test that we proposed in the pre-registration; second, the distribution of the surprise ratings was non-normal. Results of independent t-tests with surprise ratings are available in Supplementary Materials.

Here we also report the results of Student's independent samples t-tests of surprise ratings (see Table S4).

Table S4

*Study 1: Student's Independent Samples T-Test of Surprise Rating Difference between Before and After (ignore) Conditions*

| | | statistic | df | *P* | Mean difference | SE difference | Cohen's d |
|---|---|---|---|---|---|---|---|
| Event A Outcome 1 | Student's t | -2.61 | 86.00 | 0.011 | -1.15 | 0.44 | -0.56 |
| Event A Outcome 2 | Student's t | 0.32 | 83.00 | 0.748 | 0.14 | 0.44 | 0.07 |
| Event A Outcome 3 | Student's t | -0.03 | 85.00 | 0.980 | -0.01 | 0.38 | -0.01 |
| Event A Outcome 4 | Student's t | 0.19 [a] | 84.00 | 0.850 | 0.07 | 0.37 | 0.04 |
| Event B Outcome 1 | Student's t | -2.11 | 90.00 | 0.038 | -0.72 | 0.34 | -0.44 |
| Event B Outcome 2 | Student's t | 0.37 | 88.00 | 0.709 | 0.13 | 0.34 | 0.08 |
| Event B Outcome 3 | Student's t | -0.25 | 88.00 | 0.803 | -0.09 | 0.37 | -0.05 |
| Event B Outcome 4 | Student's t | -0.15 | 89.00 | 0.878 | -0.05 | 0.30 | -0.03 |

[a] Levene's test is significant ($p < .05$), suggesting a violation of the assumption of equal variances.

## Study 1: Other tests on surprise ratings

### Compare to midpoint

In one version of the pre-registration, we proposed to examine whether the mean level of surprise in the After (ignore) condition was lower than 4 (the median of the 7-point Likert rating scale). A one-sample t-test suggests that the mean ratings of surprise of the After (control) group was not significantly different from the mean of the rating scale 4 (*mean* = 3.83, *SD* = 2.06, $t(352) = -1.55$, $p = .12$, 95% CI of mean difference = [-.39, .05]). Because this test cannot rule out the influence of the extent to which the four events themselves are highly or lowly surprising, we later decided to move this analysis to Supplementary Materials.

### Correlation

In the initial pre-registration, as an exploratory hypothesis, we proposed that the correlation between probability estimates and surprise ratings would be negative in the After (ignore) conditions. We later decided to test the mediating effect of surprise, rather than looking at correlation coefficients alone. In this Supplementary Materials, we provide the results of the correlation tests. Importantly, on top of our initial exploratory hypothesis, we would like to stress that the negative correlation between probability estimates and surprise ratings should hold in both the Before condition and the After (ignore) condition. That is, regardless of the experimental condition, participants who experience a lower degree of surprise are likely to estimate the probability of the actual outcome to be lower.

We estimated the Spearman's rho and Pearson's r of the relationship between probability estimates and surprise ratings in the After (ignore) conditions. The results are shown in Table S5.

For all four events, there was a significant, negative correlation between probability estimates and levels of surprise (Event A: rho = -.53, $p < .001$; Event B: rho = -.66, $p < .001$).

Table S5

*Correlation between Probability Estimates and Surprise Ratings in the After (Ignore) Conditions*

| | | **Event A** | **Event B** |
|---|---|---|---|
| Level of Surprise - After (ignore) Conditions | Spearman's *rho* | -.53 | -.66 |
| | *p*-value | < .001 | < .001 |
| | Pearson's *r* | -0.51 | -0.61 |
| | *p*-value | < .001 | < .001 |
| | 95% CI Upper | -0.39 | -0.51 |
| | 95% CI Lower | -0.61 | -0.70 |

As mentioned earlier, theoretically, probability estimates and surprise ratings shall also be negatively correlated in the Before condition. That is, in the Before condition, participants who experienced less surprise would predict the likelihood to be higher, whereas those who experienced more surprise would predict the likelihood to be lower. The results largely supported this proposition (Event A: *rho* = -.43, $p < .001$; Event B: *rho* = -.62, $p < .001$; see Table S6).

Table S6

*Correlation between Probability Estimates and Surprise Ratings in the Before Condition*

| | | Event A | Event B |
|---|---|---|---|
| Level of Surprise - Before Condition | Spearman's rho | -0.43 | -.62 |
| | p-value | < .001 | < .001 |
| | Pearson's r | -0.40 | -0.62 |
| | p-value | < .001 | < .001 |
| | 95% CI Upper | -0.27 | -0.52 |
| | 95% CI Lower | -0.52 | -0.70 |

**Mediation**

To test Hypotheses 2(b) and 2(c), we used data from participants' responses for both events and all four outcomes. Because of the characteristics of our experimental design, participants in the After (ignore) conditions each responded to one event only. Participants in the Before condition responded to both events. We acknowledge that the data in the two events in the Before condition are nested within individuals, and used a dummy variables (Event A) to account for the confounding variance introduced by the event settings. We tested whether surprise mediates or moderates the relationship between outcome knowledge and probability estimates using the PROCESS macro for SPSS (Hayes, 2017).

Using the mediation model (Model 4 in PROCESS), controlling for Event (Event A versus Event B), we did not find support for the mediating effect of surprise in the relationship between outcome knowledge and probability estimates (indirect effect: $B = 1.48$, *boot-strapped S.E.* = 1.09, 95% CI [-0.66, 3.65]). Hypothesis 2(b) was not supported. Details of the regressions testing Hypothesis 2(b) are shown in Table S7.

Table S7

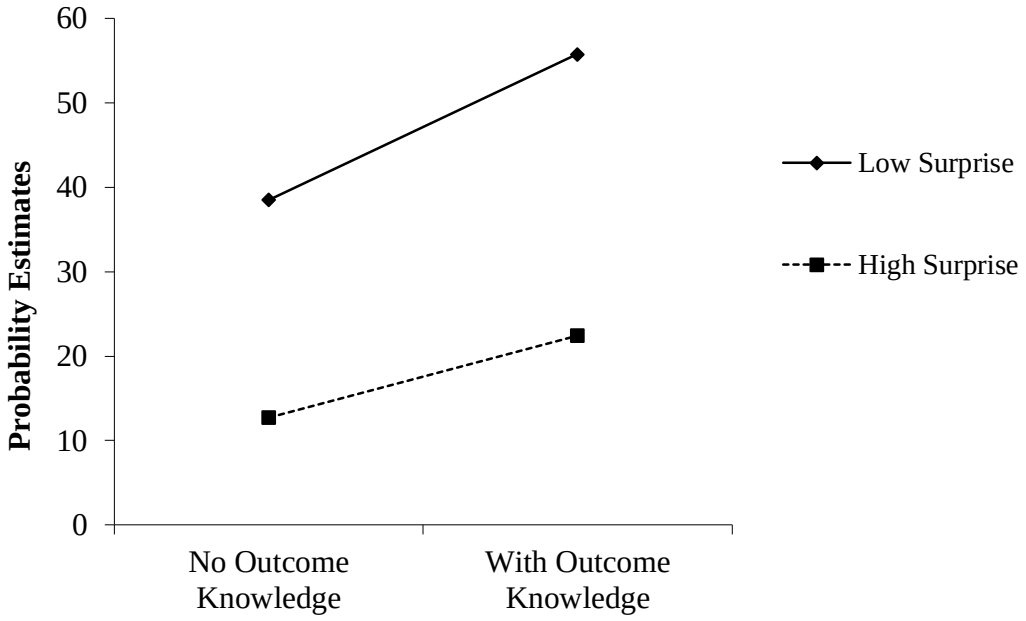
*Study 1: Regressions of Surprise Ratings and Probability Estimates*

| | *B* | *S.E.* | *p* | 95% CI Lower | 95% CI Upper |
|---|---|---|---|---|---|
| Dependent variable: Surprise | | | | | |
| constant | 4.04 | 0.14 | 0.00 | 3.77 | 4.30 |
| Outcome knowledge | -0.21 | 0.16 | 0.18 | -0.52 | 0.10 |
| EventA | 0.01 | 0.16 | 0.96 | -0.30 | 0.32 |
| Dependent variable: Probability estimates | | | | | |
| Constant | 54.54 | 2.17 | 0.00 | 50.29 | 58.79 |
| Outcome knowledge | 13.48 | 1.69 | 0.00 | 10.17 | 16.79 |
| Surprise | -6.98 | 0.40 | 0.00 | -7.76 | -6.19 |
| EventA | -2.77 | 1.68 | 0.10 | -6.08 | 0.53 |

**Moderation**

Using the moderation model (Model 1 in PROCESS) with mean-centering, controlling for Event, we found support for a moderating effect of surprise in the relationship between outcome knowledge and probability estimates ($B$ = -1.79, *S.E.* =0.80, $p$ = .03, 95% CI [-3.36, -0.22]; see Table S8, Model 1). Specifically, the relationship between outcome knowledge and probability estimates were stronger when surprise was lower (simple slope: $B$ = 17.25, *S.E.* = 2.38, $p$ < .001) rather than higher (simple slope: $B$ = 9.70, *S.E.* = 2.38, $p$ < .001; see Figure S6). However, in our original analysis when all four events were included, and when Events A, B, and C were controlled for, we did not find support for the moderating effect of surprise in the relationship between outcome knowledge and probability estimates ($B$ = -0.31, *S.E.* = 0.57, $p$ = .59, 95% CI [-1.44, 0.81]). Note that in this study, the moderator was measured but not experimentally manipulated, as we aimed to stay very close to the design of the target article. In sum, we found support for Hypothesis 2(c) when considering Events A and B only, yet we found no support for Hypothesis 2(c) in our original analysis when all four events were considered. While we have decided to remove any results related to Events C and D, which is a deliberate deviation from the preregistration, we would like to caution our readers about the conflicting findings of the moderating effect of surprise in Study 1 when different events were included in the analysis.

Figure S6
*Study 1: Moderation Effect of Surprise*




*Note.* This analysis contained data for Events A and B.

We also tested whether there was a curvilinear interaction effect between outcome knowledge and surprise, and the results were nonsignificant (see Table S8, Model 2).

Table S8

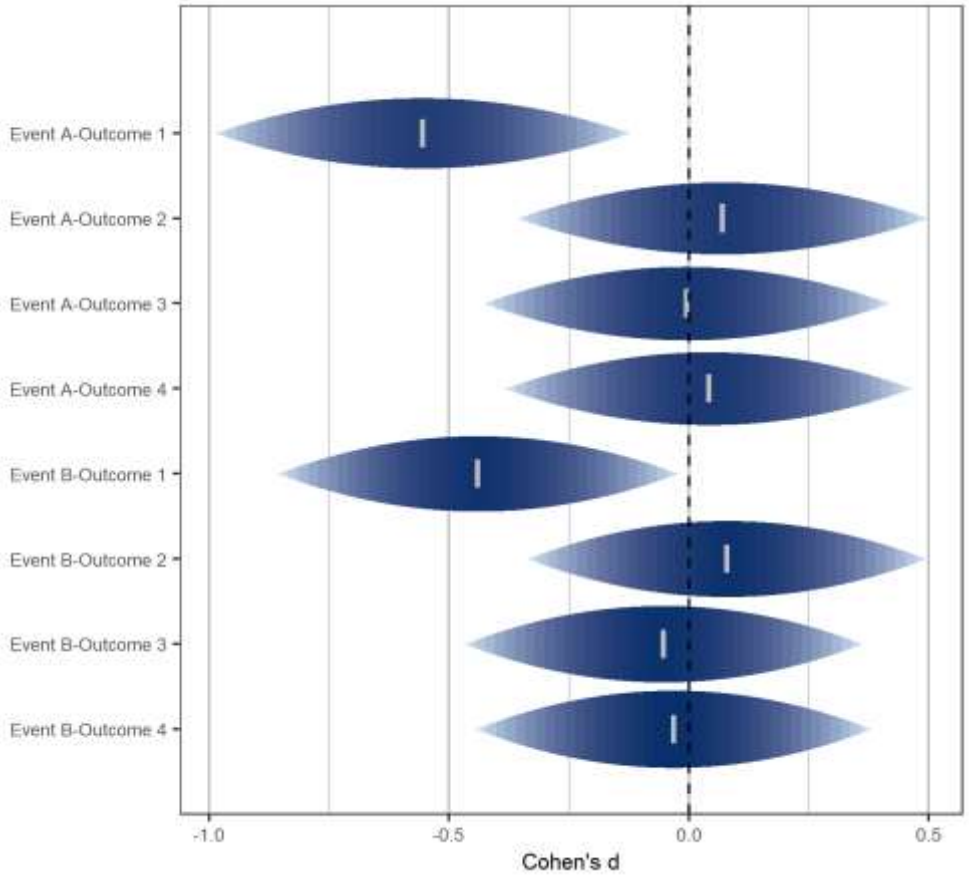
*Study 1: Moderation Analyses of Surprise Ratings on Probability Estimates*

| | Model 1 | | | | | Model 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *B* | *S.E.* | *p* | 95% LL | 95% UL | *B* | *S.E.* | *p* | 95% LL | 95% UL |
| constant | 27.00 | 1.44 | 0.00 | 24.18 | 29.82 | 23.64 | 2.11 | 0.00 | 19.50 | 27.77 |
| Outcome knowledge | 13.47 | 1.68 | 0.00 | 10.17 | 16.77 | 13.95 | 2.65 | 0.00 | 8.76 | 19.15 |
| Surprise | -6.12 | 0.55 | 0.00 | -7.20 | -5.03 | -6.21 | 0.55 | 0.00 | -7.30 | -5.13 |
| Outcome knowledge x Surprise | -1.79 | 0.80 | 0.03 | -3.36 | -0.22 | -1.61 | 0.80 | 0.04 | -3.18 | -0.04 |
| Surprise squared | | | | | | 0.65 | 0.32 | 0.04 | 0.03 | 1.28 |
| Outcome knowledge x Surprise squared | | | | | | -0.06 | 0.47 | 0.89 | -0.98 | 0.85 |
| EventA | -2.81 | 1.68 | 0.10 | -6.10 | 0.49 | -2.02 | 1.70 | 0.23 | -5.36 | 1.32 |

## Study 1: Forest plot for surprise ratings

Figure S7

*Forest Plot for Surprise Ratings*



## Study 1: Age-related analyses

We examined whether hindsight bias is contingent on age by two sets of analyses. In the first set of analyses, we conducted Mann-Whitney *U* tests in smaller samples consisting of younger participants and older participants, respectively. In the second set of analyses, we tested whether age moderated the relationship between experimental condition and outcomes such as probability estimates and surprise ratings.

For the sub-sample analyses, we reviewed previous studies to try to determine the age range of younger and older adults. In Bayen et al. (2006), the age range of younger adults was 17-28 years old, and the age range of older adults was 61-87 years old. In Bernstein et al. (2011), the age range of younger adults was 18-29 years old, and the age range of older adults was 61-95 years old. In Pohl et al. (2018), the age range of younger adults was 19-31 years old, and the age range of older adults was 60-82 years old. We therefore chose the age range of younger adults to be 18-31 years old ($n$= 224). However, because only 56 participants in our sample were equal to or above 60 years old, we had to choose a more lenient cut-off for the age range of older adults. The age range of older adults was 50 years old and above ($n$ = 175) for this set of analyses.

Table S9 showed the means and standard deviations of probability estimates and surprise ratings of younger adults. Table S10 showed the results of Mann-Whitney *U* tests of younger adults. For Event B Outcome 1, probability estimates of younger adults in the hindsight condition were significantly higher than those in the foresight condition. This finding is consistent with the prediction of hindsight bias. All other Mann-Whitney *U* tests on probability estimates of younger adults were nonsignificant after adjusting for multiple comparison using Benjamini and Hochberg's (1995) false discovery rate control method. The effect sizes Cohen's $d$ of probability estimates of younger adults ranged from -0.03 to 1.15, with a mean of $d$ = 0. 64. We therefore found weak support for Hypothesis 1 among younger adults. None of the Mann-

Whitney *U* tests on surprise ratings of younger adults were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of surprise ratings of younger adults ranged from -1.20 to 0.51, with a mean of $d = -0.40$. We therefore found no support for Hypothesis 2 among younger adults.

Table S11 reports the means and standard deviations of probability estimates and surprise ratings of older adults. Table S12 reports the results of Mann-Whitney *U* Tests of older adults. None of the Mann-Whitney *U* tests on probability estimates of older adults were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of probability estimates of older adults ranged from 0.48 to 1.38, with a mean of $d = 0.76$. However, these findings should be interpreted with caution, as the sample sizes for the analyses involving older adults only were very small (5 to 12 participants per cell). We therefore concluded little support for Hypothesis 1 among older adults, which could be due to lack of statistical power due to small sample sizes. None of the Mann-Whitney *U* tests on surprise ratings of older adults were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of surprise ratings of older adults ranged from -0.58 to 0.17, with a mean of $d = -0.20$. We therefore found no support for Hypothesis 2 among older adults. These findings also need to be interpreted with caution due to the small sample sizes.

Table S9

*Study 1: Means and Standard Deviations of Probability Estimates and Surprise Ratings (Younger Participants Aged Between 18 and 31 Years Old)*

| Experimental Condition | Variable | Outcome Evaluated | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Outcome 1 | | | Outcome 2 | | | Outcome 3 | | | Outcome 4 | | |
| | | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* |
| | Event A: British-Gurka struggle | | | | | | | | | | | | |
| Probability | Before | 7 | 23.57 | 17.73 | 7 | 33.57 | 26.10 | 7 | 27.86 | 21.19 | 7 | 15.00 | 14.43 |
| Probability | After | 11 | 39.18 | 31.57 | 12 | 40.83 | 19.17 | 13 | 27.31 | 12.18 | 7 | 40.00 | 28.43 |
| Surprise | Before | 7 | 5.29 | 1.70 | 7 | 4.29 | 2.14 | 7 | 3.14 | 1.57 | 7 | 4.43 | 1.90 |
| Surprise | After | 11 | 3.18 | 1.78 | 12 | 3.92 | 2.02 | 13 | 4.00 | 1.73 | 7 | 4.14 | 1.22 |
| | Event B: Near riot in Atlanta | | | | | | | | | | | | |
| Probability | Before | 13 | 10.46 | 11.19 | 13 | 27.46 | 24.54 | 13 | 13.08 | 15.75 | 13 | 49.00 | 24.20 |
| Probability | After | 15 | 35.07 | 27.34 | 14 | 41.79 | 26.36 | 9 | 28.33 | 30.41 | 9 | 66.67 | 22.36 |
| Surprise | Before | 13 | 5.69 | 1.18 | 13 | 3.54 | 1.66 | 13 | 5.08 | 1.44 | 13 | 2.62 | 1.50 |
| Surprise | After | 15 | 4.47 | 1.64 | 14 | 3.07 | 1.44 | 9 | 5.11 | 2.09 | 9 | 1.33 | 0.71 |

*Note:* The foresight ratings of all four outcomes came from the same participants in the foresight condition. The hindsight ratings of the four outcomes came from participants in the four hindsight conditions, respectively. Following discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

Table S10

*Study 1: Mann-Whitney U Tests of Probability Estimates Difference and Surprise Difference between Before and After Conditions (Younger Participants Aged Between 18 and 31 Years Old)*

| **After - Before** | ***Probability Estimates*** | | | | | | ***Surprise*** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Mean Difference (Rank)* | ***U*** | ***z*** | $p_{exact}$ | $p_{adjusted}$ | ***d*** | *Mean Difference (Rank)* | ***U*** | ***z*** | $p_{exact}$ | $p_{adjusted}$ | ***d*** |
| Event A Outcome 1 | 2.69 | 27 | 1.05 | .311 | .415 | 0.57 | -5.61 | 14.5 | -2.21 | .029 | .125 | -1.20 |
| Event A Outcome 2 | 1.70 | 34.5 | 0.64 | .546 | .624 | 0.33 | -1.24 | 36.5 | -0.47 | .665 | .887 | -0.18 |
| Event A Outcome 3 | 0.22 | 44.5 | 0.08 | .951 | .951 | -0.03 | 3.08 | 31.5 | 1.13 | .271 | .542 | 0.51 |
| Event A Outcome 4 | 4.57 | 8.5 | 2.06 | .038 | .152 | 1.11 | 0.29 | 23.5 | 0.13 | .897 | .897 | -0.18 |
| Event B Outcome 1 | 8.54 | 38.0 | 2.77 | .004 | .032 | 1.15 | -6.10 | 55.0 | -2.00 | .047 | .125 | -0.85 |
| Event B Outcome 2 | 5.19 | 56.0 | 1.70 | .091 | .182 | 0.56 | -2.60 | 73.5 | -0.87 | .397 | .635 | -0.30 |
| Event B Outcome 3 | 3.67 | 39.0 | 1.33 | .194 | .310 | 0.67 | 0.75 | 54.5 | 0.28 | .790 | .897 | 0.02 |
| Event B Outcome 4 | 4.89 | 32.5 | 1.75 | .083 | .182 | 0.75 | -5.36 | 30.0 | -2.09 | .043 | .125 | -1.03 |

*Note.* When the sample size is small, SPSS reports *p* values from the exact tests (Dineen & Blakesley, 1973), which does not require the assumption of normal distribution; the exact *p* values are also corrected for ties (IBM, 2020a, 2020b). We also calculated the adjusted *p* values due to multiple comparisons using the Benjamini and Hochberg (1995) false discovery rate control method. Cohen's *d* was calculated based on independent-samples *t* tests. Following discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

Table S11

*Study 1: Means and Standard Deviations of Probability Estimates and Surprise Ratings (Older Participants >= 50 Years Old)*

| Experimental Condition | Variable | Outcome Evaluated | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Outcome 1 | | | Outcome 2 | | | Outcome 3 | | | Outcome 4 | | |
| | | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* |
| | Event A: British-Gurka struggle | | | | | | | | | | | | |
| Probability | Before | 10 | 33.00 | 23.24 | 10 | 26.00 | 25.47 | 10 | 15.50 | 15.36 | 10 | 25.50 | 19.36 |
| Probability | After | 5 | 52.00 | 32.71 | 9 | 45.22 | 23.86 | 6 | 41.67 | 24.22 | 9 | 40.00 | 39.37 |
| Surprise | Before | 10 | 2.60 | 1.26 | 10 | 5.30 | 1.83 | 10 | 4.10 | 0.88 | 10 | 3.80 | 1.69 |
| Surprise | After | 5 | 2.40 | 2.61 | 9 | 5.11 | 1.54 | 6 | 3.33 | 2.50 | 9 | 4.00 | 2.18 |
| | Event B: Near riot in Atlanta | | | | | | | | | | | | |
| Probability | Before | 12 | 5.58 | 6.43 | 12 | 30.00 | 31.41 | 12 | 9.42 | 10.85 | 12 | 55.00 | 27.72 |
| Probability | After | 7 | 10.00 | 10.41 | 7 | 52.29 | 30.49 | 9 | 17.22 | 12.53 | 10 | 76.50 | 26.15 |
| Surprise | Before | 12 | 6.08 | 1.73 | 12 | 2.67 | 1.78 | 12 | 6.17 | 1.03 | 12 | 1.75 | 1.48 |
| Surprise | After | 7 | 5.71 | 2.14 | 7 | 3.00 | 2.38 | 9 | 5.56 | 1.81 | 10 | 1.10 | 0.32 |

*Note:* The foresight ratings of all four outcomes came from the same participants in the foresight condition. The hindsight ratings of the four outcomes came from participants in the four hindsight conditions, respectively. Following discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

Table S12

*Study 1: Mann-Whitney U Tests of Probability Estimates Difference and Surprise Difference between Before and After Conditions (Older Participants >= 50 Years Old)*

| **After - Before** | ***Probability Estimates*** | | | | | | ***Surprise*** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Mean Difference (Rank)* | ***U*** | ***z*** | $p_{exact}$ | $p_{adjusted}$ | ***d*** | *Mean Difference (Rank)* | ***U*** | ***z*** | $p_{exact}$ | $p_{adjusted}$ | ***d*** |
| Event A Outcome 1 | 2.40 | 17 | 0.99 | .081 | .162 | 0.72 | -2.55 | 16.5 | -1.07 | .300 | .837 | -0.11 |
| Event A Outcome 2 | 4.54 | 23.5 | 1.76 | .022 | .152 | 0.78 | -0.95 | 40.5 | -0.38 | .741 | .847 | -0.11 |
| Event A Outcome 3 | 5.60 | 9 | 2.30 | .407 | .465 | 1.38 | -2.67 | 20 | -1.10 | .314 | .837 | -0.46 |
| Event A Outcome 4 | 2.22 | 34.5 | 0.87 | .038 | .152 | 0.48 | 1.16 | 39.5 | 0.46 | .672 | .847 | 0.10 |
| Event B Outcome 1 | 1.92 | 33.5 | 0.75 | .488 | .488 | 0.55 | -1.24 | 36.5 | -0.51 | .654 | .847 | -0.20 |
| Event B Outcome 2 | 4.41 | 22.5 | 1.65 | .104 | .166 | 0.72 | 0.34 | 40.5 | 0.13 | .933 | .933 | 0.17 |
| Event B Outcome 3 | 3.89 | 34 | 1.46 | .152 | .203 | 0.67 | -1.75 | 45 | -0.68 | .537 | .847 | -0.43 |
| Event B Outcome 4 | 5.13 | 32 | 1.87 | .063 | .162 | 0.80 | -2.75 | 45 | -1.35 | .248 | .837 | -0.58 |

*Note*. When the sample size is small, SPSS reports *p* values from the exact tests (Dineen & Blakesley, 1973), which does not require the assumption of normal distribution; the exact *p* values are also corrected for ties (IBM, 2020a, 2020b). We also calculated the adjusted *p* values due to multiple comparisons using the Benjamini and Hochberg (1995) false discovery rate control method. Cohen's *d* was calculated based on independent-samples *t* tests. Following discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

Table S13 showed a series of moderation analyses using experimental condition as the independent variable, probability estimates and surprise ratings as dependent variables, and age as the moderator. One participant whose self-reported age was 5 years old was removed from the analyses, and the remaining sample size was 441. None of the moderation analyses were significant (even before adjusting the *p* values for multiple testing, $n$ = 85~92 for each moderation analysis). The findings therefore suggest that the hindsight bias (based on probability estimates) and the null findings about surprise ratings in Study 1 were not contingent on participants' age.

Table S13

*Study 1: Age as Moderator (n = 441)*

| | | Outcome 1 | | | Outcome 2 | | | Outcome 3 | | | Outcome 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *B* | *SD* | *p* | *B* | *SD* | *p* | *B* | *SD* | *p* | *B* | *SD* | *p* |
| Event A - Probability | Constant | 20.22 | 3.75 | .000 | 39.57 | 3.94 | .000 | 24.65 | 3.05 | .000 | 15.56 | 3.14 | .000 |
| | Condition | 25.51 | 5.20 | .000 | 4.03 | 5.54 | .469 | 7.78 | 4.33 | .076 | 11.37 | 4.42 | .012 |
| | Age | 0.48 | 0.32 | .140 | -0.40 | 0.34 | .247 | -0.48 | 0.26 | .075 | 0.39 | 0.27 | .155 |
| | Condition * Age | -0.33 | 0.45 | .472 | 0.49 | 0.46 | .289 | 0.78 | 0.41 | .064 | -0.12 | 0.41 | .769 |
| Event A - Surprise | Constant | 4.53 | 0.31 | .000 | 3.86 | 0.31 | .000 | 3.30 | 0.27 | .000 | 4.62 | 0.26 | .000 |
| | Condition | -1.35 | 0.43 | .003 | 0.23 | 0.44 | .599 | 0.07 | 0.38 | .849 | -0.02 | 0.37 | .961 |
| | Age | -0.07 | 0.03 | .007 | 0.04 | 0.03 | .154 | 0.05 | 0.02 | .047 | -0.04 | 0.02 | .131 |
| | Condition * Age | 0.06 | 0.04 | .103 | -0.01 | 0.04 | .843 | -0.06 | 0.04 | .112 | 0.04 | 0.03 | .297 |
| Event B - Probability | Constant | 7.64 | 2.49 | .003 | 25.68 | 3.96 | .000 | 13.23 | 3.53 | .000 | 53.45 | 4.05 | .000 |
| | Condition | 16.71 | 3.53 | .000 | 23.44 | 5.76 | .000 | 16.47 | 5.03 | .002 | 11.94 | 5.78 | .042 |
| | Age | -0.18 | 0.22 | .414 | 0.23 | 0.35 | .518 | -0.30 | 0.31 | .333 | 0.26 | 0.36 | .475 |
| | Condition * Age | -0.52 | 0.31 | .098 | -0.54 | 0.52 | .308 | -0.07 | 0.43 | .865 | 0.02 | 0.48 | .967 |
| Event B - Surprise | Constant | 5.87 | 0.24 | .000 | 2.82 | 0.24 | .000 | 5.42 | 0.26 | .000 | 1.98 | 0.21 | .000 |
| | Condition | -0.62 | 0.33 | .065 | 0.14 | 0.34 | .680 | -0.05 | 0.37 | .884 | -0.04 | 0.30 | .897 |
| | Age | 0.02 | 0.02 | .380 | -0.03 | 0.02 | .125 | 0.03 | 0.02 | .156 | -0.02 | 0.02 | .182 |
| | Condition * Age | 0.03 | 0.03 | .324 | 0.05 | 0.03 | .120 | -0.01 | 0.03 | .744 | 0.01 | 0.02 | .719 |

*Note.* $n$ = 85~92 for each moderation analysis. Following discussion with lead original author and editor Events C and D about therapy have been removed from reporting due to problematic stimuli in the target article.

# Study 2

## Study 2: Transparency report

### PREREGISTRATION SECTION

(4) Prior to analyzing the complete data set, a time-stamped preregistration was posted in an independent, third-party registry for the data analysis plan. **Yes**

(5) The manuscript includes a URL to all preregistrations that concern the present study. **Yes**

(6) The study was preregistered… **before any data were collected**

**The preregistration fully describes…**

(14) all inclusion and exclusion criteria for participation (e.g., English speakers who achieved a certain cutoff score in a language test). **Yes**

(15) all procedures for assigning participants to conditions. **Yes**

(16) all procedures for randomizing stimulus materials. **Yes**

(17) any procedures for ensuring that participants, experimenters, and data-analysts were kept naive (blinded) to potentially biasing information. **Yes**

(18) a rationale for the sample size used (e.g., an a priori power analysis). **Yes**

(19) the measures of interest (e.g., friendliness). **Yes**

(20) all operationalizations for the measures of interest (e.g., a questionnaire measuring friendliness). **Yes**

(21) the data preprocessing plans (e.g., transformed, cleaned, normalized, smoothed). **Yes**

(22) how missing data (e.g., dropouts) were planned to be handled. **Yes**

(23) the intended statistical analysis for each research question (this may require, for example, information about the sidedness of the tests, inference criteria, corrections for multiple testing, model selection criteria, prior distributions etc.). **Yes**

**Comments about your Preregistration**

No comments.

## METHODS SECTION

**The manuscript fully describes…**

(26) the rationale for the sample size used (e.g., an a priori power analysis). **Yes**

(27) how participants were recruited. **Yes**

(28) how participants were selected (e.g., eligibility criteria). **Yes**

(29) what compensation was offered for participation. **No**

(30) how participant dropout was handled (e.g., replaced, omitted, etc.). **Yes**

(31) how participants were assigned to conditions. **Yes**

(32) how stimulus materials were randomized. **Yes**

(33) whether (and, if so, how) participants, experimenters, and data-analysts were kept naive to potentially biasing information. **NA**

(34) the study design, procedures, and materials to allow independent replication. **Yes**

(35) the measures of interest (e.g., friendliness). **Yes**

(36) all operationalizations for the measures of interest (e.g., a questionnaire measuring friendliness). **Yes**

(37) any changes to the preregistration (such as changes in eligibility criteria, group membership cutoffs, or experimental procedures)? **Yes**

**Comments about your Methods section**

No comments.

---

## RESULTS AND DISCUSSION SECTION

**The manuscript…**

(32) distinguishes explicitly between “confirmatory” (i.e., prespecified) and “exploratory” (i.e., not prespecified) analyses. **Yes**

(33) describes how violations of statistical assumptions were handled. **Yes**

(34) justifies all statistical choices (e.g., including or excluding covariates; applying or not applying transformations; use of multi-level models vs. ANOVA). **Yes**

(35) reports the sample size for each cell of the design. **Yes**

(36) reports how incomplete or missing data were handled. **Yes**

(37) presents protocols for data preprocessing (e.g., cleaning, discarding of cases and items, normalizing, smoothing, artifact correction). **Yes**

**Comments about your Results and Discussion**

No comments.

---

## DATA, CODE, AND MATERIALS AVAILABILITY SECTION

**The following have been made publicly available…**

(37) the (processed) data, on which the analyses of the manuscript were based. **Yes**

(38) all code and software (that is not copyright protected). **Yes**

(39) all instructions, stimuli, and test materials (that are not copyright protected). **Yes**

(40) Are the data properly archived (i.e., would a graduate student with relevant background knowledge be able to identify each variable and reproduce the analysis)? **Yes**

(41) The manuscript includes a statement concerning the availability and location of all research items, including data, materials, and code relevant to the study. **Yes**

**Comments about your Data, Code, and Materials**

No comments.

## Study 2: Power analysis

### Original sample size and p-values

In the original study, 184 participants were recruited to finish the questionnaire. These participants were assigned to three conditions: Foresight condition, Hindsight A condition, Hindsight B condition. However, the exact actual number of participants per condition with/without exclusion was not revealed.

Results of the original experiment 1 showed that the mean estimated probability in Foresight condition was significantly different from that in the Hindsight condition at the significance levels of .001, .01 and .05.

### Estimation of effect size in the original study

The effect size of the original experiment was estimated using the following procedure in Mavis (Hamilton et al., 2016).

- Estimated number of participants per condition = 184/3 = 61 (rounded)
- Largest *p*-value in the table of mean probabilities of the original study = 0.05 (for a conservative estimation)
- **Cohen's *d* = 0.36, 95% CI [0, 0.72]**

### Calculation of the minimum sample size required

We estimated the minimum sample size required using the effect size of $d = .36$, power of .95, and significance level of .05 (one-tailed) in G*Power 3.1 (Faul et al. 2009). The minimum sample size required was 336.

Table S14

*Study 2: Sample Size Calculation*

| **t tests**- Means: Difference between two independent means (two groups) | | |
|---|---|---|
| **Analysis**: | A priori: Compute required sample size | |
| **Input**: | Tail(s) | = One |
| | Effect size d | = 0.36 |
| | α err prob | = 0.05 |
| | Power (1-βerr prob) | = 0.95 |
| | Allocation ratio N2/N1 | = 1 |
| **Output**: | Noncentrality parameter δ | = 3.2994545 |
| | Critical t | = 1.6494286 |
| | Df | = 334 |
| | Sample size group 1 | = 168 |
| | Sample size group 2 | = 168 |
| | Total sample size | = 336 |
| | Actual power | = 0.9503142 |

## Study 2: Materials and scales used in the replication + extension experiment

The stimuli and the questions of the replication are taken from the original study (Slovic & Fischhoff, 1977), with a few adaptations made to present them clearer. See the main text for details of the adaptations made.

**Scenario 1: Virgin rat**

Several researchers intend to perform the following experiment: They will inject blood from a mother rat into a virgin rat immediately after the mother rat has given birth. After the injection, the virgin rat will be placed in a cage with the newly born baby rats, after removal of the actual mother.

The possible outcomes were:

(a) the virgin rat exhibited maternal behaviour, or

(b) the virgin rat failed to exhibit maternal behavior.

*Comprehension check:*

- What will be injected blood from mother rat?
    - Mother rat
    - Virgin rat with mother rat blood injection

***Foresight Condition (Outcomes A & B)***

1. Try and estimate, what are the probabilities of the following outcomes (these probabilities should total 100%)

Virgin rat **will exhibit** maternal behavior : _____ (1)

Virgin rat **will NOT exhibit** maternal behavior : _____ (2)

Total : _____

2. If the virgin rat **does exhibit** maternal behavior, what is the probability that in a replication of this experiment with 10 additional virgin female rats (these probabilities should total 100%)

a. All will exhibit maternal behavior? : _____ (1)

b. Some will exhibit maternal behavior? : _____ (2)

c. None will exhibit maternal behavior? : _____ (3)

Total : _____

3. If the virgin rat **does exhibit** maternal behavior, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised

4. If the virgin rat **does NOT exhibit** maternal behavior, what is the probability that in a replication of this experiment with 10 additional virgin female rats (these probabilities should total 100%)

a. All will exhibit maternal behavior? : _____

b. Some will exhibit maternal behavior? : _____

c. None will exhibit maternal behavior? : _____

Total : _____

5. If the virgin rat **does NOT exhibit** maternal behavior, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised

6. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Virgin Rat experiment**? 0 = Extremely not confident … 6 = Extremely confident

***Hindsight Condition (Outcome A)***

| Outcome: The initial virgin rat **exhibited** maternal behavior in the first trial. |
|---|

1. What is the probability that in a replication of this experiment with 10 additional virgin female rats (these probabilities should total 100%)

a. All will exhibit maternal behavior? : _____

b. Some will exhibit maternal behavior? : _____

c. None will exhibit maternal behavior? : _____

Total : _____

2. Do you think the finding that the virgin rat **exhibited** maternal behavior is surprising? 1 = Not surprising at all … 5 = Extremely surprising

3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Virgin Rat experiment**? 0 = Extremely not confident … 6 = Extremely confident

***Hindsight Condition (Outcome B)***

| Outcome: The initial virgin rat **did NOT exhibit** maternal behavior in the first trial. |
|---|

1. What is the probability that in a replication of this experiment with 10 additional virgin female rats (these probabilities should total 100%)

a. All will exhibit maternal behavior? : _____

b. Some will exhibit maternal behavior? : _____

c. None will exhibit maternal behavior? : _____

Total : _____

2. Do you think the finding that the virgin rat **did not exhibit** maternal behavior is surprising? 1 = Not surprising at all … 5 = Extremely surprising

3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Virgin Rat experiment**? 0 = Extremely not confident … 6 = Extremely confident

**Scenario 2: Hurricane seeding**

A team of government meteorologists recently seeded a tropical storm, which had reached hurricane status, with large quantities of silver-iodide crystals (the same type of crystals that are used to seed clouds in attempts to produce rain).

The possible outcomes were:

(a) the hurricane increased in intensity, or

(b) the hurricane decreased in intensity.

*Comprehension checks:*

- What did the meteorologists recently seed?
    - Tropical storm
    - Lawn
- What was the purpose of silver-iodide crystals?
    - To produce rain
    - To stop the rain

***Foresight Condition (Outcomes A & B)***

1. Try and estimate, what are the probabilities of the following outcomes (these probabilities should total 100%)

Hurricane will **increase** in intensity : _____ (1)

Hurricane will **decrease** in intensity : _____ (2)

Total : _____

2. If the hurricane does **increase** in intensity, what is the probability that in a replication of this experiment with 6 additional hurricanes (these probabilities should total 100%)

a. All will increase in intensity? : _____ (1)

b. Some will increase in intensity? : _____ (2)

c. None will increase in intensity? : _____ (3)

Total : _____

3. If the hurricane does **increase** in intensity, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised

4. If the hurricane does **decrease** in intensity, what is the probability that in a replication of this experiment with 6 additional hurricanes (these probabilities should total 100%)

a. All will weaken in intensity? : _____ (1)

b. Some will weaken in intensity? : _____ (2)

c. None will weaken in intensity? : _____ (3)

Total : _____

5. If the hurricane does **decrease** in intensity, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised

6. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Hurricane Seeding experiment**? 0 = Extremely not confident … 6 = Extremely confident

***Hindsight Condition (Outcome A)***

| |
|---|
| Outcome: The initial hurricane **increased** in intensity in the first trial. |

1. What is the probability that in a replication of this experiment with 6 additional hurricanes (these probabilities should total 100%)

a. All will increase in intensity? : _____ (1)

b. Some will increase in intensity? : _____ (2)

c. None will increase in intensity? : _____ (3)

Total : _____

2. Do you think the finding that the hurricane **increased** in intensity is surprising? 1 = Not surprising at all … 5 = Extremely surprising

3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Hurricane Seeding experiment**? 0 = Extremely not confident … 6 = Extremely confident

***Hindsight Condition (Outcome B)***

| Outcome: The initial hurricane **decreased** in intensity in the first trial. |
|---|

1. What is the probability that in a replication of this experiment with 6 additional hurricanes (these probabilities should total 100%)

a. All will weaken in intensity? : _____ (1)

b. Some will weaken in intensity? : _____ (2)

c. None will weaken in intensity? : _____ (3)

Total : _____

2. Do you think the finding that the hurricane **decreased** in intensity is surprising? 1 = Not surprising at all … 5 = Extremely surprising

3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Hurricane Seeding experiment**? 0 = Extremely not confident … 6 = Extremely confident

**Scenario 3: Gosling imprinting**

A goose egg was placed in a soundproof, heated box from time of laying to time of cracking. Approximately 2 days before it cracked, the experimenter began intermittently to play sounds of ducks quacking into the box. On the day after birth, the gosling was placed on a smooth floor equidistant from a duck and a goose, each of which was in a wire cage. The gosling was observed for 2 minutes.

The possible outcomes were

(a) the gosling approached the caged duck, or

(b) the gosling approached the caged goose.

*Comprehension check:*

- What sounds were played into the box by the experimenter?
    - Goose sounds
    - Duck sounds

***Foresight Condition (Outcomes A & B)***

1. Try and estimate, what are the probabilities of the following outcomes (these probabilities should total 100%)

Gosling will approach the caged **duck** : _____ (1)

Gosling will approach the caged **goose** : _____ (2)

Total : _____

2. If the gosling does approach the caged **goose**, what is the probability that in a replication of this experiment with 10 additional goslings (these probabilities should total 100%)

a. All will approach the caged goose? : _____ (1)

b. Some will approach the caged goose? : _____ (2)

c. None will approach the caged goose? : _____ (3)

Total : _____

3. If the gosling does approach the caged **goose**, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised

4. If the gosling does approach the caged **duck**, what is the probability that in a replication of this experiment with 10 additional goslings (these probabilities should total 100%)

a. All will approach the caged duck? : _____ (1)

b. Some will approach the caged duck? : _____ (2)

c. None will approach the caged duck? : _____ (3)

Total : _____

5. If the hurricane does **decrease** in intensity, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised

6. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Gosling Imprinting experiment**? 0 = Extremely not confident … 6 = Extremely confident

***Hindsight Condition (Outcome A)***

| |
|---|
| Outcome: Outcome: The initial gosling approached the caged **duck**. |

1. What is the probability that in a replication of this experiment with 10 additional goslings (these probabilities should total 100%)

a. All will approach the caged duck? : _____ (1)

b. Some will approach the caged duck? : _____ (2)

c. None will approach the caged duck? : _____ (3)

Total : _____

2. Do you think the finding that the gosling approached the caged **duck** is surprising? 1 = Not surprising at all … 5 = Extremely surprising

3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Gosling Imprinting experiment**? 0 = Extremely not confident … 6 = Extremely confident

***Hindsight Condition (Outcome B)***

| Outcome: The initial gosling approached the caged **goose**. |
|---|

1. What is the probability that in a replication of this experiment with 10 additional goslings (these probabilities should total 100%)

a. All will approach the caged goose? : _____ (1)

b. Some will approach the caged goose? : _____ (2)

c. None will approach the caged goose? : _____ (3)

Total : _____

2. Do you think the finding that the gosling approached the caged **goose** is surprising? 1 = Not surprising at all … 5 = Extremely surprising

3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Gosling Imprinting experiment**? 0 = Extremely not confident … 6 = Extremely confident

**Scenario 4: The Y-Test**

In the pretest that she intends to run in the future, an experimenter placed a 4-year-old child in front of an easel with a large Y on it, with a dot in the lower left-hand third of the letter. The child was then taken around to the back of easel where he saw another Y. He was asked to draw a dot in the "same position" on that Y as the one he had just seen.

The possible outcomes were

(a) the child placed a dot in Area A (the lower left-hand third), or

(b) the child placed a dot in Area B (the upper third). The lower right hand was labeled Area C.

*Comprehension check:*

- Where is the dot on the large Y (in the front)?
    - Lower left-hand third of the letter Y
    - Lower right-hand corner of the letter Y

***Foresight Condition (Outcomes A & B)***

1. Try and estimate, what are the probabilities of the following outcomes (these probabilities should total 100%)

Child would place a dot in Area A (the **lower left-hand** corner of the letter Y) : _____ (1)

Child would place a dot in Area C (the **lower right-hand** corner of the letter Y) : _____ (2)

Total : _____

2. If the child placed a dot in **Area A (the lower left-hand third of the letter Y)**, what is the probability that in a replication of this experiment with one additional child (these probabilities should total 100%)

a. Places dot in Area A (the lower left-hand corner of the letter Y) : _____ (1)

b. Places dot in Area B (the upper corner of the letter Y) : _____ (2)

c. Places dot in Area C (the lower right-hand corner of the letter Y) : _____ (3)

Total : _____

3. If the child placed a dot in **Area A (the lower left-hand corner of the letter Y)**, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised

4. If the child placed a dot in **Area C (the lower right-hand corner of the letter Y)**, what is the probability that in a replication of this experiment with one additional child (these probabilities should total 100%)

a. Places dot in Area A (the lower left-hand corner of the letter Y) : _____ (1)

b. Places dot in Area B (the upper corner of the letter Y) : _____ (2)

c. Places dot in Area C (the lower right-hand corner of the letter Y) : _____ (3)

Total : _____

5. If the child placed a dot in **Area C (the lower right-hand corner of the letter Y)**, how surprised would you be? 1 = Not surprised at all … 5 = Extremely surprised

6. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Y test**? 0 = Extremely not confident … 6 = Extremely confident

***Hindsight Condition (Outcome A)***

Outcome: The initial child placed the dot in **Area A (the lower left-hand third)**.

1. What is the probability that in a replication of this experiment with one additional child (these probabilities should total 100%)

a. Places in Area A (the lower left-hand corner of the letter Y)? : _____ (1)

b. Places in Area B (the upper corner of the letter Y)? : _____ (2)

c. Places in Area C (the lower right-hand corner of the letter Y) : _____ (3)

Total : _____

2. Do you think the finding that the child placed the dot in **Area A (the lower left-hand third)** is surprising? 1 = Not surprising at all … 5 = Extremely surprising

3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Y-test**? 0 = Extremely not confident … 6 = Extremely confident

***Hindsight Condition (Outcome B)***

Outcome: The initial child placed the dot in **Area C (the lower right-hand corner of the letter Y).**

1. What is the probability that in a replication of this experiment with one additional child (these probabilities should total 100%)

a. Places in Area A (the lower left-hand corner of the letter Y)? : _____ (1)

b. Places in Area B (the upper corner of the letter Y)? : _____ (2)

c. Places in Area C (the lower right-hand corner of the letter Y)? : _____ (3)

Total : _____

2. Do you think the finding that the child placed the dot in **Area C (the lower right-hand corner of the letter Y)** is surprising? 1 = Not surprising at all … 5 = Extremely surprising

3. How confident are you about the accuracy of your predictions on the probability of the future outcomes of the **Y-test**? 0 = Extremely not confident … 6 = Extremely confident

## Study 2: Changes made after the pre-registration

3. Study material: Outcome B option for the Y Test was changed from “Area B” to “Area C” in order to prevent a perception in hindsight outcome B participants that it was made for a setup of bias.
4. Study material: Open-ended questions (e.g., “why do you think this happened”) were removed from the final survey. This is because the reasons were not the main interest of the study, plus it would prolong the time for finishing the survey.
5. Exploratory hypotheses: While we proposed to test the exploratory hypotheses using basic statistical methods such independent samples *t*-tests in the pre-registration, we added more sophisticated analyses of mediating and moderating effects in the manuscript.

## Study 2: Sample characteristics

Most of the participants ($n$ = 596, 98.68%) were born in the United States, and the others were born in India, Philippines, Korea, and the United Kingdom. When asked about their family's social class, 31 participants self-identified as lower class (5.13%), 133 as working class (22.02%), 114 as lower middle class (18.87%), 268 as middle class (44.37%), 55 as upper middle class (9.11%), and three as upper class (0.50%).

## Study 2: Additional analyses

### Study 2: Results with the sample after applying exclusion criteria

In the preregistrations, we proposed to analyze the data using a sample after applying a set of pre-specified exclusion criteria:

- All participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale)
- Participants who were not serious about filling in the survey (self-report < 4, on a 1-5 scale).
- Participants who correctly guessed the hypothesis of this study in the funneling section. We screened all participants' answers to the purpose of the study, and removed all cases that mentioned hindsight bias or how outcome knowledge could affect probability estimates or surprise.

The results are summarized in Table S15 and Table S16, and are highly similar to those obtained using the full sample. Regarding task difficulty, the difference between foresight condition ($M$ = 4.98, $SD$ = 1.44) and hindsight outcome A condition ($M$ = 4.41, $SD$ = 1.62) was significant, $t(376) = -3.59$, $p < .001$, $d = -.37$; the difference between foresight condition and hindsight outcome B condition ($M$ = 4.37, $SD$ = 1.53) was also significant, $t(378) = -4.02$, $p < .001$, $d = -.41$. Readers interested in reproducing the analyses can visit our OSF webpage, look for the file 20189-PSYC-INC-Slovic-Fischhoff+1977_Mann-Whitney U test.omv, and use the filter function in JAMOVI to conduct the analyses.

Table S15

*Independent Samples Student's t-tests of Probability Estimates between Foresight and Hindsight Conditions (After Exclusion)*

| **Hindsight vs. Foresight** | ***t*** | ***df*** | ***p*** | **Cohen's *d*** |
|---|---|---|---|---|
| **Virgin rat experiment** | | | | |
| Outcome A: Shows maternal behavior | | | | |
| **a. All show maternal behavior*** ** | 3.28 | 376 | 0.001 | 0.34 |
| b. Some show maternal behavior | 0.79 | 376 | 0.429 | 0.08 |
| c. None show maternal behavior*** | -3.96[a] | 376 | < .001 | -0.41 |
| Outcome B: Fails to show maternal behavior | | | | |
| a. All show maternal behavior | -1.47 | 378 | 0.141 | -0.15 |
| b. Some show maternal behavior | -0.95 | 378 | 0.342 | -0.10 |
| **c. None show maternal behavior** | 1.71 | 378 | 0.088 | 0.18 |
| **Hurricane seeding experiment** | | | | |
| Outcome A: Intensity increases | | | | |
| **a. All increases** | 0.40 | 376 | 0.687 | 0.04 |
| b. Some increases | 0.58 | 376 | 0.560 | 0.06 |
| c. None increases | -1.38 | 376 | 0.169 | -0.14 |
| Outcome B: Intensity weakens | | | | |
| **a. All weaken** | 2.54 | 378 | 0.011 | 0.26 |
| b. Some weaken** | 2.22 | 378 | 0.027 | 0.23 |
| c. None weaken*** | -4.41[a] | 378 | < .001 | -0.45 |
| **Gosling imprinting experiment** | | | | |
| Outcome A: Approaches duck | | | | |
| **a. All approach duck** | 2.04[a] | 376 | 0.042 | 0.21 |
| b. Some approach duck | 0.37 | 376 | 0.713 | 0.04 |
| c. None approach duck** | -3.23[a] | 376 | 0.001 | -0.33 |
| Outcome B: Approaches goose | | | | |
| **a. All approach goose*** | 2.56[a] | 378 | 0.011 | 0.26 |
| b. Some approach goose | -1.01 | 378 | 0.313 | -0.10 |
| c. None approach goose** | -2.32 | 378 | 0.021 | -0.24 |
| **Y-test experiment** | | | | |
| Outcome A: Places dot in Area A | | | | |
| **a. Places in Area A** | 1.20 | 376 | 0.231 | 0.12 |
| b. Places in Area B | 0.87 | 376 | 0.385 | 0.09 |
| c. Places in Area C* | -2.48 | 376 | 0.013 | -0.26 |
| Outcome B: Places dot in Area C | | | | |
| a. Places in Area A | -1.64 | 378 | 0.103 | -0.17 |
| b. Places in Area B | -0.32 | 378 | 0.748 | -0.03 |
| **c. Places in Area C*** | 2.00 | 378 | 0.046 | 0.21 |

*Note.* Bolded options indicate the pairs of comparisons of interest. a. Levene's test was significant. *$p < .05$, **$p < .01$, ***$p < .001$.

Table S16

*Independent Samples Student's t-tests of Surprise and Confidence Ratings between Foresight and Hindsight Conditions (After Exclusion)*

| **Hindsight vs. Foresight** | ***t*** | ***df*** | ***p*** | **Cohen's *d*** |
|---|---|---|---|---|
| **Surprise** | | | | |
| **Outcome A** | | | | |
| a. Virgin rat | -1.43[a] | 376 | 0.153 | -.15 |
| b. Hurricane seeding | 0.93 | 376 | 0.354 | .10 |
| c. Gosling imprinting | -1.15 | 376 | 0.250 | -.12 |
| d. Y-test | -1.64 | 376 | 0.103 | -.17 |
| **Outcome B** | | | | |
| a. Virgin rat* | -1.77 | 378 | 0.078 | -.18 |
| b. Hurricane seeding** | -3.07 | 378 | 0.002 | -.32 |
| c. Gosling imprinting** | -2.21 | 378 | 0.028 | -.23 |
| d. Y-test** | -3.02[a] | 378 | 0.003 | -.31 |
| **Confidence** | | | | |
| **Outcome A** | | | | |
| a. Virgin rat** | -2.61 | 376 | 0.009 | -0.27 |
| b. Hurricane seeding | 0.84 | 376 | 0.403 | 0.09 |
| c. Gosling imprinting | 0.52 | 376 | 0.603 | 0.05 |
| d. Y-test | 1.01 | 376 | 0.315 | 0.10 |
| **Outcome B** | | | | |
| a. Virgin rat* | 2.29 | 378 | 0.022 | 0.24 |
| b. Hurricane seeding | -0.05[a] | 378 | 0.960 | -0.01 |
| c. Gosling imprinting | 1.91 | 378 | 0.057 | 0.20 |
| **d.** Y-test | -0.99 | 378 | 0.323 | -0.10 |

*Note*. a. Levene's test was significant. *$p < .05$, ** $p < .01$, *** $p < .001$.

## Study 2: Codes for calculating confidence intervals of Cohen's *d*

We used the R package psych (Revelle, 2019) and the following codes to calculate the confidence intervals of the effect size Cohen's *d*.

```
library(psych)
# Probability Estimates
# Virgin rat Outcome A
cohen.d.ci(d = 0.323, n1 = 197, n2 = 204, alpha = .05)
cohen.d.ci(d = 0.078, n1 = 197, n2 = 204, alpha = .05)
cohen.d.ci(d = -0.391, n1 = 197, n2 = 204, alpha = .05)
# Virgin rat Outcome B
cohen.d.ci(d = -0.169, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = -0.092, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = 0.182, n1 = 197, n2 = 203, alpha = .05)
# Hurricane seeding Outcome A
cohen.d.ci(d = 0.055, n1 = 197, n2 = 204, alpha = .05)
cohen.d.ci(d = 0.048, n1 = 197, n2 = 204, alpha = .05)
cohen.d.ci(d = -0.142, n1 = 197, n2 = 204, alpha = .05)
# Hurricane seeding Outcome B
cohen.d.ci(d = 0.17, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = 0.274, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = -0.406, n1 = 197, n2 = 203, alpha = .05)
```

```
#Gosling imprinting Outcome A

cohen.d.ci(d = 0.21, n1 = 197, n2 = 204, alpha = .05)

cohen.d.ci(d = 0.042, n1 = 197, n2 = 204, alpha = .05)

cohen.d.ci(d = -0.34, n1 = 197, n2 = 204, alpha = .05)

#Gosling imprinting Outcome B

cohen.d.ci(d = 0.261, n1 = 197, n2 = 203, alpha = .05)

cohen.d.ci(d = -0.093, n1 = 197, n2 = 203, alpha = .05)

cohen.d.ci(d = -0.246, n1 = 197, n2 = 203, alpha = .05)

# Y-test Outcome A

cohen.d.ci(d = 0.1, n1 = 197, n2 = 204, alpha = .05)

cohen.d.ci(d = 0.118, n1 = 197, n2 = 204, alpha = .05)

cohen.d.ci(d = -0.248, n1 = 197, n2 = 204, alpha = .05)

# Y-test Outcome B

cohen.d.ci(d = -0.169, n1 = 197, n2 = 203, alpha = .05)

cohen.d.ci(d = -0.062, n1 = 197, n2 = 203, alpha = .05)

cohen.d.ci(d = 0.224, n1 = 197, n2 = 203, alpha = .05)



# Surprise Outcome A (VR, HS, GI, YT)

cohen.d.ci(d = -0.15, n1 = 197, n2 = 204, alpha = .05)

cohen.d.ci(d = 0.09, n1 = 197, n2 = 204, alpha = .05)

cohen.d.ci(d = -0.1, n1 = 197, n2 = 204, alpha = .05)

cohen.d.ci(d = -0.15, n1 = 197, n2 = , alpha = .05)
```

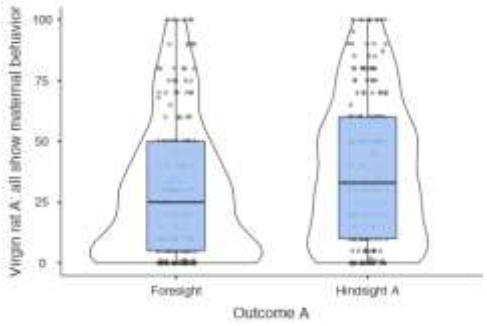
```
# Surprise Outcome B (VR, HS, GI, YT)
cohen.d.ci(d = -0.18, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = -0.28, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = -0.23, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = -0.29, n1 = 197, n2 = 203, alpha = .05)

# Confidence Outcome A (VR, HS, GI, YT)
cohen.d.ci(d = -0.28, n1 = 197, n2 = 204, alpha = .05)
cohen.d.ci(d = 0.07, n1 = 197, n2 = 204, alpha = .05)
cohen.d.ci(d = 0.05, n1 = 197, n2 = 204, alpha = .05)
cohen.d.ci(d = 0.08, n1 = 197, n2 = 204, alpha = .05)
# Confidence Outcome B (VR, HS, GI, YT)
cohen.d.ci(d = 0.2, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = -0.01, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = 0.17, n1 = 197, n2 = 203, alpha = .05)
cohen.d.ci(d = -0.12, n1 = 197, n2 = 204, alpha = .05)

# Task difficulty outcome A
cohen.d.ci(d = -0.38, n1 = 197, n2 = 204, alpha = .05)
# Task difficulty outcome B
cohen.d.ci(d = -0.40, n1 = 197, n2 = 204, alpha = .05)
```

## Study 2: Violin plots of probability estimates

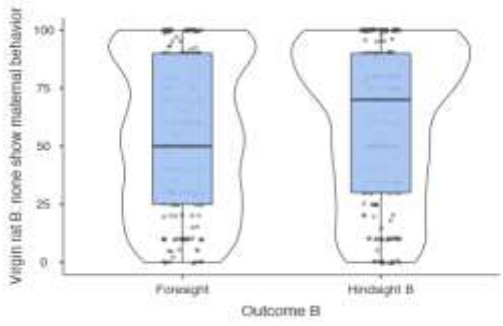
Figure S8a-h

*Violin Plots of Probability Estimates in Foresight and Hindsight Conditions*

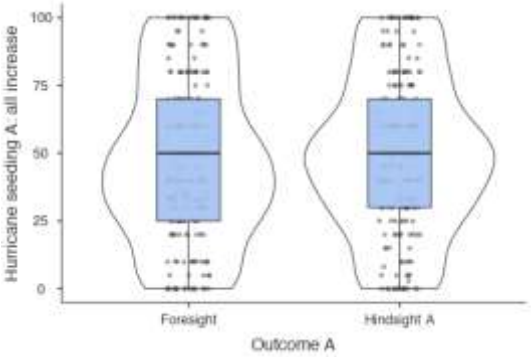
a. Virgin Rat Outcome A



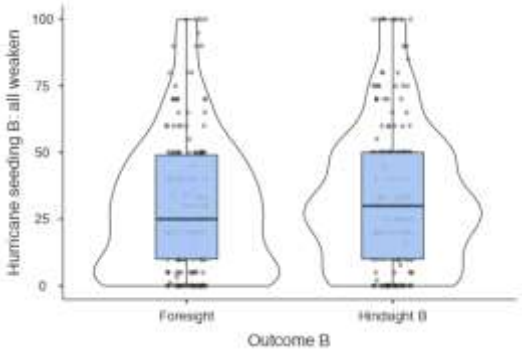

b. Virgin Rat Outcome B



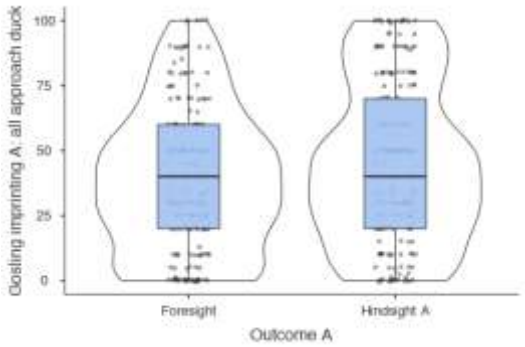

c. Hurricane Seeding Outcome A



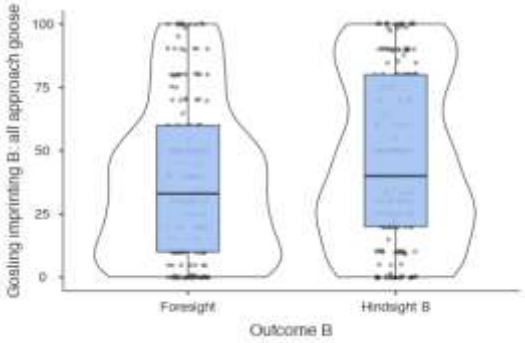

d. Hurricane Seeding Outcome B



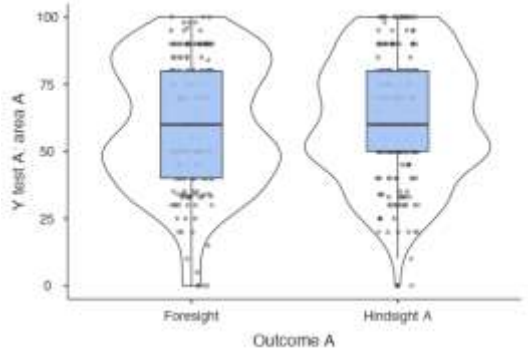

e. Gosling Imprinting Outcome A



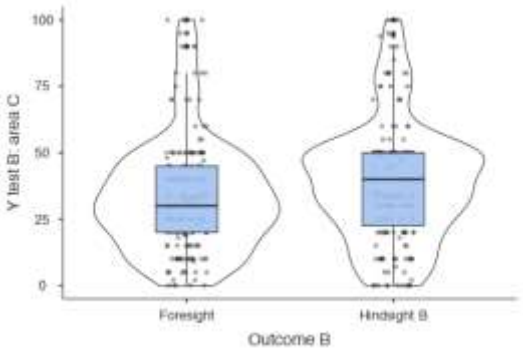

f. Gosling Imprinting Outcome B




g. Y Test Outcome A




h. Y Test Outcome B

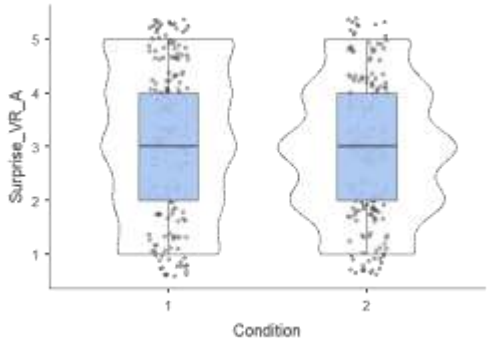


## Study 2: Violin plots of surprise ratings

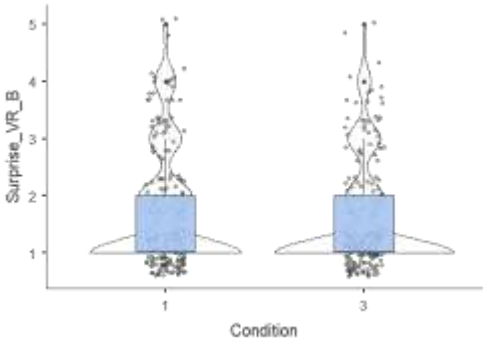
Figure S9a-h

*Violin Plots of Surprise Ratings in Foresight and Hindsight Conditions*

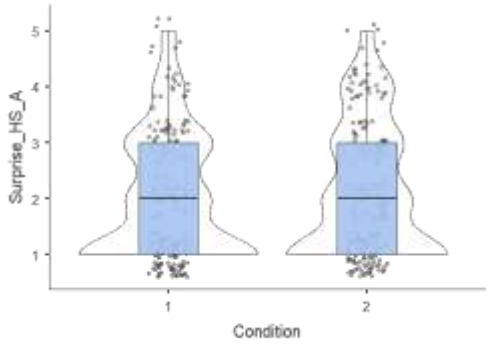
a. Virgin Rat Outcome A



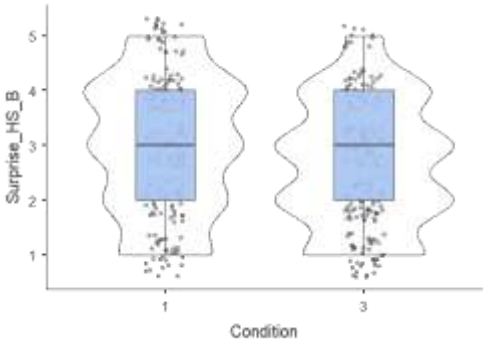

b. Virgin Rat Outcome B



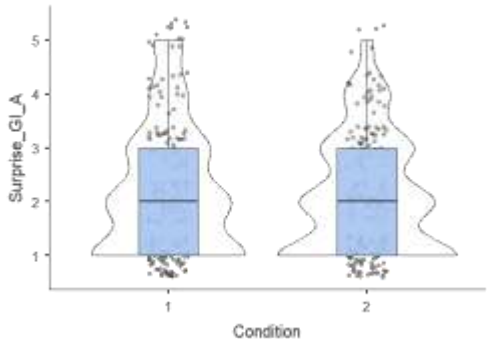

c. Hurricane Seeding Outcome A



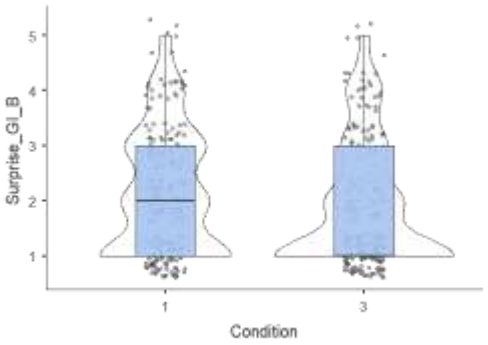

d. Hurricane Seeding Outcome B



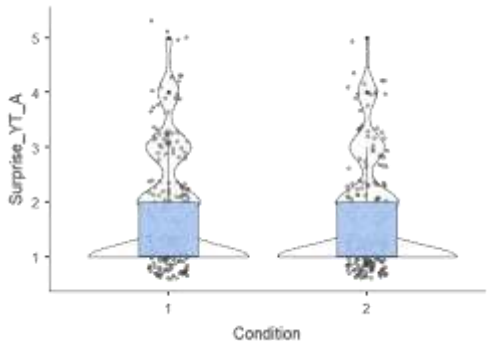

e. Gosling Imprinting Outcome A



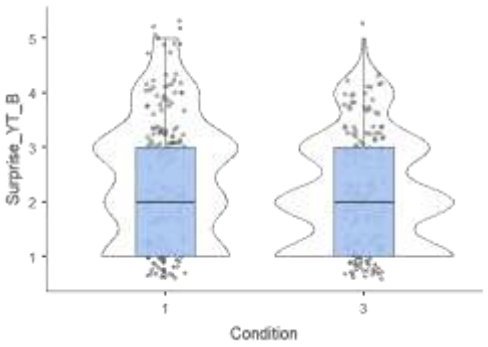

f. Gosling Imprinting Outcome B




g. Y Test Outcome A




h. Y Test Outcome B



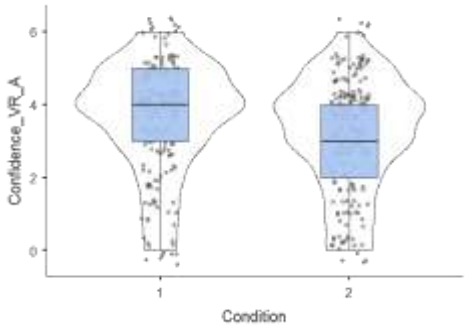

*Note.* 1 = foresight condition, 2 = hindsight condition.

## Study 2: Violin plots of confidence ratings

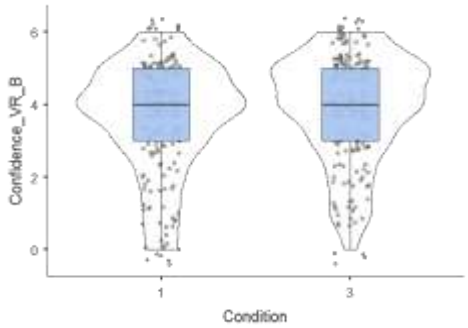
Figure S10a-h

*Violin Plots of Confidence Ratings in Foresight and Hindsight Conditions*

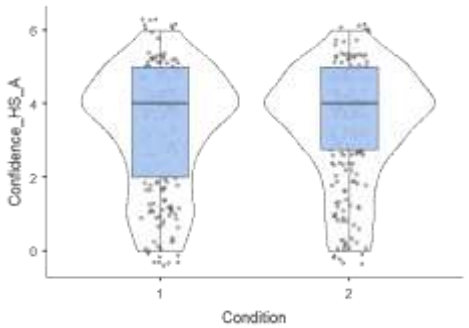
a. Virgin Rat Outcome A



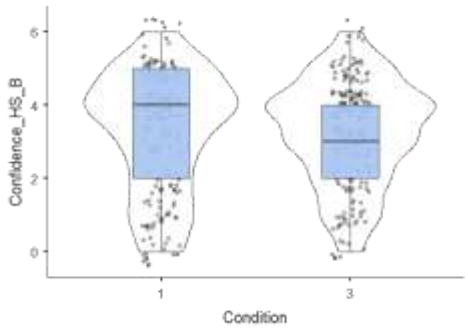

b. Virgin Rat Outcome B



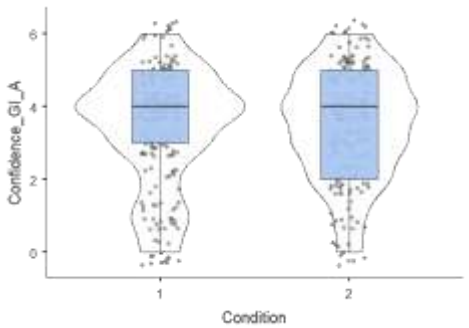

c. Hurricane Seeding Outcome A



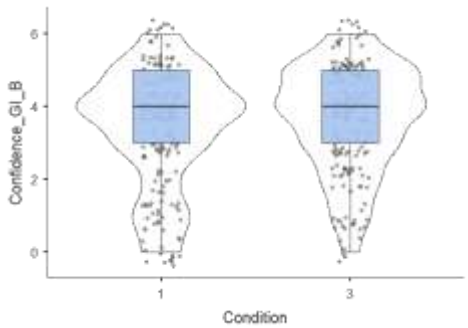

d. Hurricane Seeding Outcome B



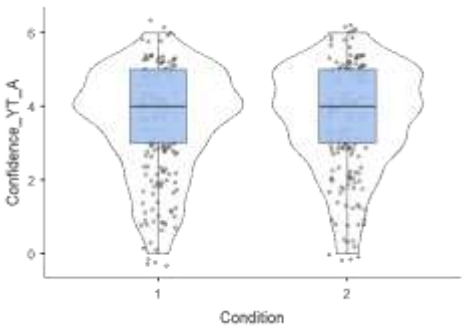

e. Gosling Imprinting Outcome A



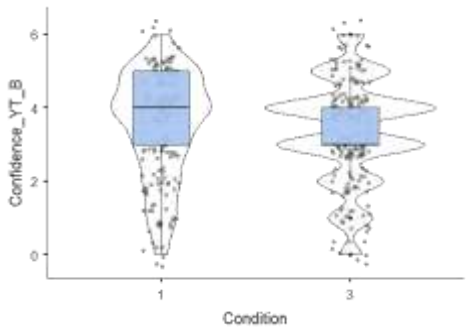

f. Gosling Imprinting Outcome B




g. Y Test Outcome A




h. Y Test Outcome B

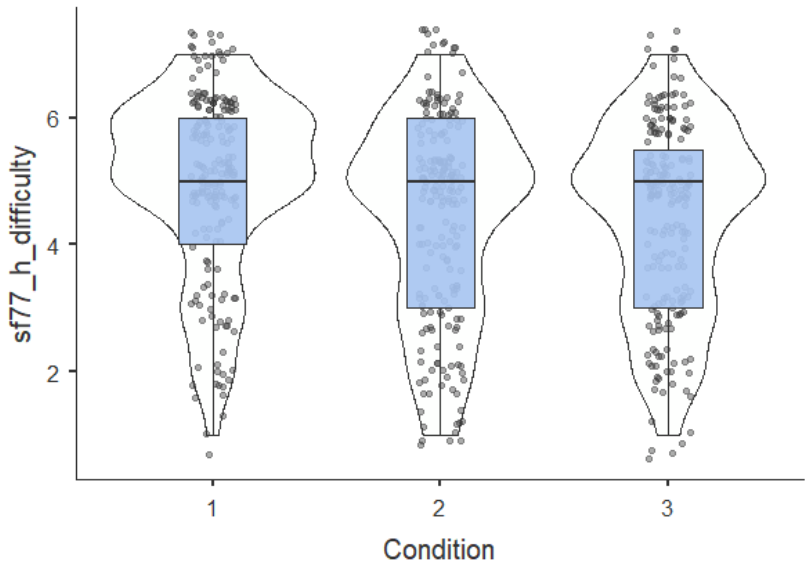



*Note.* 1 = foresight condition, 2 = hindsight condition.

## Study 2: Violin plots of task difficulty

Figure S11

*Violin Plot of Task Difficulty*




1 = foresight condition, 2 = hindsight outcome A condition, 3 = hindsight outcome B condition.

## Study 2: Mann-Whitney U tests of probability estimates

As a robustness check, we tested the hypotheses regarding probability estimates using Mann-Whitney U tests (see Table S17). These results were largely similar to those found with Student's t-tests. Exceptions include: (1) For the virgin rat scenario, the comparison between the foresight condition and the hindsight outcome B condition changed from being marginally significant to being significant; (2) For the gosling imprinting scenario, the comparison between the foresight condition and the hindsight outcome A condition changed from being significant to being marginally significant.

Table S17

*Mann-Whitney U tests of Probability Estimates between Foresight and Hindsight Conditions*

| Hindsight vs. Foresight | *U* | *df* | *p* | Cohen's *d* |
|---|---|---|---|---|
| **Virgin rat experiment** | | | | |
| Outcome A: Shows maternal behavior | | | | |
| **d. All show maternal behavior*** ** | 16152.5 | 399 | < .001 | 0.32 |
| e. Some show maternal behavior | 18919.5 | 399 | 0.310 | 0.08 |
| f. None show maternal behavior*** | 15991 | 399 | < .001 | -0.39 |
| Outcome B: Fails to show maternal behavior | | | | |
| d. All show maternal behavior | 17511 | 398 | 0.026 | -0.17 |
| e. Some show maternal behavior | 18594.5 | 398 | 0.223 | -0.09 |
| **f. None show maternal behavior** | 17868 | 398 | 0.065 | 0.18 |
| **Hurricane seeding experiment** | | | | |
| Outcome A: Intensity increases | | | | |
| **d. All increases** | 19344 | 399 | 0.517 | 0.05 |
| e. Some increases | 19542.5 | 399 | 0.634 | 0.05 |
| f. None increases | 18627 | 399 | 0.200 | -0.14 |
| Outcome B: Intensity weakens | | | | |
| **d. All weaken** | 18054.5 | 398 | 0.092 | 0.17 |
| e. Some weaken** | 16696 | 398 | 0.004 | 0.27 |
| f. None weaken*** | 15761 | 398 | < .001 | -0.41 |
| **Gosling imprinting experiment** | | | | |
| Outcome A: Approaches duck | | | | |
| **d. All approach duck** | 17936.5 | 399 | 0.062 | .21 |
| e. Some approach duck | 19559.5 | 399 | 0.644 | .04 |
| f. None approach duck** | 16520 | 399 | 0.002 | -.34 |
| Outcome B: Approaches goose | | | | |
| **d. All approach goose*** | 17237.5 | 398 | 0.017 | 0.26 |
| e. Some approach goose | 18780.5 | 398 | 0.292 | -0.09 |
| f. None approach goose** | 16691.5 | 398 | 0.004 | -0.25 |
| **Y-test experiment** | | | | |
| Outcome A: Places dot in Area A | | | | |
| **d. Places in Area A** | 19090 | 399 | 0.385 | 0.10 |
| e. Places in Area B | 19516.5 | 399 | 0.614 | 0.12 |
| f. Places in Area C* | 17472.5 | 399 | 0.023 | -0.25 |
| Outcome B: Places dot in Area C | | | | |
| d. Places in Area A | 18133 | 398 | 0.105 | -0.17 |
| e. Places in Area B | 18898.5 | 398 | 0.334 | -0.06 |
| **f. Places in Area C*** | 16828 | 398 | 0.006 | 0.22 |

*Note.* Bolded options indicate the pairs of comparisons of interest. *$p < .05$, **$p < .01$, ***$p < .001$.

## Study 2: Mann-Whitney U tests of surprise ratings and confidence ratings

As a robustness check, we tested Hypotheses 5a and 6a regarding surprise ratings and confidence ratings using Mann-Whitney U tests (see Table S18). These results were largely similar to those found with Student's t-tests. An exceptions is that, for the virgin rat scenario, the comparison of surprise ratings between the foresight condition and the hindsight outcome A condition changed from being marginally significant to being significant.

Table S18

*Mann-Whitney U tests of Surprise and Confidence Ratings between Foresight and Hindsight Conditions*

| **Hindsight vs. Foresight** | ***U*** | ***df*** | ***p*** | **Cohen's *d*** |
|---|---|---|---|---|
| **Surprise** | | | | |
| **Outcome A** | | | | |
| e. Virgin rat | 18398.5 | 399 | 0.135 | -.15 |
| f. Hurricane seeding | 19150 | 399 | 0.390 | .09 |
| g. Gosling imprinting | 19304 | 399 | 0.477 | -.10 |
| h. Y-test | 18633.5 | 399 | 0.159 | -.15 |
| **Outcome B** | | | | |
| e. Virgin rat* | 18019 | 398 | 0.047 | -.18 |
| f. Hurricane seeding** | 16878 | 398 | 0.006 | -.28 |
| g. Gosling imprinting** | 17144.5 | 398 | 0.009 | -.23 |
| h. Y-test** | 17038 | 398 | 0.008 | -.29 |
| **Confidence** | | | | |
| **Outcome A** | | | | |
| e. Virgin rat** | 16706 | 399 | 0.003 | -.28 |
| f. Hurricane seeding | 19399 | 399 | 0.540 | .07 |
| g. Gosling imprinting | 19967 | 399 | 0.911 | .05 |
| h. Y-test | 19256.5 | 399 | 0.459 | .08 |
| **Outcome B** | | | | |
| e. Virgin rat* | 17690 | 398 | 0.041 | .20 |
| f. Hurricane seeding | 19240 | 398 | 0.504 | -.01 |
| g. Gosling imprinting | 18315.5 | 398 | 0.137 | .17 |
| **h.** Y-test | 18104.5 | 398 | 0.093 | -.12 |

*Note*. *$p < .05$, ** $p < .01$, *** $p < .001$.

## Study 2: Mediation analyses

To test the mediation and the moderation hypotheses, we collapsed participants' responses across all four scenarios, forming a data set of 3204 scenario-outcome responses from 604 individuals. Because the responses were nested within participants, we conducted the analyses using the complex model in Mplus (Muthén & Muthén, 2015), with response ID being the cluster variable. The complex model estimates the robust standard errors using the sandwich estimator, which is more accurate than the standard errors estimated in mono-level linear regressions (Heck & Thomas, 2015). We also controlled for the effects of scenarios and outcomes using dummy variables.

To test Hypotheses 4(b), 5(b), and 6(b), we entered the three mediators in the same model. As shown in Table 13, we found a negative effect of hindsight condition on surprise (*B* = -.18, *S.E.* = .06, *p* = .001, 95% CI [-.29, -.07]) and task difficulty (*B* = -.58, *S.E.* = .13, *p* < .001, 95% CI [-.83, -.33]), but not on confidence (*B* = .03, *S.E.* = .11, *p* = .78, 95% CI [-.19, .25]). In addition, we found support for surprise (*B* = -9.10, *S.E.* = .48, *p* < .001, 95% CI [-10.04, -8.17]) and confidence (*B* = 2.49, *S.E.* = .39, *p* < .001, 95% CI [1.72, 3.26]) on probability estimates, but a nonsignificant effect of task difficulty on probability estimates (*B* = .49, *S.E.* = .43, *p* = .25, 95% CI [-.35, 1.32]). When we regressed probability estimates on surprise, confidence, task difficulty, and hindsight condition, the effect of hindsight condition (*B* = 3.92, *S.E.* = 1.30, *p* = .003, 95% CI [1.37, 6.47]) remained significant.

The indirect effect of hindsight condition on probability estimates via surprise was significant (*B* = 1.66, *boot-strapped S.E.* = .52, p = .001, 95% CI [.64, 2.68]). The indirect effects via confidence (*B* = .08, *boot-strapped S.E.* = .28, p = .78, 95% CI [-.46, .62]) and task difficulty (*B* = -.29, *boot-strapped S.E.* = .26, *p* = .26, 95% CI [-.78, .21]) were nonsignificant. The results of person-level linear regression and scenario-level linear regression were similar to those

reported here (see Supplementary Materials). Overall, the results suggest that surprise partially mediated the relationship between hindsight (vs. foresight) and probability estimates, supporting H4(a). There was no support for the mediating effects of confidence in H5(a) or task difficulty in H6(a).

Table S19

*Study 2: Complex Model Mediation Analyses (604 individuals with 3204 scenario-outcome responses)*

| **DV: Surprise** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **95% CI LL** | **95% CI UL** |
|---|---|---|---|---|---|---|
| Constant | 2.17 | .09 | 24.67 | .000 | 2.00 | 2.34 |
| Hindsight (vs. foresight) | -.18 | .06 | -3.30 | .001 | -.29 | -.07 |
| Scenario_VR (Dummy) | .33 | .05 | 6.50 | .000 | .23 | .43 |
| Scenario_HS (Dummy) | .44 | .05 | 9.25 | .000 | .35 | .53 |
| Scenario_GI (Dummy) | .07 | .04 | 1.49 | .136 | -.02 | .15 |
| Outcome A (Dummy) | -.04 | .05 | -.86 | .393 | -.13 | .05 |
| **DV: Confidence** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **95% CI LL** | **95% CI UL** |
| Constant | 3.39 | .13 | 25.66 | .000 | 3.14 | 3.65 |
| Hindsight (vs. foresight) | .03 | .11 | .28 | .779 | -.19 | .25 |
| Scenario_VR (Dummy) | .07 | .06 | 1.11 | .267 | -.06 | .20 |
| Scenario_HS (Dummy) | -.21 | .07 | -3.16 | .002 | -.34 | -.08 |
| Scenario_GI (Dummy) | -.01 | .06 | -.17 | .867 | -.13 | .11 |
| Outcome A (Dummy) | .06 | .06 | 1.04 | .300 | -.06 | .18 |
| **DV: Task difficulty** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **95% CI LL** | **95% CI UL** |
| Constant | 4.99 | .16 | 32.07 | .000 | 4.69 | 5.30 |
| Hindsight (vs. foresight) | -.58 | .13 | -4.55 | .000 | -.83 | -.33 |
| Scenario_VR (Dummy) | .00 | .00 | -16.91 | .000 | .00 | .00 |
| Scenario_HS (Dummy) | .00 | .00 | -15.01 | .000 | .00 | .00 |
| Scenario_GI (Dummy) | .00 | .00 | -16.36 | .000 | .00 | .00 |
| Outcome A (Dummy) | .00 | .08 | -.05 | .959 | -.16 | .15 |
| **DV: Probability estimates** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **95% CI LL** | **95% CI UL** |
| Constant | 61.37 | 3.53 | 17.40 | .000 | 54.46 | 68.29 |
| Hindsight (vs. foresight) | 3.92 | 1.30 | 3.01 | .003 | 1.37 | 6.47 |
| Surprise | -9.11 | .48 | -19.15 | .000 | -10.04 | -8.17 |
| Confidence | 2.49 | .39 | 6.33 | .000 | 1.72 | 3.26 |
| Task difficulty | .49 | .43 | 1.15 | .251 | -.35 | 1.32 |
| Scenario_VR (Dummy) | -.21 | 1.23 | -.17 | .865 | -2.62 | 2.20 |
| Scenario_HS (Dummy) | -3.85 | 1.25 | -3.08 | .002 | -6.31 | -1.40 |
| Scenario_GI (Dummy) | -5.66 | 1.31 | -4.31 | .000 | -8.23 | -3.08 |
| Outcome A (Dummy) | -4.97 | 1.03 | -4.83 | .000 | -6.99 | -2.96 |
| **Indirect effects** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **95% CI LL** | **95% CI UL** |
| Surprise | 1.66 | .52 | 3.20 | .001 | .64 | 2.68 |
| Confidence | .08 | .28 | .28 | .778 | -.46 | .62 |
| Task Difficulty | -.29 | .26 | -1.12 | .264 | -.78 | .21 |
| **Total effects (direct + indirect)** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **95% CI LL** | **95% CI UL** |
| | 5.38 | 1.23 | 4.38 | .000 | 2.97 | 7.78 |

## Study 2: Moderation analyses

To test Hypotheses 4(c), 5(c), and 6(c), we mean-centered surprise, confidence, and task difficulty, and tested one moderator at a time to reduce concerns about potential multicollinearity. As shown in Table 14, the results failed to provide support for the moderating effects of surprise or task difficulty in the relationship between hindsight condition and probability estimates (surprise: $B$ = .39, *S.E.* = .94, $p$ = .68, 95% CI [-1.46, 2.23]; task difficulty: $B$ = -1.08, *S.E.* = .79, $p$ = .17, 95% CI [-2.62, .47]).

We found support for an interaction between hindsight and confidence on probability estimates ($B$ = 4.93, *S.E.* = .74, $p$ < .001, 95% CI [3.48, 6.39]). As shown in Figure S12, the relationship between hindsight condition and probability estimates was positive and significant when confidence was high (simple slope analysis: $B$ = 17.20, *S.E.* = 2.26, $p$ < .001), but it became negative and significant when confidence was low (simple slope analysis: $B$ = -6.61, *S.E.* = 2.07, $p$ < .001). The correlation between hindsight condition and confidence was .01 ($p$ =.57), reducing the concern about multicollinearity in this interaction effect. The results of the moderation analyses at the scenario-level and the person-level are similar to those reported here (see details in the Supplementary Materials). Overall, the results suggest that confidence moderated the relationship between hindsight (vs. foresight) and probability estimates, supporting H5(b). There was no support for the moderating effects of surprise in H4(b) or task difficulty in H6(b).

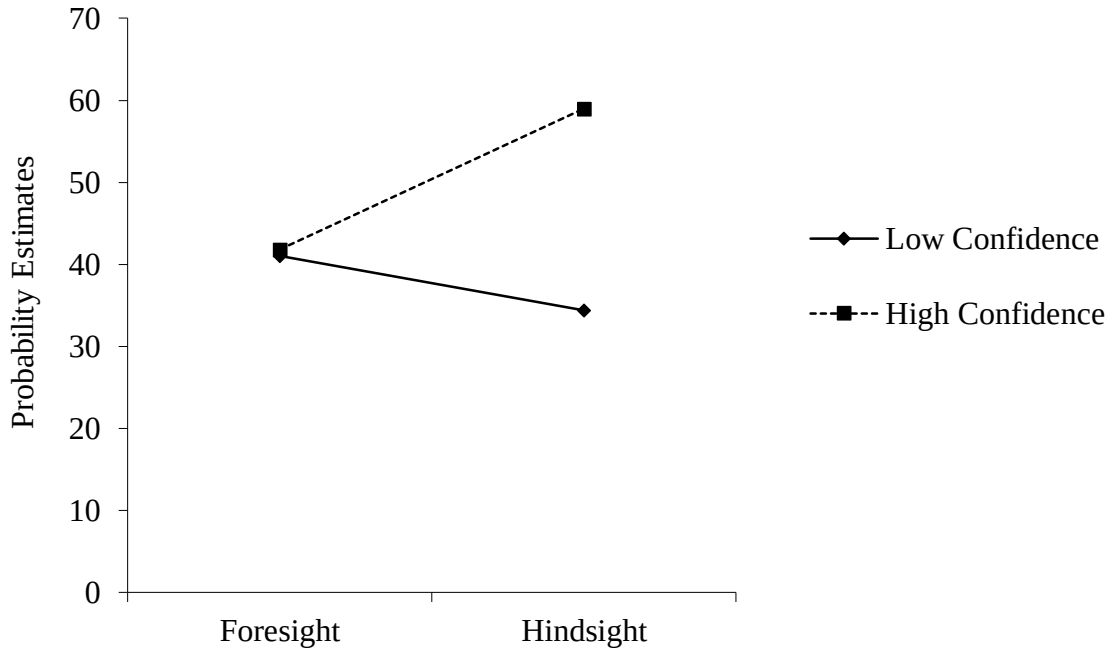
Table S21

*Study 2: Complex Model Moderation Analyses (604 individuals with 3204 scenario-outcome responses)*

| **DV: Probability Estimates** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **95% CI LL** | **95% CI UL** |
|---|---|---|---|---|---|---|
| Constant | 52.01 | 1.82 | 28.53 | .000 | 48.43 | 55.58 |
| Hindsight (vs. foresight) | 3.70 | 1.29 | 2.87 | .004 | 1.17 | 6.23 |
| Surprise | -9.36 | .68 | -13.83 | .000 | -10.68 | -8.03 |
| Hindsight x Surprise | .39 | .94 | .41 | .683 | -1.46 | 2.23 |
| Scenario_VR (Dummy) | -.02 | 1.26 | -.02 | .988 | -2.49 | 2.45 |
| Scenario_HS (Dummy) | -4.36 | 1.25 | -3.49 | .000 | -6.81 | -1.91 |
| Scenario_GI (Dummy) | -5.68 | 1.32 | -4.29 | .000 | -8.28 | -3.09 |
| Outcome A (Dummy) | -4.80 | 1.04 | -4.61 | .000 | -6.85 | -2.76 |
| **DV: Probability Estimates** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **95% CI LL** | **95% CI UL** |
| Constant | 52.97 | 1.91 | 27.76 | .000 | 49.23 | 56.70 |
| Hindsight (vs. foresight) | 5.29 | 1.22 | 4.34 | .000 | 2.90 | 7.68 |
| Confidence | .16 | .49 | .34 | .737 | -.79 | 1.11 |
| Hindsight x Confidence | 4.93 | .74 | 6.65 | .000 | 3.48 | 6.39 |
| Scenario_VR (Dummy) | -3.16 | 1.25 | -2.54 | .011 | -5.60 | -.72 |
| Scenario_HS (Dummy) | -7.90 | 1.22 | -6.49 | .000 | -10.29 | -5.52 |
| Scenario_GI (Dummy) | -6.51 | 1.31 | -4.97 | .000 | -9.07 | -3.94 |
| Outcome A (Dummy) | -4.77 | 1.11 | -4.30 | .000 | -6.94 | -2.60 |
| **DV: Probability Estimates** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **95% CI LL** | **95% CI UL** |
| Constant | 52.53 | 1.94 | 27.15 | .000 | 48.74 | 56.32 |
| Hindsight (vs. foresight) | 5.05 | 1.23 | 4.09 | .000 | 2.63 | 7.46 |
| Task Difficulty | -.04 | .52 | -.07 | .946 | -1.06 | .99 |
| Hindsight x Task Difficulty | -1.08 | .79 | -1.37 | .172 | -2.62 | .47 |
| Scenario_VR (Dummy) | -3.02 | 1.31 | -2.31 | .021 | -5.58 | -.46 |
| Scenario_HS (Dummy) | -8.36 | 1.23 | -6.80 | .000 | -10.77 | -5.95 |
| Scenario_GI (Dummy) | -6.29 | 1.33 | -4.72 | .000 | -8.90 | -3.67 |
| Outcome A (Dummy) | -4.46 | 1.11 | -4.02 | .000 | -6.63 | -2.28 |

Figure S12

*Study 2: Interaction Between Hindsight (vs. Foresight) Condition and Confidence on Probability Estimates.*



## Study 2: Forest plots of surprise ratings and confidence ratings

Figure S13

*Study 2: Forest Plot of the Effect Size of Surprise Ratings*

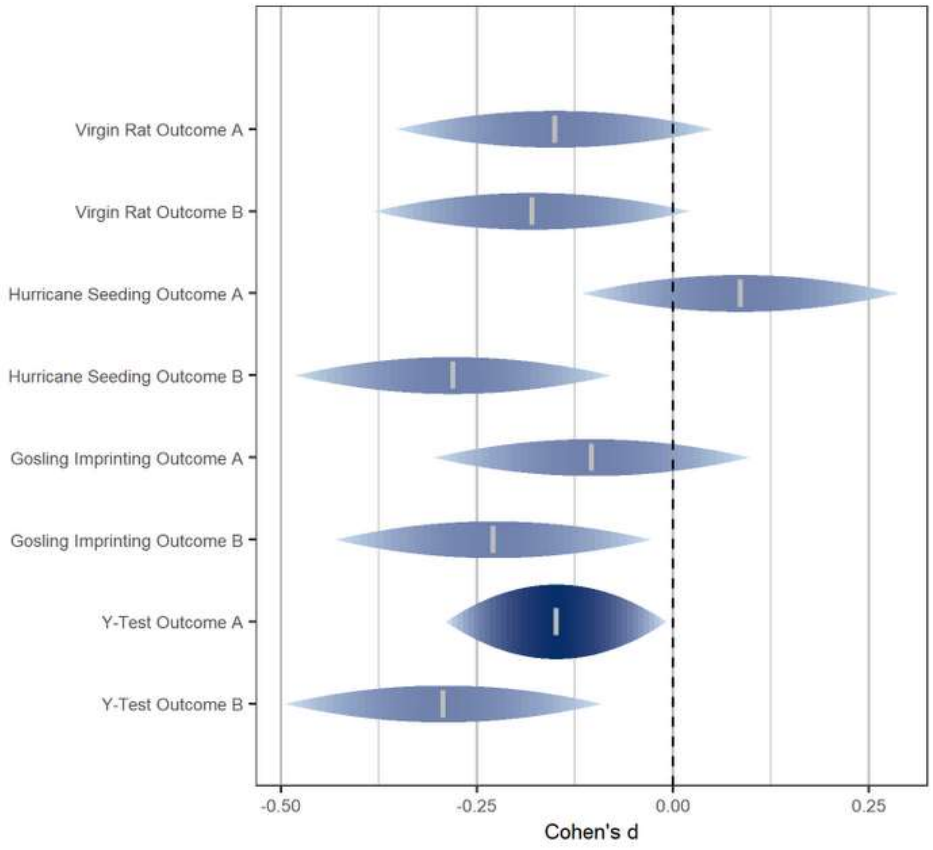
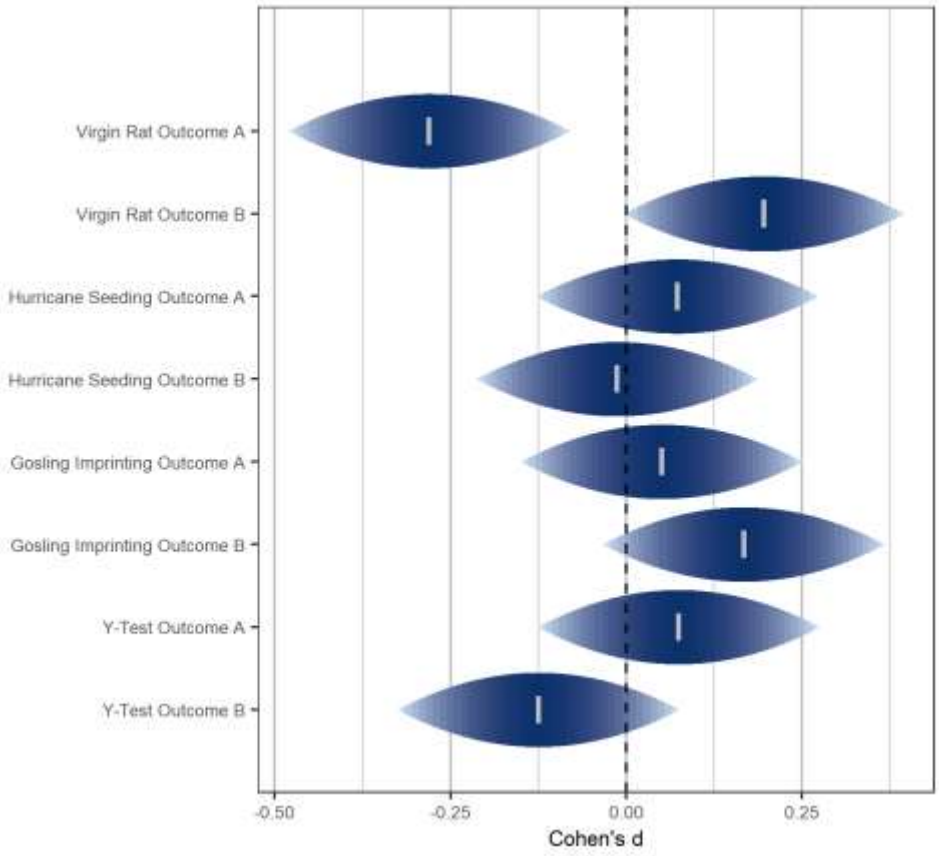


Figure S14

*Study 2: Forest Plot of the Effect Size of Confidence Ratings*



## Study 2: Age-related analyses

We examined whether hindsight bias is contingent on age by two sets of analyses. In the first set of analyses, we conducted independent-samples *t* tests in smaller samples consisting of younger participants and older participants, respectively. In the second set of analyses, we tested whether age moderates the relationship between experimental condition and outcomes such as probability estimates and surprise ratings.

As in Study 1, we chose 18-31 years old as the age range of younger adults ($n$= 195), and 50 years old and above as the age range of older adults ($n$ = 115). Table S23 showed the means and standard deviations of probability estimates of younger adults. As shown in Table S24, none of the independent-samples *t* tests on probability estimates were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of probability estimates for the key comparisons ranged from 0.03 to 0.39, with a mean of $d$ = 0.16. The findings therefore provided no support for Hypothesis 3 among younger participants.

Table S25 showed the means and standard deviations of surprise ratings, confidence ratings, and task difficulty of younger adults. As shown in Table S26, none of the independent-samples *t* tests on surprise ratings were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of surprise ratings ranged from -0.45 to 0.30, with a mean of $d$ = -0.07. The findings therefore provided no support for Hypothesis 4a among younger participants. Also, none of the independent-samples *t* tests on confidence ratings were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of confidence ratings ranged from -0.28 to 0.23, with a mean of $d$ = 0.03. The findings therefore provided no support for Hypothesis 5a among younger participants. In addition, neither of the independent-samples *t* tests on task difficulty were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of task difficulty ranged from -0.29 to -0.24, with a mean of $d$ = -0.27. The findings therefore provided no support for Hypothesis 6a among younger participants.

Table S27 showed the means and standard deviations of probability estimates of older adults. As shown in Table S28, none of the independent-samples *t* tests on probability estimates were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of probability estimates ranged from 0.01 to 0.49, with a mean of $d = 0.24$. The findings therefore provided no support for Hypothesis 3 among older participants.

Table S29 showed the means and standard deviations of surprise ratings, confidence ratings, and task difficulty of older adults. As shown in Table S30, none of the independent-samples *t* tests on surprise ratings were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of surprise ratings ranged from -0.57 to 0.49, with a mean of $d = -0.25$. The findings therefore provided no support for Hypothesis 4a among older participants. Also, none of the independent-samples *t* tests on confidence ratings were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of confidence ratings ranged from -0.13 to 0.43, with a mean of $d = 0.15$. The findings therefore provided no support for Hypothesis 5a among older participants. In addition, neither of the independent-samples *t* tests on task difficulty were nonsignificant after the Benjamini and Hochberg (1995) adjustment. The effect sizes Cohen's *d* of task difficulty ranged from -0.39 to -0.30, with a mean of $d = -0.35$. The findings therefore provided no support for Hypothesis 6a among older participants.

Table S23

*Study 2: Mean Probabilities in Future Trials (in percentage %) (Younger Participants Aged Between 18 and 31 Years Old)*

| | Foresight | | | Hindsight | | |
|---|---|---|---|---|---|---|
| Initial result and kind of replication | *N* | *Mean* | *SD* | *N* | *Mean* | *SD* |
| **Virgin rat experiment** | | | | | | |
| Outcome A: Shows maternal behavior | | | | | | |
| **a. All show maternal behavior** | | **29.59** | 25.29 | | **40.44** | 29.72 |
| b. Some show maternal behavior | 59 | 35.02 | 20.98 | 80 | 32.51 | 21.93 |
| c. None show maternal behavior | | 35.39 | 27.85 | | 27.05 | 26.08 |
| Outcome B: Fails to show maternal behavior | | | | | | |
| a. All show maternal behavior | | 18.71 | 21.83 | | 18.13 | 23.38 |
| b. Some show maternal behavior | 59 | 29.46 | 20.48 | 56 | 26.75 | 21.35 |
| **c. None show maternal behavior** | | **51.83** | 30.93 | | **55.13** | 31.05 |
| **Hurricane seeding experiment** | | | | | | |
| Outcome A: Intensity increases | | | | | | |
| **a. All increase** | | **49.19** | 28.29 | | **50.20** | 27.61 |
| b. Some increase | 59 | 32.03 | 20.66 | 80 | 35.39 | 24.91 |
| c. None increase | | 18.78 | 20.34 | | 14.41 | 15.84 |
| Outcome B: Intensity weakens | | | | | | |
| **a. All weaken** | | **27.63** | 22.88 | | **31.97** | 22.07 |
| b. Some weaken | 59 | 35.63 | 21.19 | 56 | 41.61 | 22.06 |
| c. None weaken | | 36.75 | 28.91 | | 26.42 | 21.59 |
| **Gosling imprinting experiment** | | | | | | |
| Outcome A: Approaches duck | | | | | | |
| **a. All approach duck** | | **36.05** | 25.09 | | **43.68** | 30.77 |
| b. Some approach duck | 59 | 43.08 | 25.06 | 80 | 39.04 | 27.53 |
| c. None approach duck | | 20.86 | 17.86 | | 17.29 | 18.36 |
| Outcome B: Approaches goose | | | | | | |
| **a. All approach goose** | | **37.27** | 27.48 | | **41.11** | 30.13 |
| b. Some approach goose | 59 | 41.37 | 26.43 | 56 | 35.88 | 23.49 |
| c. None approach goose | | 21.36 | 20.83 | | 23.02 | 22.79 |
| **Y-test experiment** | | | | | | |
| Outcome A: Places dot in Area A | | | | | | |
| **a. Places in Area A** | | **58.24** | 25.70 | | **59.05** | 23.76 |
| b. Places in Area B | 59 | 16.22 | 15.63 | 80 | 18.45 | 19.06 |
| c. Places in Area C | | 25.54 | 18.69 | | 22.50 | 17.36 |
| Outcome B: Places dot in Area C | | | | | | |
| a. Places in Area A | | 47.91 | 24.23 | | 46.74 | 21.02 |
| b. Places in Area B | 59 | 17.05 | 15.77 | 56 | 16.51 | 14.80 |
| **c. Places in Area C** | | **35.04** | 21.76 | | **36.76** | 19.17 |

*Note.* Options and numbers marked in bold represent the kind of replication that was reported to have occurred in the initial trial (hindsight) or could possibly occur in the initial trial (foresight). The foresight ratings of both outcome A and outcome B came from the same participants in the foresight condition. The hindsight ratings came from participants in the hindsight outcome A condition or the hindsight outcome B condition, respectively.

Table S24

*Study 2: Independent Samples Student's T-Tests of Probability Estimates between Foresight and Hindsight (Outcome A/B) Conditions (Younger Participants Aged Between 18 and 31 Years Old)*

| **Hindsight vs. Foresight** | ***Mean Difference*** | ***t*** | ***df*** | ***p*** | $p_{adjusted}$ | **Cohen's d** |
|---|---|---|---|---|---|---|
| Virgin rat experiment | | | | | | |
| Outcome A: Shows maternal behavior | | | | | | |
| **a. All show maternal behavior** | 10.84 | 2.26 | 137 | .025 | .396 | 0.39 |
| b. Some show maternal behavior | -2.50 | -0.68 | 137 | .499 | .749 | -0.12 |
| c. None show maternal behavior | -8.34 | -1.81 | 137 | .072 | .576 | -0.31 |
| Outcome B: Fails to show maternal behavior | | | | | | |
| a. All show maternal behavior | -0.59 | -0.14 | 113 | .890 | .890 | -0.03 |
| b. Some show maternal behavior | -2.71 | -0.69 | 113 | .489 | .749 | -0.13 |
| **c. None show maternal behavior** | 3.29 | 0.57 | 113 | .570 | .805 | 0.11 |
| Hurricane seeding experiment | | | | | | |
| Outcome A: Intensity increases | | | | | | |
| **a. All increases** | 1.01 | 0.21 | 137 | .833 | .887 | 0.04 |
| b. Some increases | 3.35 | 0.84 | 137 | .401 | .749 | 0.14 |
| c. None increases | -4.37 | -1.42 | 137 | .157 | .628 | -0.24 |
| Outcome B: Intensity weakens | | | | | | |
| **a. All weaken** | 4.34 | 1.03 | 113 | .303 | .749 | 0.19 |
| b. Some weaken | 5.99 | 1.48 | 113 | .141 | .628 | 0.28 |
| c. None weaken a | -10.33 | -2.16 | 113 | .033 | .396 | -0.40 |
| Gosling imprinting experiment | | | | | | |
| Outcome A: Approaches duck | | | | | | |
| **a. All approach duck** | 7.62 | 1.56 | 137 | .121 | .628 | 0.27 |
| b. Some approach duck | -4.05 | -0.89 | 137 | .375 | .749 | -0.15 |
| c. None approach duck | -3.58 | -1.15 | 137 | .253 | .749 | -0.20 |
| Outcome B: Approaches goose | | | | | | |
| **a. All approach goose** | 3.84 | 0.71 | 113 | .477 | .749 | 0.13 |
| b. Some approach goose | -5.50 | -1.18 | 113 | .242 | .749 | -0.22 |

| | | | | | | |
|---|---|---|---|---|---|---|
| c. None approach goose | 1.66 | 0.41 | 113 | .684 | .864 | 0.08 |
| Y-test experiment | | | | | | |
| Outcome A: Places dot in Area A | | | | | | |
| **a. Places in Area A** | 0.81 | 0.19 | 137 | .849 | .887 | 0.03 |
| b. Places in Area B | 2.23 | 0.74 | 137 | .463 | .749 | 0.13 |
| c. Places in Area C | -3.04 | -0.99 | 137 | .325 | .749 | -0.17 |
| Outcome B: Places dot in Area C | | | | | | |
| a. Places in Area A | -1.17 | -0.28 | 113 | .782 | .887 | -0.05 |
| b. Places in Area B | -0.54 | -0.19 | 113 | .850 | .887 | -0.04 |
| **c. Places in Area C** | 1.72 | 0.45 | 113 | .655 | .864 | 0.08 |

*Note.* Bolded options indicate the pairs of comparisons of interest. The Levene's test of equal variance was nonsignificant for all pairs of comparison. *p* values were adjusted using the Benjamini and Hochberg (1995) false discovery rate control method.

Table S25

*Study 2: Means and Standard Deviations of Surprise Ratings, Confidence Ratings, and Task Difficulty (Younger Participants Aged Between 18 and 31 Years Old)*

| **Scenario** | **Outcome A** | | | | **Outcome B** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Foresight** | | **Hindsight** | | **Foresight** | | **Hindsight** | |
| **Surprise** | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| Virgin rat | 3.19 | 1.31 | 3.05 | 1.25 | 1.76 | 1.01 | 1.80 | 1.13 |
| Hurricane seeding | 1.98 | 1.15 | 2.35 | 1.24 | 3.05 | 1.14 | 2.77 | 1.19 |
| Goose imprinting | 2.29 | 1.05 | 2.29 | 1.13 | 2.25 | 1.14 | 2.07 | 1.19 |
| Y-test | 1.88 | 1.12 | 1.91 | 1.06 | 2.76 | 1.13 | 2.27 | 1.05 |
| **Confidence** | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| Virgin rat | 3.63 | 1.473 | 3.31 | 1.42 | 3.63 | 1.473 | 3.96 | 1.44 |
| Hurricane seeding | 3.32 | 1.59 | 3.53 | 1.32 | 3.32 | 1.59 | 3.25 | 1.40 |
| Goose imprinting | 3.42 | 1.43 | 3.74 | 1.42 | 3.42 | 1.43 | 3.64 | 1.39 |
| Y-test | 3.61 | 1.39 | 3.70 | 1.33 | 3.61 | 1.39 | 3.21 | 1.41 |
| | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| **Task Difficulty** | 4.68 | 1.44 | 4.25 | 1.48 | 4.68 | 1.44 | 4.30 | 1.65 |

*Note*. Surprise ratings: 1 = *not surprising at all*, 5 = *extremely surprising*. Confidence ratings: 0 = *extremely not confident*, 6 = *extremely confidence*. Task difficulty: 1 = *extremely easy*, 7 = *extremely difficult*. The foresight ratings of both outcome A and outcome B came from the same participants in the foresight condition. The hindsight ratings came from participants in the hindsight outcome A condition or the hindsight outcome B condition, respectively. Hindsight participants only rated their surprise levels of the outcome which they knew had occurred in the initial trial.

Table S26

*Study 2: Independent Samples Student's T-Tests of Surprise, Confidence, and Task Difficulty between Foresight and Hindsight (Outcome A/B) Conditions (Younger Participants Aged Between 18 and 31 Years Old)*

| **Hindsight vs. Foresight** | ***Mean Difference*** | ***t*** | ***df*** | ***p*** | $p_{adjusted}$ | **Cohen's *d*** |
|---|---|---|---|---|---|---|
| **Surprise** | | | | | | |
| Virgin rat – Outcome A | -0.14 | -0.62 | 137 | .534 | .854 | -0.11 |
| Virgin rat – Outcome B | 0.04 | 0.20 | 113 | .838 | .991 | 0.04 |
| Hurricane seeding – Outcome A | 0.37 | 1.77 | 137 | .078 | .312 | 0.30 |
| Hurricane seeding – Outcome B | -0.28 | -1.30 | 113 | .195 | .520 | -0.24 |
| Goose imprinting – Outcome A | 0.00 | 0.00 | 137 | .997 | .997 | 0.00 |
| Goose imprinting – Outcome B | -0.18 | -0.84 | 113 | .401 | .802 | -0.16 |
| Y-test – Outcome A | 0.03 | 0.17 | 137 | .867 | .991 | 0.03 |
| Y-test – Outcome B | -0.49 | -2.42 | 113 | .017 | .136 | -0.45 |
| **Confidence** | | | | | | |
| Virgin rat –Outcome A | -0.31 | -1.27 | 137 | .206 | .434 | -0.22 |
| Virgin rat – Outcome B | 0.34 | 1.24 | 113 | .217 | .434 | 0.23 |
| Hurricane seeding – Outcome A | 0.20 | 0.82 | 137 | .413 | .551 | 0.14 |
| Hurricane seeding – Outcome B | -0.07 | -0.26 | 113 | .798 | .798 | -0.05 |
| Goose imprinting – Outcome A | 0.31 | 1.28 | 137 | .201 | .434 | 0.22 |
| Goose imprinting – Outcome B | 0.22 | 0.83 | 113 | .407 | .551 | 0.16 |
| Y-test – Outcome A | 0.09 | 0.39 | 137 | .699 | .798 | 0.07 |
| Y-test – Outcome B | -0.40 | -1.52 | 113 | .132 | .434 | -0.28 |
| **Task difficulty** | | | | | | |
| Outcome A | -0.43 | -1.70 | 137 | .091 | - | -0.29 |
| Outcome B | -0.37 | -1.30 | 113 | .197 | - | -0.24 |

*Note.* Bolded options indicate the pairs of comparisons of interest. The Levene's test of equal variance was nonsignificant for all pairs of comparison. *p* values were adjusted using the Benjamini and Hochberg (1995) false discovery rate control method.

Table S27

*Study 2: Mean Probabilities in Future Trials (in percentage %) (Older Participants >= 50 Years Old)*

| | Foresight | | | Hindsight | | |
|---|---|---|---|---|---|---|
| Initial result and kind of replication | *N* | *Mean* | *SD* | *N* | *Mean* | *SD* |
| **Virgin rat experiment** | | | | | | |
| Outcome A: Shows maternal behavior | | | | | | |
| **d. All show maternal behavior** | | **36.00** | 31.93 | | **47.77** | 35.50 |
| e. Some show maternal behavior | 35 | 30.83 | 28.17 | 39 | 30.92 | 27.82 |
| f. None show maternal behavior | | 33.17 | 33.84 | | 21.31 | 29.96 |
| Outcome B: Fails to show maternal behavior | | | | | | |
| d. All show maternal behavior | | 14.91 | 20.32 | | 16.56 | 30.44 |
| e. Some show maternal behavior | 35 | 24.20 | 25.08 | 41 | 22.34 | 26.49 |
| **f. None show maternal behavior** | | **60.89** | 32.74 | | **61.10** | 38.00 |
| **Hurricane seeding experiment** | | | | | | |
| Outcome A: Intensity increases | | | | | | |
| **d. All increase** | | **45.29** | 31.10 | | **48.92** | 32.83 |
| e. Some increase | 35 | 33.31 | 23.69 | 39 | 32.69 | 27.56 |
| f. None increase | | 21.40 | 24.56 | | 18.38 | 27.12 |
| Outcome B: Intensity weakens | | | | | | |
| **d. All weaken** | | **32.94** | 28.57 | | **37.46** | 32.89 |
| e. Some weaken | 35 | 34.37 | 23.00 | 41 | 39.80 | 33.33 |
| f. None weaken | | 32.69 | 29.66 | | 22.73 | 29.39 |
| **Gosling imprinting experiment** | | | | | | |
| Outcome A: Approaches duck | | | | | | |
| **d. All approach duck** | | **45.97** | 31.11 | | **49.18** | 34.40 |
| e. Some approach duck | 35 | 30.29 | 20.90 | 39 | 36.85 | 29.58 |
| f. None approach duck | | 23.74 | 29.61 | | 13.97 | 22.03 |
| Outcome B: Approaches goose | | | | | | |
| **d. All approach goose** | | **34.31** | 32.30 | | **52.76** | 41.34 |
| e. Some approach goose | 35 | 40.46 | 28.22 | 41 | 31.44 | 35.38 |
| f. None approach goose | | 25.23 | 25.07 | | 15.80 | 27.26 |
| **Y-test experiment** | | | | | | |
| Outcome A: Places dot in Area A | | | | | | |
| **d. Places in Area A** | | **62.12** | 23.18 | | **68.82** | 20.63 |
| e. Places in Area B | 35 | 9.32 | 10.37 | 39 | 12.46 | 17.63 |
| f. Places in Area C | | 28.55 | 20.29 | | 18.72 | 13.65 |
| Outcome B: Places dot in Area C | | | | | | |
| d. Places in Area A | | 57.64 | 23.59 | | 51.22 | 26.94 |
| e. Places in Area B | 35 | 10.18 | 11.91 | 41 | 7.66 | 12.44 |
| **f. Places in Area C** | | **32.18** | 20.60 | | **41.12** | 27.56 |

*Note.* Options and numbers marked in bold represent the kind of replication that was reported to have occurred in the initial trial (hindsight) or could possibly occur in the initial trial (foresight). The foresight ratings of both outcome A and outcome B came from the same participants in the foresight condition. The hindsight ratings came from participants in the hindsight outcome A condition or the hindsight outcome B condition, respectively.

Table S28

*Study 2: Independent Samples Student's T-Tests of Probability Estimates between Foresight and Hindsight (Outcome A/B) Conditions (Older Participants >= 50 Years Old)*

| **Hindsight vs. Foresight** | ***Mean Difference*** | ***t*** | ***df*** | ***p*** | $p_{adjusted}$ | **Cohen's d** |
|---|---|---|---|---|---|---|
| Virgin rat experiment | | | | | | |
| Outcome A: Shows maternal behavior | | | | | | |
| **a. All show maternal behavior** | 11.77 | 1.49 | 72 | .140 | .441 | 0.35 |
| b. Some show maternal behavior | 0.09 | 0.01 | 72 | .988 | .988 | 0.00 |
| c. None show maternal behavior | -11.86 | -1.60 | 72 | .114 | .441 | -0.37 |
| Outcome B: Fails to show maternal behavior | | | | | | |
| a. All show maternal behavior | 1.65 | 0.27 | 74 | .786 | .898 | 0.06 |
| b. Some show maternal behavior | -1.86 | -0.31 | 74 | .756 | .898 | -0.07 |
| **c. None show maternal behavior** | 0.21 | 0.03 | 74 | .979 | .988 | 0.01 |
| Hurricane seeding experiment | | | | | | |
| Outcome A: Intensity increases | | | | | | |
| **a. All increases** | 3.64 | 0.49 | 72 | .627 | .836 | 0.11 |
| b. Some increases | -0.62 | -0.10 | 72 | .918 | .988 | -0.02 |
| c. None increases | -3.02 | -0.50 | 72 | .619 | .836 | -0.12 |
| Outcome B: Intensity weakens | | | | | | |
| **a. All weaken** | 4.52 | 0.63 | 74 | .528 | .792 | 0.15 |
| b. Some weaken a | 5.43 | 0.81 | 74 | .419 | .670 | 0.19 |
| c. None weaken | -9.95 | -1.47 | 74 | .147 | .441 | -0.34 |
| Gosling imprinting experiment | | | | | | |
| Outcome A: Approaches duck | | | | | | |
| **a. All approach duck** | 3.21 | 0.42 | 72 | .677 | .855 | 0.10 |
| b. Some approach duck a | 6.56 | 1.09 | 72 | .279 | .558 | 0.25 |
| c. None approach duck a | -9.77 | -1.62 | 72 | .109 | .441 | -0.38 |
| Outcome B: Approaches goose | | | | | | |
| **a. All approach goose*** a | 18.44 | 2.14 | 74 | .036 | .432 | 0.49 |
| b. Some approach goose | -9.02 | -1.21 | 74 | .229 | .550 | -0.28 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| c. | None approach goose | -9.42 | -1.56 | 74 | .123 | .441 | -0.36 |
| Y-test experiment | | | | | | | |
| Outcome A: Places dot in Area A | | | | | | | |
| **a.** | **Places in Area A** | 6.70 | 1.32 | 72 | .193 | .515 | 0.31 |
| b. | Places in Area B a | 3.14 | 0.92 | 72 | .361 | .638 | 0.21 |
| c. | Places in Area C | -9.84 | -2.47 | 72 | .016 | .384 | -0.57 |
| Outcome B: Places dot in Area C | | | | | | | |
| a. | Places in Area A | -6.42 | -1.10 | 74 | .277 | .558 | -0.25 |
| b. | Places in Area B | -2.52 | -0.90 | 74 | .372 | .638 | -0.21 |
| **c.** | **Places in Area C** | 8.94 | 1.58 | 74 | .119 | .441 | 0.36 |

*Note.* Bolded options indicate the pairs of comparisons of interest. a. Levene's test of equal variance was significant. *p* values were adjusted using the Benjamini and Hochberg (1995) false discovery rate control method.

Table S29

*Study 2: Means and Standard Deviations of Surprise Ratings, Confidence Ratings, and Task Difficulty (Older Participants >= 50 Years Old)*

| **Scenario** | **Outcome A** | | | | **Outcome B** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Foresight** | | **Hindsight** | | **Foresight** | | **Hindsight** | |
| **Surprise** | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| Virgin rat | 3.03 | 1.44 | 2.51 | 1.41 | 1.86 | 1.19 | 1.54 | 0.90 |
| Hurricane seeding | 2.14 | 1.09 | 2.05 | 1.21 | 2.74 | 1.34 | 2.66 | 1.32 |
| Goose imprinting | 2.29 | 1.41 | 1.79 | 1.06 | 2.26 | 1.24 | 1.85 | 1.22 |
| Y-test | 1.69 | 1.08 | 1.59 | 0.97 | 2.49 | 1.27 | 2.07 | 1.21 |
| **Confidence** | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| Virgin rat | 3.51 | 1.36 | 3.38 | 1.70 | 3.51 | 1.36 | 4.17 | 1.63 |
| Hurricane seeding | 2.91 | 1.74 | 3.03 | 1.95 | 2.91 | 1.74 | 3.46 | 1.70 |
| Goose imprinting | 3.26 | 1.75 | 3.59 | 1.76 | 3.26 | 1.75 | 3.95 | 1.77 |
| Y-test | 3.80 | 1.37 | 3.79 | 1.69 | 3.80 | 1.37 | 3.61 | 1.46 |
| | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| **Task Difficulty** | 5.23 | 1.42 | 4.74 | 1.77 | 5.23 | 1.42 | 4.63 | 1.61 |

*Note*. Surprise ratings: 1 = *not surprising at all*, 5 = *extremely surprising*. Confidence ratings: 0 = *extremely not confident*, 6 = *extremely confidence*. Task difficulty: 1 = *extremely easy*, 7 = *extremely difficult*. The foresight ratings of both outcome A and outcome B came from the same participants in the foresight condition. The hindsight ratings came from participants in the hindsight outcome A condition or the hindsight outcome B condition, respectively. Hindsight participants only rated their surprise levels of the outcome which they knew had occurred in the initial trial.

Table S30

*Study 2: Independent Samples Student's T-Tests of Surprise, Confidence, and Task Difficulty between Foresight and Hindsight (Outcome A/B) Conditions (Older Participants >= 50 Years Old)*

| **Hindsight vs. Foresight** | ***Mean Difference*** | ***t*** | ***df*** | ***p*** | $p_{adjusted}$ | **Cohen's *d*** |
|---|---|---|---|---|---|---|
| **Surprise** | | | | | | |
| Virgin rat –Outcome A | -0.52 | -1.55 | 72 | .125 | .298 | -0.36 |
| Virgin rat – Outcome B | -0.32 | -1.34 | 74 | .186 | .298 | -0.31 |
| Hurricane seeding – Outcome A | -0.09 | -0.34 | 72 | .735 | .783 | -0.08 |
| Hurricane seeding – Outcome B | -0.08 | -0.28 | 74 | .783 | .783 | -0.06 |
| Goose imprinting – Outcome A | -0.49 | -1.71 | 72 | .092 | .298 | -0.40 |
| Goose imprinting – Outcome B | -0.40 | -1.43 | 74 | .158 | .298 | -0.33 |
| Y-test – Outcome A | -0.10 | -0.40 | 72 | .687 | .783 | -0.09 |
| Y-test – Outcome B | -0.41 | -1.45 | 74 | .152 | .298 | -0.33 |
| **Confidence** | | | | | | |
| Virgin rat –Outcome A | -0.13 | -0.36 | 72 | .720 | .911 | -0.08 |
| Virgin rat – Outcome B | 0.66 | 1.89 | 74 | .063 | .368 | 0.43 |
| Hurricane seeding – Outcome A | 0.11 | 0.26 | 72 | .797 | .911 | 0.06 |
| Hurricane seeding – Outcome B | 0.55 | 1.39 | 74 | .170 | .453 | 0.32 |
| Goose imprinting – Outcome A | 0.33 | 0.81 | 72 | .419 | .838 | 0.19 |
| Goose imprinting – Outcome B | 0.69 | 1.71 | 74 | .092 | .368 | 0.39 |
| Y-test – Outcome A | -0.01 | -0.01 | 72 | .989 | .989 | 0.00 |
| Y-test – Outcome B | -0.19 | -0.58 | 74 | .562 | .899 | -0.13 |
| **Task difficulty** | | | | | | |
| Outcome A | -0.48 | -1.29 | 72 | .201 | - | -0.30 |
| Outcome B | -0.59 | -1.70 | 74 | .094 | - | -0.39 |

*Note.* Bolded options indicate the pairs of comparisons of interest. a. Levene's test of equal variance was significant. *$p$ < .05, **$p$ < .01, ***$p$ < .001. $p$ values were adjusted using the Benjamini and Hochberg (1995) false discovery rate control method.

Table S31 showed a series of moderation analyses using experimental condition as the independent variable, probability estimates, surprise ratings, confidence ratings, and task difficulty as dependent variables, and age as the moderator. None of the moderation analyses were significant (even before adjusting the *p* values for multiple testing, $n$ = 400~401 for each moderation analysis). The findings therefore suggest that the hindsight bias (based on probability estimates) and the findings related to surprise ratings, confidence ratings, and task difficulty in Study 2 were not contingent on participants' age.

Table S31

*Study 2: Age as Moderator (n = 604)*

| | | Probability (Key Comparison) | | | Surprise | | | Confidence | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *B* | *SD* | *p* | *B* | *SD* | *p* | *B* | *SD* | *p* |
| Virgin rat – Outcome A | Constant | 29.13 | 2.04 | .000 | 3.13 | 0.09 | .000 | 3.61 | 0.11 | .000 |
| | Condition | 9.41 | 2.86 | .001 | -0.20 | 0.13 | .126 | -0.43 | 0.16 | .006 |
| | Age | 0.14 | 0.19 | .463 | 0.00 | 0.01 | .803 | 0.00 | 0.01 | .916 |
| | Condition * Age | 0.05 | 0.24 | .853 | -0.01 | 0.01 | .454 | 0.01 | 0.01 | .623 |
| Virgin rat – Outcome B | Constant | 54.13 | 2.35 | .000 | 1.75 | 0.07 | .000 | 3.61 | 0.11 | .000 |
| | Condition | 6.00 | 3.29 | .070 | -0.18 | 0.10 | .081 | 0.30 | 0.15 | .050 |
| | Age | 0.39 | 0.21 | .071 | 0.00 | 0.01 | .985 | 0.00 | 0.01 | .914 |
| | Condition * Age | -0.22 | 0.29 | .434 | -0.01 | 0.01 | .351 | 0.00 | 0.01 | .837 |
| Hurricane seeding – Outcome A | Constant | 47.73 | 2.10 | .000 | 2.03 | 0.08 | .000 | 3.27 | 0.12 | .000 |
| | Condition | 1.63 | 2.95 | .581 | 0.10 | 0.12 | .403 | 0.12 | 0.16 | .481 |
| | Age | 0.06 | 0.19 | .770 | 0.00 | 0.01 | .983 | -0.01 | 0.01 | .417 |
| | Condition * Age | -0.04 | 0.25 | .871 | -0.01 | 0.01 | .517 | 0.00 | 0.01 | .982 |
| Hurricane seeding – Outcome B | Constant | 29.56 | 1.85 | .000 | 3.01 | 0.09 | .000 | 3.27 | 0.11 | .000 |
| | Condition | 4.41 | 2.60 | .091 | -0.34 | 0.12 | .005 | -0.03 | 0.16 | .860 |
| | Age | 0.14 | 0.17 | .400 | -0.01 | 0.01 | .186 | -0.01 | 0.01 | .395 |
| | Condition * Age | -0.10 | 0.23 | .660 | 0.01 | 0.01 | .449 | 0.02 | 0.01 | .258 |
| Goose imprinting – Outcome A | Constant | 39.10 | 2.08 | .000 | 2.20 | 0.08 | .000 | 3.41 | 0.11 | .000 |
| | Condition | 6.32 | 2.91 | .030 | -0.12 | 0.12 | .289 | 0.08 | 0.16 | .625 |
| | Age | 0.24 | 0.19 | .204 | 0.00 | 0.01 | .715 | 0.00 | 0.01 | .671 |
| | Condition * Age | 0.00 | 0.25 | .987 | -0.01 | 0.01 | .411 | 0.00 | 0.01 | .839 |
| Goose imprinting – Outcome B | Constant | 38.08 | 2.26 | .000 | 2.16 | 0.08 | .000 | 3.41 | 0.11 | .000 |
| | Condition | 8.13 | 3.18 | .011 | -0.26 | 0.11 | .024 | 0.26 | 0.16 | .095 |
| | Age | 0.07 | 0.21 | .740 | 0.00 | 0.01 | .768 | 0.00 | 0.01 | .666 |
| | Condition * Age | 0.27 | 0.28 | .323 | 0.00 | 0.01 | .762 | 0.01 | 0.01 | .532 |
| Y-test – Outcome A | Constant | 59.61 | 1.65 | .000 | 1.81 | 0.07 | .000 | 3.52 | 0.10 | .000 |
| | Condition | 2.55 | 2.32 | .272 | -0.16 | 0.10 | .109 | 0.12 | 0.15 | .416 |
| | Age | 0.10 | 0.15 | .508 | -0.01 | 0.01 | .279 | 0.01 | 0.01 | .368 |
| | Condition * Age | 0.19 | 0.20 | .337 | 0.00 | 0.01 | .941 | 0.00 | 0.01 | .797 |
| Y-test – Outcome B | Constant | 33.81 | 1.58 | .000 | 2.46 | 0.08 | .000 | 3.52 | 0.10 | .000 |
| | Condition | 4.86 | 2.22 | .029 | -0.32 | 0.11 | .004 | -0.18 | 0.14 | .221 |
| | Age | -0.18 | 0.14 | .218 | -0.01 | 0.01 | .302 | 0.01 | 0.01 | .358 |
| | Condition * Age | 0.29 | 0.19 | .140 | 0.00 | 0.01 | .948 | 0.00 | 0.01 | .874 |

| | Task Difficulty (Outcome A) | | | Task Difficulty (Outcome B) | | |
|---|---|---|---|---|---|---|
| | *B* | *SD* | *p* | *B* | *SD* | *p* |
| Constant | 4.98 | .11 | .000 | 4.98 | .10 | .000 |
| Condition | -.56 | .15 | .000 | -.59 | .15 | .000 |
| Age | .02 | .01 | .082 | .02 | .01 | .072 |
| Condition * Age | .00 | .01 | .980 | -.01 | .01 | .629 |

*Note.* The sample size for moderation analyses involving Outcome A was 401, and the sample size for Outcome B was 400.

## Studies 1 & 2: Summary of Extension Hypotheses and Exclusion Criteria in Pre-registrations

For Study 1, two groups of students independently pre-registered the replication and extension experiment. For Study 2, four students independently pre-registered the replication and extension experiment. We consider all pre-registrations equally important, and therefore included all of them in Supplementary Materials and OSF. We see this design of having multiple independent pre-registrations as a strength, as it helps us cross-check pre-registrations and help students learn from each other at the data analysis stage.

Among the different versions of pre-registrations, students agreed on the power analyses and the proposed methods for the main analyses. But the exclusion criteria and exploratory analyses differed. Our policy is that as long as one pre-registration included the exclusion criteria, we would perform the analyses and report the results in this Supplementary Materials. See Table S32 for a summary of these differences.

Note that most of the extension statistics proposed in students' pre-registrations were Mann-Whitney *U* tests, *t* tests, and correlations, which are basic and limited in their capability to reveal underlying mechanisms. In our final manuscript, we decided to test the moderating and mediating effects of surprise, confidence, and task difficulty, as explained in the *Changes after pre-registration* of this Supplementary Materials. Tests of the extension hypotheses proposed in students' original pre-registrations are reported in the *Additional analyses* section of this Supplementary Materials.

Table S32

*Studies 1 and 2: Extension Hypotheses and Exclusion Criteria in Pre-registrations*

| Pre-registrations | Additional variables | Hypotheses | Exclusion criteria |
|---|---|---|---|
| **Study 1** | | | |
| Au, S. S. Y | Surprise | • Compared with participants in the Before group, participants in the after (ignore) groups will have lower surprise ratings regarding the outcome that they were provided with outcome knowledge, even when they were told to answer as if they had not known what happened.<br>• For after (ignore) groups, there will be a negative correlation between surprise ratings and probability estimates. | - English proficiency (smaller than 5 on a 1-7 point Likert scale).<br>- Serious about study (smaller than 4 on a 1-5 point Likert scale).<br>- Correctly guessed the hypothesis of this study in the funneling section.<br>- Failed to complete the survey. |
| Choi, H. Y. & Hayley, A. | Surprise | • There is a difference in the level of surprise between the After (ignore) groups and the Before group on the same outcome of the same event.<br>• Participants in the After (ignore) groups have a low level of surprise (compared to the midpoint, $\mu < 4$).<br>• There is a relationship between the level of surprise and probability estimates on outcome knowledge ($\rho \neq 0$). | Same as above. |
| **Study 2** | | | |
| Kwan, L. C. | Surprise; task difficulty | • Participants in the foresight group will feel more surprised than participants in the hindsight groups. High levels of surprise will also lead to a decrease or reversal of hindsight bias.<br>• Participants in the foresight group will feel it is more difficult to estimate the outcome than participants in the hindsight groups. | - English proficiency (smaller than 5 on a 1-7 point Likert scale).<br>- Serious about study (smaller than 4 on a 1-5 point Likert scale).<br>- Correctly guessed the hypothesis of this study in the funneling section.<br>- Failed to complete the survey. |
| Lo, Y. C. | Confidence | • Participants in the hindsight conditions will report higher confidence in the accuracy of their predictions than participants in the foresight condition. | Same as above. |
| Ma, L. L. Y. | Estimated probabilities of outcomes in the first trial (dropped) [a]; confidence | • Participants in the hindsight condition will estimate the probability of outcome of the first trial to be higher than participants in the foresight condition do, even if they are asked to answer as if they had not known the outcome of the first trial (dropped).<br>• Participants in the hindsight condition will feel more confident about their answers than participants in the foresight condition. | No exclusion criteria were specified. |
| Tsang, Chi Ho | Not applicable. | No extension hypotheses. | Same as Kwan, L. C.'s. |

*Note*. a. In Study 2, we dropped this proposed measure in the actual survey, and focused on surprise, confidence, and task difficulty.

# Study 3

## Study 3: Transparency report

### PREREGISTRATION SECTION

(7) Prior to analyzing the complete data set, a time-stamped preregistration was posted in an independent, third-party registry for the data analysis plan. **Yes**

(8) The manuscript includes a URL to all preregistrations that concern the present study. **Yes**

(9) The study was preregistered… **before any data were collected**

**The preregistration fully describes…**

(24) all inclusion and exclusion criteria for participation (e.g., English speakers who achieved a certain cutoff score in a language test). **Yes**

(25) all procedures for assigning participants to conditions. **Yes**

(26) all procedures for randomizing stimulus materials. **Yes**

(27) any procedures for ensuring that participants, experimenters, and data-analysts were kept naive (blinded) to potentially biasing information. **Yes**

(28) a rationale for the sample size used (e.g., an a priori power analysis). **Yes**

(29) the measures of interest (e.g., friendliness). **Yes**

(30) all operationalizations for the measures of interest (e.g., a questionnaire measuring friendliness). **Yes**

(31) the data preprocessing plans (e.g., transformed, cleaned, normalized, smoothed). **Yes**

(32) how missing data (e.g., dropouts) were planned to be handled. **Yes**

(33) the intended statistical analysis for each research question (this may require, for example, information about the sidedness of the tests, inference criteria, corrections for multiple testing, model selection criteria, prior distributions etc.). **Yes**

**Comments about your Preregistration**

Hypothesis 9 was added after the pre-registration.

---

## METHODS SECTION

**The manuscript fully describes…**

(38) the rationale for the sample size used (e.g., an a priori power analysis). **Yes**

(39) how participants were recruited. **Yes**

(40) how participants were selected (e.g., eligibility criteria). **Yes**

(41) what compensation was offered for participation. **No**

(42) how participant dropout was handled (e.g., replaced, omitted, etc.). **Yes**

(43) how participants were assigned to conditions. **Yes**

(44) how stimulus materials were randomized. **Yes**

(45) whether (and, if so, how) participants, experimenters, and data-analysts were kept naive to potentially biasing information. **NA**

(46) the study design, procedures, and materials to allow independent replication. **Yes**

(47) the measures of interest (e.g., friendliness). **Yes**

(48) all operationalizations for the measures of interest (e.g., a questionnaire measuring friendliness). **Yes**

(49) any changes to the preregistration (such as changes in eligibility criteria, group membership cutoffs, or experimental procedures)? **Yes**

**Comments about your Methods section**

No comments.

---

## RESULTS AND DISCUSSION SECTION

**The manuscript…**

(38) distinguishes explicitly between “confirmatory” (i.e., prespecified) and “exploratory” (i.e., not prespecified) analyses. **Yes**

(39) describes how violations of statistical assumptions were handled. **Yes**

(40) justifies all statistical choices (e.g., including or excluding covariates; applying or not applying transformations; use of multi-level models vs. ANOVA). **Yes**

(41) reports the sample size for each cell of the design. **Yes**

(42) reports how incomplete or missing data were handled. **Yes**

(43) presents protocols for data preprocessing (e.g., cleaning, discarding of cases and items, normalizing, smoothing, artifact correction). **Yes**

**Comments about your Results and Discussion**

No comments.

---

## DATA, CODE, AND MATERIALS AVAILABILITY SECTION

**The following have been made publicly available…**

(42) the (processed) data, on which the analyses of the manuscript were based. **Yes**

(43) all code and software (that is not copyright protected). **Yes**

(44) all instructions, stimuli, and test materials (that are not copyright protected). **Yes**

(45) Are the data properly archived (i.e., would a graduate student with relevant background knowledge be able to identify each variable and reproduce the analysis)? **Yes**

(46) The manuscript includes a statement concerning the availability and location of all research items, including data, materials, and code relevant to the study. **Yes**

**Comments about your Data, Code, and Materials**

No comments.

## Study 3: Power analysis

We estimated the effect size through pretests. We conducted two pretests, and recruited about 30 participants for each pretest from CloudResearch. In the first pretest, we tested the earliest version of the materials. In the second pretest, like Slovic and Fischhoff (1977), we added the questions asking participants to write down the reasons for their predictions. We included these questions to reinforce participants' sense-making activities. The second pretest is different from our formal study in that in the second pretest, participants in the hindsight conditions reported their surprise about the known outcome only, while in the formal study, participants in the hindsight conditions reported their surprise about both the known outcome and the other outcome.

Table S33 presents the means, standard deviations, and Cohen's *d* of the two pretests regarding probability estimates. We expected that participants in the Hindsight Outcome Success condition would have the highest predictions of a successful replication, followed by those in the Foresight condition, and lastly those in the Hindsight Outcome Fail condition. The absolute values of Cohen's *d*s ranged from 0.24 to 0.83. Five out of the six pairwise comparisons had the expected sign, although none of the pairwise comparisons were significant judged by 95% confidence interval of Cohen's *d* (which is understandable given the very small sample sizes and low statistical power). We took the average of the Cohen's *d*s, which equals (|0.24| + |-0.27| + |0.53| - |-0.32| + |-0.83| + |0.42|)/6 = 0.33. Note that we minus the Cohen's *d* of the Hindsight Outcome Success vs Foresight comparison in the second pretest, because its sign was opposite to our expectation. We suspected that such incidents of opposite signs would be rare, and decided to calculate the required sample size based on an estimated effect size of $d = 0.4$.

Table S33

*Study 3: Means, Standard Deviations, and Effect Sizes of Pretests*

| **Pilot Samples** | Foresight | | | Hindsight Outcome Success | | | Hindsight Outcome Failure | | |
|---|---|---|---|---|---|---|---|---|---|
| | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* | *n* | *Mean* | *SD* |
| First pretest | 12 | 58.33 | 20.71 | 9 | 63.00 | 17.42 | 9 | 52.78 | 20.93 |
| Second pretest | 13 | 65.38 | 14.21 | 9 | 60.00 | 20.62 | 6 | 50.83 | 23.75 |

| **Pairwise Comparisons** | Cohen's *d* | 95% CI LL of *d* | 95% CI UL of *d* |
|---|---|---|---|
| **First pretest** | | | |
| Hindsight Outcome Success vs Foresight | 0.24 | -0.63 | 1.11 |
| Hindsight Outcome Fail vs Foresight | -0.27 | -1.13 | 0.60 |
| Hindsight Outcome Success vs Hindsight Outcome Fail | 0.53 | -0.42 | 1.46 |
| **Second pretest** | | | |
| Hindsight Outcome Success vs Foresight | -0.32 | -1.17 | 0.54 |
| Hindsight Outcome Fail vs Foresight | -0.83 | -1.82 | 0.19 |
| Hindsight Outcome Success vs Hindsight Outcome Fail | 0.42 | -0.63 | 1.46 |

We used G*Power 3.1 (Faul et al., 2009) to estimate the required effect size (see Table S34). To achieve a power of .95 with an alpha of .05 (two-tailed), the sample size required per group is 164. Because there are three conditions (Foresight, Hindsight Outcome A, Hindsight Outcome B), the total sample size required is 164 * 3 = 492. In anticipation of careless responses and expectancies, we plan to recruit about 10 more participants per condition. The total planned number of participants is 520.

Table S34

*Study 3: Sample Size Calculation*

| **t tests**- Means: Difference between two independent means (two groups) | | | |
|---|---|---|---|
| **Analysis**: | A priori: Compute required sample size | | |
| **Input**: | Tail(s) | = | Two |
| | Effect size d | = | 0.4 |
| | α err prob | = | 0.05 |
| | Power (1-βerr prob) | = | 0.95 |

| | | |
|---|---|---|
| | Allocation ratio N2/N1 | = 1 |
| **Output**: | Noncentrality parameter δ | = 3.6221541 |
| | Critical t | = 1.9672675 |
| | Df | = 326 |
| | Sample size group 1 | = 164 |
| | Sample size group 2 | = 164 |
| | Total sample size | = 328 |
| | Actual power | = 0.9506816 |

## Study 3: Changes made after the pre-registration

In Study 3, H9 but no other hypotheses were added after the pre-registration. Before the pre-registration, we ran some small-sample pretests of about 70 people to test and improve the study materials. In those pretests, we saw strong evidence for H7 (the probability estimate of a successful replication will be higher than chance) and H8 (the probability estimate of a successful replication will be higher in the Hindsight Outcome A condition than in the Hindsight Outcome B condition). However, when comparing the probability estimates between the hindsight conditions and the foresight condition in the pretest data, we initially misspecified the model by comparing the mean probability estimates of a successful replication in the two hindsight conditions and the mean probability estimate of a successful replication in the foresight condition. Such a comparison was nonsignificant in the pretest data, leading us to suspect that the effect size might be a very small one and thus did not pre-register any hypothesis about the comparison between the hindsight conditions and the foresight condition. We only realized that this comparison was misspecified and would be nonsignificant because a positive difference between Hindsight Outcome A condition and Foresight condition and a negative difference between Hindsight Outcome B condition and Foresight condition would cancel off each other after we have completed the pre-registration. We therefore added a hypothesis about the correctly specified model (H9: the probability estimate of a successful replication will be higher in the Hindsight Outcome A condition than in the Foresight condition) to this study after the pre-registration.

## Study 3: Study materials

### Foresight Condition

Material

In recent years, the discipline of psychology has undergone a **replication crisis**, where many famous, long-established phenomena—ideas written in textbooks and presented in TED Talks—were found to be non-replicable.

**Hindsight bias** is a long-established phenomenon that has not been tested for replicability. It refers to people's tendency to perceive an event as more predictable after being informed of its outcome.

**Fischhoff's (1975) study** was among the first to investigate hindsight bias. In Fischhoff's (1975) study, participants were invited to read the background information of an event, and then estimate the probability of four possible outcomes. Participants were assigned to one of the two conditions: those in the ***Foresight*** condition did not know which outcome actually occurred; those in the ***Hindsight*** condition were informed of the actual outcome, but were asked to answer as if they had not known the actual outcome. The study found that participants in the ***Hindsight*** condition perceived the known outcome to be more probable, compared to participants in the ***Foresight*** condition. This finding suggests that receiving outcome knowledge makes people assign a greater likelihood to the known outcome than they would otherwise do, demonstrating hindsight bias.

A group of researchers intends to perform a **replication study** of Fischhoff (1975). There are two possible outcomes:

a) the hindsight bias effect will be successfully replicated, or

b) the hindsight bias effect will fail to replicate.

Comprehension Checks

To make sure you read and understood the paragraphs above, please answer the following comprehension questions:

In **Fischhoff's (1975) original study**, which of the following group knew the actual outcome:

- o Participants in the Foresight condition.
- o Participants in the Hindsight condition.
- o The paragraphs did not tell.

What is the outcome of the **replication study**?

- o Successful replication.
- o Failed replication.
- o The paragraphs did not tell.

Reminder Message

*(Appearing on the top of the pages for the reasons and probability estimates questions.)*

In Fischhoff's (1975) study, participants in the ***Hindsight*** condition perceived the known outcome to be more probable, compared to participants in the ***Foresight*** condition. This finding suggests that receiving outcome knowledge makes people assign a greater likelihood to the known outcome than they would otherwise do, demonstrating hindsight bias.

A group of researchers intends to perform a replication study of Fischhoff (1975).

Reasons

*(The order of the following two questions about reasons were randomized inr the foresight condition.)*

1. Please write down the reasons why the replication may be **successful** in one or two sentences.

2. Please write down the reasons that the replication may **fail** in one or two sentences.

Probability Estimates

3. In light of the information appearing in the paragraphs provided, please estimate the probabilities of occurrence of the two possible outcomes in the **replication study**. There are no right or wrong answers, answer based on your intuition.

(The probabilities should sum to 100%).

The hindsight bias effect will be successfully replicated. : _____

The hindsight bias effect will fail to replicate. : _____

Total : _____

Other ratings

4. If the hindsight bias effect is **successfully replicated,** how surprised would you be?

*1 = Not surprised at all, 5 = Extremely surprised*

5. If the hindsight bias effect **fails to replicate**, how surprised would you be?

*1 = Not surprised at all, 5 = Extremely surprised*

6. How confident are you about the accuracy of your predictions on the outcome of the replication study?

*0 = Extremely not confident, 6 = Extremely confident*

7. How difficult was it to make estimations of outcomes probabilities?

*1 = Extremely easy, 7 = Extremely difficult*

**Hindsight Outcome A Condition**

Material

*Same as the information for the Foresight condition, plus the following sentence:*

The outcome of the replication study is a **successful** replication, demonstrating hindsight bias.

Comprehension Checks

*Same questions as those for the Foresight condition, but the correct answer of the second question "What is the outcome of the replication study" differed.*

Reminder Message

*Same as the information for the Foresight condition, plus the following sentence:*

The outcome of the replication study is a **successful replication**.

Reasons

1. Please write down the reasons why the replication may be **successful** in one or two sentences.

2. Please write down the reasons that the replication may **fail** in one or two sentences.

Probability Estimates

3. In light of the information appearing in the paragraphs provided, please estimate the probabilities of occurrence of the two possible outcomes in the **replication study**. There are no

right or wrong answers, answer based on your intuition.

(The probabilities should sum to 100%).

**Answer as if you do not know the outcome**, estimating the probabilities at that time before the replication study was launched.

The hindsight bias effect will be successfully replicated. : _____

The hindsight bias effect will fail to replicate. : _____

Total : _____

Other ratings

4. How surprised are you by the outcome of a **successful replication**?

*1 = Not surprised at all, 5 = Extremely surprised*

5. How surprised would you be if the replication study **fails**?

*1 = Not surprised at all, 5 = Extremely surprised*

6. How confident are you about the accuracy of your predictions on the outcome of the replication study?

*0 = Extremely not confident, 6 = Extremely confident*

7. How difficult was it to make estimations of outcomes probabilities?

*1 = Extremely easy, 7 = Extremely difficult*

**Hindsight Outcome B Condition**

Material

*Same as the information for the Foresight condition, plus the following sentence:*

The outcome is a **failed replication**. There is no evidence for hindsight bias in the replication study.

Comprehension Checks

*Same questions as those for the Foresight condition, but the correct answer of the second question "What is the outcome of the replication study" differed.*

Reminder Message

*Same as the information for the Foresight condition, plus the following sentence:*

The outcome of the replication study is a **failed replication**.

Reasons

1. Please write down the reasons that the replication may **fail** in one or two sentences.

2. Please write down the reasons why the replication may be **successful** in one or two sentences.

Probability Estimates

3. In light of the information appearing in the paragraphs provided, please estimate the probabilities of occurrence of the two possible outcomes in the **replication study**. There are no right or wrong answers, answer based on your intuition.

(The probabilities should sum to 100%).

**Answer as if you do not know the outcome**, estimating the probabilities at that time before the replication study was launched.

The hindsight bias effect will be successfully replicated. : _____

The hindsight bias effect will fail to replicate. : _____

Total : _____

Other ratings

4. How surprised are you by the outcome of a **failed replication**?

*1 = Not surprised at all, 5 = Extremely surprised*

5. How surprised would you be if the outcome is a **successful replication**?

*1 = Not surprised at all, 5 = Extremely surprised*

6. How confident are you about the accuracy of your predictions on the outcome of the replication study?

*0 = Extremely not confident, 6 = Extremely confident*

7. How difficult was it to make estimations of outcomes probabilities?

*1 = Extremely easy, 7 = Extremely difficult*

## Study 3: Sample characteristics

The majority of the participants (96.54%) were born in the United States. The rest were born in Albania, Bahamas, China, Germany, Ghana, India, Indonesia, Italy, Japan, Mexico, Philippines, Singapore, South Korea, and Syria. In terms of ethnicity, most of the participants are White/Caucasian (71.15%), followed by Black/African (11.73%), Asian (10.19%), Hispanic/Latino (5.19%), American Indian/Alaskan Native (0.58%), and Other (1.15%). Most of the participants held a Bachelor's degree (56.35%), followed by Master's degree (20.58%), high school diploma (16.92%), doctoral degree (2.12%), and Other (4.04%).

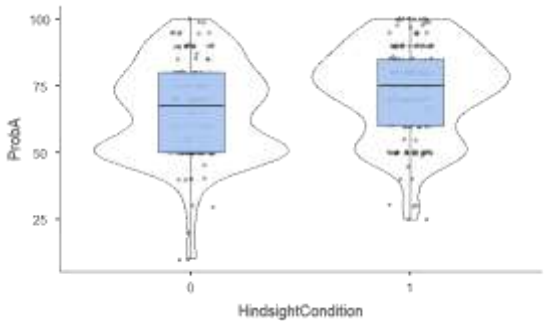
## Study 3: Additional analyses

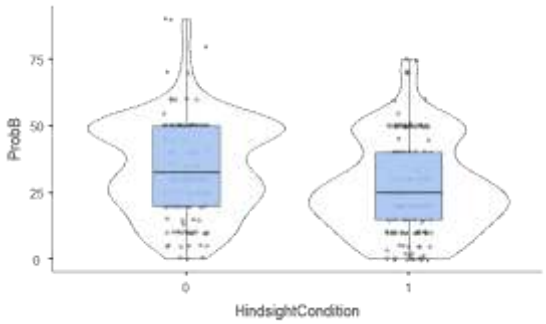
## Study 3: Violin plots

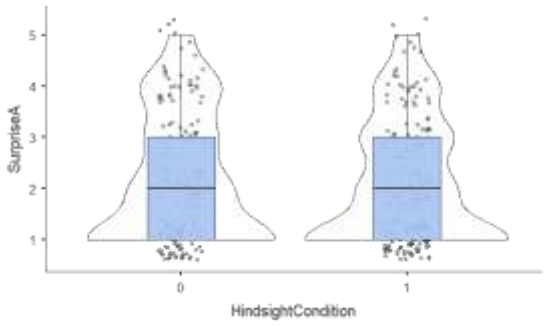
Figure S15

*Study 3: Violin Plots for Outcomes A and B*

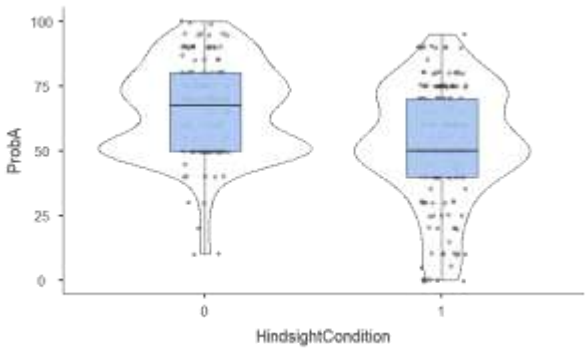
Outcome A

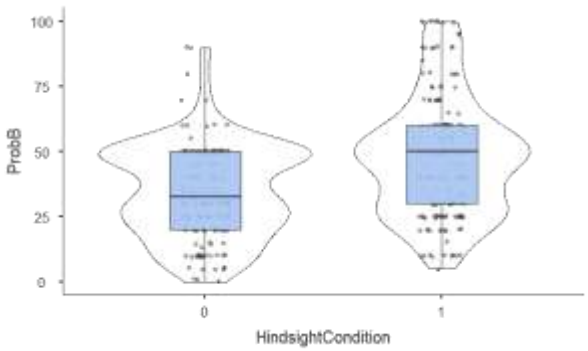


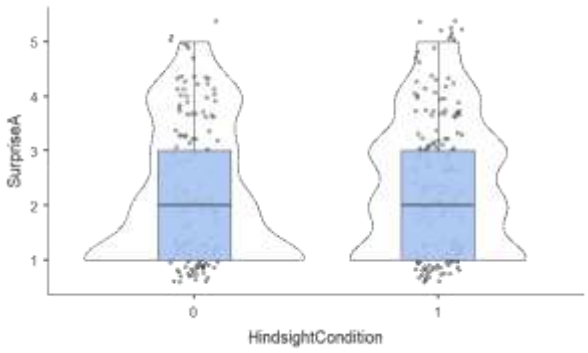



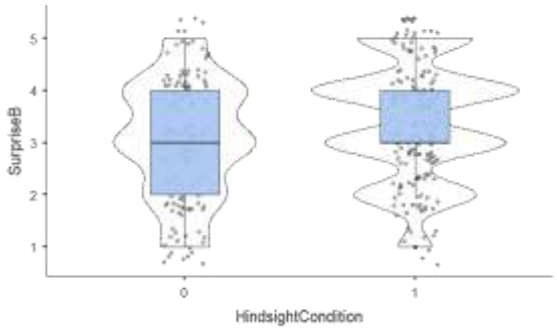



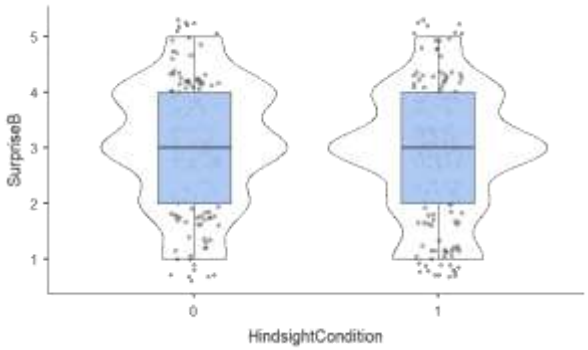

Outcome B

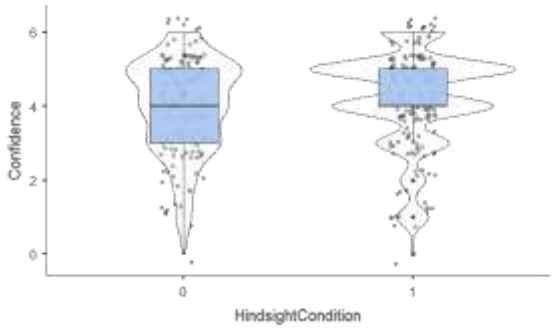


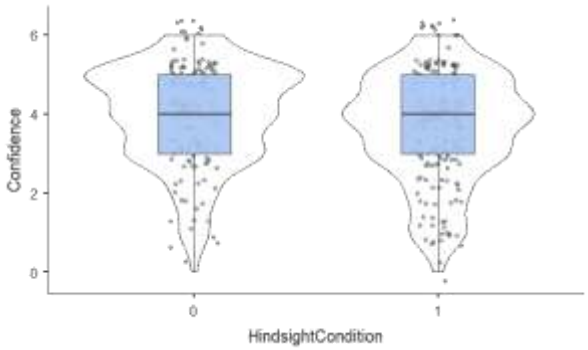



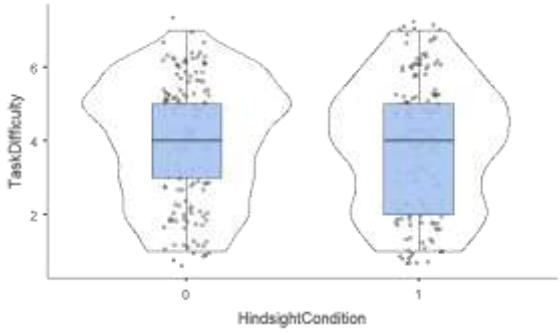



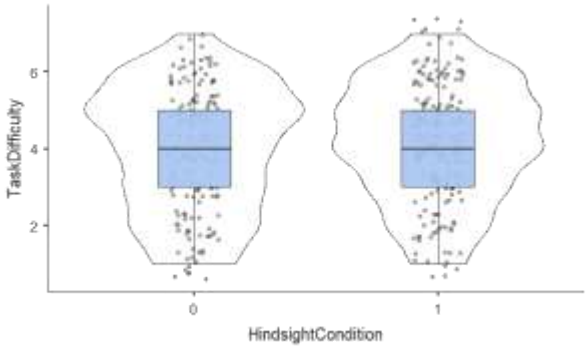

SurpriseB
HindsightCondition
0
1


SurpriseB
HindsightCondition
0
1


Confidence
HindsightCondition
0
1


Confidence
HindsightCondition
0
1


TaskDifficulty
HindsightCondition
0
1


TaskDifficulty
HindsightCondition
0
1

## Study 3: Codes for calculating confidence intervals

```
library(psych)

# Probability Estimates
cohen.d.ci(d = 0.43, n1 = 154, n2 = 178, alpha = .05)
cohen.d.ci(d = -0.64, n1 = 154, n2 = 188, alpha = .05)
cohen.d.ci(d = 1.03, n1 = 178, n2 = 188, alpha = .05)

# Surprise about successful replication
cohen.d.ci(d = -0.05, n1 = 154, n2 = 178, alpha = .05)
cohen.d.ci(d = 0.16, n1 = 154, n2 = 188, alpha = .05)
cohen.d.ci(d = -.21, n1 = 178, n2 = 188, alpha = .05)

# Surprise about failed replication
cohen.d.ci(d = .28, n1 = 154, n2 = 178, alpha = .05)
cohen.d.ci(d = -.14, n1 = 154, n2 = 188, alpha = .05)
cohen.d.ci(d = .43, n1 = 178, n2 = 188, alpha = .05)

# Confidence
cohen.d.ci(d = .14, n1 = 154, n2 = 178, alpha = .05)
cohen.d.ci(d = -.26, n1 = 154, n2 = 188, alpha = .05)
cohen.d.ci(d = .40, n1 = 178, n2 = 188, alpha = .05)

# Task difficulty
cohen.d.ci(d = -.05, n1 = 154, n2 = 178, alpha = .05)
cohen.d.ci(d = .13, n1 = 154, n2 = 188, alpha = .05)
cohen.d.ci(d = -.18, n1 = 178, n2 = 188, alpha = .05)
```

## Study 3: Results after exclusion

As stated in the preregistration, we used the full sample for the main analyses, and report the results with a restricted sample meeting the following criteria in Supplementary Materials:

1. Being serious about the study (>= 4 on a 1-5 point Likert scale).

2. Understood the English used in the study (>= 5 on a 1-5 point Likert scale)

3. Had not seen the materials used in this study

The total number of participants remained was 480 (n = 141 for Foresight condition, n = 164 for Hindsight Outcome A condition, and n = 175 for Hindsight Outcome B condition). Readers interested in reproducing the analyses can visit our OSF webpage, look for the file Fischhoff Replicability_JC.omv, and use the filter function in JAMOVI to conduct the analyses.

In a one-sample t-test, we found that participants who were informed of Outcome A (successful replication) estimated the probability of a successful replication (65.86%) to be higher than chance (50%), $t(140) = 10.55$, $p = .000$, Cohen's d = 0.89. Hypothesis 7 is supported.

In a set of independent samples t-tests (see Table S35), we found that participants who were informed of Outcome A (successful replication) estimated a successful replication to be *more* probable than participants who did not know the outcome, $t(303) = 3.92$, $p = .000$, Cohen's $d = 0.45$. In contrast, participants who were informed of Outcome B (failed replication) estimated a successful replication to be *less* probable than participants who did not know the outcome, $t(314) = -5.60$, $p = .000$, Cohen's $d = -0.63$. In addition, participants who were informed of Outcome A (successful replication) estimated a successful replication to be *more* probable than participants who were informed of Outcome B (failed replication), $t(337) = 9.51$, $p = .000$, Cohen's $d = 1.03$. The results therefore provided strong support for Hypotheses 8 and 9.

Table S35

*Study 3: Independent Samples Student's T-Tests of Estimations of Outcomes of a Replication of Fischhoff (1975)-After Exclusion*

| **Hindsight vs. Foresight** | ***Mean Difference*** | ***t*** | ***df*** | ***p*** | **Cohen's *d*** |
|---|---|---|---|---|---|
| **Estimated probabilities of successful replication** | | | | | |
| Hindsight Outcome A vs. Foresight | 7.71 | 3.92 | 303 | .000 | 0.45 |
| Hindsight Outcome B vs. Foresight | -13.15 | - 5.60 | 314 | .000 | -0.63 |
| Hindsight Outcome A vs. Hindsight Outcome B | 21.15 | 9.51 | 337 | .000 | 1.03 |
| **Surprise about successful replication** | | | | | |
| Hindsight Outcome A vs. Foresight | -0.06 | -0.78 | 303 | .437 | -0.09 |
| Hindsight Outcome B vs. Foresight | 0.20 | 1.41 | 314 | .161 | 0.16 |
| Hindsight Outcome A vs. Hindsight Outcome B | -0.32 | -2.32 | 337 | .021 | -0.25 |
| **Surprise about failed replication** | | | | | |
| Hindsight Outcome A vs. Foresight | 0.32 | 2.38 | 303 | .018 | 0.27 |
| Hindsight Outcome B vs. Foresight | -0.16 | -1.33 | 314 | .184 | -0.15 |
| Hindsight Outcome A vs. Hindsight Outcome B | 0.48 | 3.88 | 337 | .000 | 0.42 |
| **Confidence** | | | | | |
| Hindsight Outcome A vs. Foresight | 0.19 | 0.84 | 303 | .401 | 0.10 |
| Hindsight Outcome B vs. Foresight | -0.35 | -2.73 | 314 | .007 | -0.31 |
| Hindsight Outcome A vs. Hindsight Outcome B | 0.54 | 3.65 | 337 | .000 | 0.40 |
| **Task difficulty** | | | | | |
| Hindsight Outcome A vs. Foresight | -0.09 | -0.43 | 303 | .670 | -0.05 |
| Hindsight Outcome B vs. Foresight | 0.21 | 1.27 | 314 | .204 | 0.14 |
| Hindsight Outcome A vs. Hindsight Outcome B | -0.32 | 1.75 | 337 | .081 | -0.19 |

*Note.* Levene's test was nonsignificant for all comparisons. *$p < .05$, **$p < .01$, ***$p < .001$.

## Study 3: Mediation and moderation analyses combining Outcomes A and B

**Study 3: Mediation analyses Combining Outcomes A and B**

### Exploratory hypotheses

We added measures of surprise, confidence, and task difficulty, and adopted the same operationalizations of confidence and task difficulty as in Study 2. For surprise, we measured participants' surprise about the outcome (success or failure). Like in Study 2, we tested both the mediating and the moderating effects of these variables.

*(H10a) Surprise mediates the relationship between the hindsight condition and probability estimates. (exploratory)*

*(H10b) Surprise moderates the relationship between hindsight condition and probability estimates. (exploratory)*

*(H11a) Confidence mediates the relationship between the hindsight condition and probability estimates. (exploratory)*

*(H11b) Confidence moderates the relationship between hindsight condition and probability estimates. (exploratory)*

*(H12a) Task difficulty mediates the relationship between the hindsight condition and probability estimates. (exploratory)*

*(H12b) Task difficulty moderates the relationship between hindsight condition and probability estimates. (exploratory)*

### Mediation analyses

We conducted two sets of mediation analyses: one set for the comparison between Hindsight Outcome Success condition and Foresight condition, the other for the comparison

between Hindsight Outcome Fail condition and Foresight condition. The dependent variable was the probability of the informed outcome (i.e., successful replication for Hindsight Outcome Success condition, failed replication for Hindsight Outcome Fail condition). We used the PROCESS macro for SPSS (Hayes, 2017) to test the mediation model, and tested all four possible mediators simultaneously.

For the comparison between Hindsight Outcome Success condition and Foresight condition, we found support for the mediating effect of surprise about the other outcome (*B* = 1.18, bootstrapped *S.E.* = .57, *p* = .04, 95% CI [0.25, 2.40]). Participants in the Hindsight Outcome Success condition were more likely to perceive the other outcome as more surprising, which was in turn associated with increased probability estimates of Outcome Success. The effect of Hindsight Outcome Success condition on the probability estimates of Outcome Success remained significant even after we controlling for the mediators (*B* = 5.65, *S.E.* = 1.63, *p* = .00, 95% CI [2.43, 8.86]), suggesting that the mediating effect of surprise about the other outcome is partial. There was no evidence for the mediating effects of surprise about the informed outcome (*B* = .32, *boot-strapped S.E.* = .76, *p* = .67, 95% CI [-1.18, 1.82]), confidence (*B* = .59, *boot-strapped S.E.* = .47, *p* = .21, 95% CI [-0.28, 1.58]), or task difficulty (*B* = .02, *boot-strapped S.E.* = .14, p = .87, 95% CI [-0.35, 0.26]).

For the comparison between Hindsight Outcome Fail condition and Foresight condition, we found no support for any of the mediating effects (surprise about Outcome A: *B* = 1.30, *boot-strapped S.E.* = .93, *p* = .16, 95% CI [-0.42, 3.18]; surprise about Outcome B: *B* = 1.05, *boot-strapped S.E.* = .81, *p* = .20, 95% CI [-0.50, 2.75]; confidence: *B* = .88, *boot-strapped S.E.* = .53, *p* = .10, 95% CI [0.09, 2.14]; task difficulty: *B* = -.14, *boot-strapped S.E.* = .20, *p* = .49, 95% CI

[-0.63, 0.16]). Overall, the results provide some support for H10(a), and no support for H11(a) or H12(a).

Table S36

*Study 3: Mediation Analyses*

| **Outcome A (*n* = 332)** | | | | | **Outcome B (*n* = 342)** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **DV: Surprise A** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **DV: Surprise A** | ***B*** | ***S.E.*** | ***t*** | ***p*** |
| Constant | 2.22 | .10 | 21.85 | .000 | Constant | 2.22 | .10 | 21.74 | .000 |
| Hindsight A (vs. Foresight) | -.06 | .14 | -.42 | .677 | Hindsight B (vs. Foresight) | .20 | .14 | 1.45 | .149 |
| **DV: Surprise B** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **DV: Surprise B** | ***B*** | ***S.E.*** | ***t*** | ***p*** |
| Constant | 3.06 | .09 | 33.63 | .000 | Constant | 3.06 | .09 | 33.35 | .000 |
| Hindsight A (vs. Foresight) | .32 | .12 | 2.56 | .011 | Hindsight B (vs. Foresight) | -.16 | .12 | -1.33 | .184 |
| **DV: Confidence** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **DV: Confidence** | ***B*** | ***S.E.*** | ***t*** | ***p*** |
| Constant | 3.99 | .10 | 38.28 | .000 | Constant | 3.99 | .11 | 36.87 | .000 |
| Hindsight A (vs. Foresight) | .19 | .14 | 1.31 | .192 | Hindsight B (vs. Foresight) | -.35 | .15 | -2.40 | .017 |
| **DV: Task difficulty** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **DV: Task difficulty** | ***B*** | ***S.E.*** | ***t*** | ***p*** |
| Constant | 3.98 | .14 | 29.07 | .000 | Constant | 3.98 | .13 | 30.54 | .000 |
| Hindsight A (vs. Foresight) | -.09 | .19 | -.50 | .620 | Hindsight B (vs. Foresight) | .21 | .18 | 1.17 | .243 |
| **DV: Probability – Outcome A** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **DV: Probability – Outcome B** | ***B*** | ***S.E.*** | ***t*** | ***p*** |
| Constant | 52.53 | 4.49 | 11.71 | .000 | Constant | 52.26 | 4.72 | 11.07 | .000 |
| Hindsight A (vs. Foresight) | 5.65 | 1.63 | 3.45 | .001 | Hindsight B (vs. Foresight) | 10.05 | 1.84 | 5.47 | .000 |
| Surprise A | -5.48 | .72 | -7.62 | .000 | Surprise A | 6.53 | .77 | 8.53 | .000 |
| Surprise B | 3.71 | .78 | 4.78 | .000 | Surprise B | -6.36 | .84 | -7.60 | .000 |
| Confidence | 3.17 | .69 | 4.57 | .000 | Confidence | -2.51 | .73 | -3.43 | .001 |
| Task difficulty | .25 | .56 | .45 | .656 | Task difficulty | -.66 | .61 | -1.09 | .279 |
| **Indirect effects** | ***B*** | ***Boot S.E.*** | ***t*** | ***p*** | **Indirect effects** | ***B*** | ***Boot S.E.*** | ***t*** | ***p*** |
| Surprise A | .32 | .76 | .42 | .675 | Surprise A | 1.30 | .93 | 1.41 | .160 |
| Surprise B | 1.18 | .57 | 2.06 | .040 | Surprise B | 1.05 | .81 | 1.29 | .198 |
| Confidence | .59 | .47 | 1.26 | .209 | Confidence | .88 | .53 | 1.66 | .098 |
| Task Difficulty | -.02 | .14 | -.17 | .865 | Task Difficulty | -.14 | .20 | -.69 | .491 |
| **Total effects (direct + indirect)** | ***B*** | ***Boot S.E.*** | ***t*** | ***p*** | **Total effects (direct + indirect)** | ***B*** | ***Boot S.E.*** | ***t*** | ***p*** |
| | 2.06 | 1.14 | 1.81 | .071 | | 3.09 | 1.40 | 2.22 | .027 |

*Note.* The left panel shows results of mediation for the comparison between Hindsight Outcome Success condition and Foresight condition on the probability estimates of Outcome A. The right panel shows results of mediation for the comparison between Hindsight Outcome Fail condition and Foresight condition on the probability estimates of Outcome B. A positive relationship between hindsight condition and probability estimates indicates hindsight bias.

## Moderation analyses

To test Hypotheses 10(b), 11(b), and 11(b), we mean-centered hindsight condition, surprise about Outcome Success, surprised about Outcome Fail, confidence, and task difficulty, and tested one moderator at a time to reduce concerns about potential multicollinearity. As shown in Table S37, we found a significant interaction effect between Hindsight Outcome Success condition and surprise about Outcome Fail on the probability estimate of Outcome A (*B* = -3.26, *S.E.* = 1.60, *p* = .04, 95% CI [-6.41, -0.11]). As shown in Figure S16, the relationship between Hindsight Outcome Success condition and the probability estimate of Outcome Success was positive and significant when surprise about Outcome Fail was low (simple slope analysis: *B* = 9.47, *S.E.* = 2.55, *p* = .000), and it became nonsignificant when surprise about Outcome Fail was high (simple slope analysis: *B* = 2.05, *S.E.* = 2.60, *p* = .43). The correlation between hindsight condition and confidence was .14 (*p* =.01), which was a small to medium effect, alleviating concerns about multicollinearity in this interaction effect. We did not find support for all other hypothesized moderating effects. Overall, the results provided some support for H10(b), but no support for the H11(b) or H12(b).

Table S37

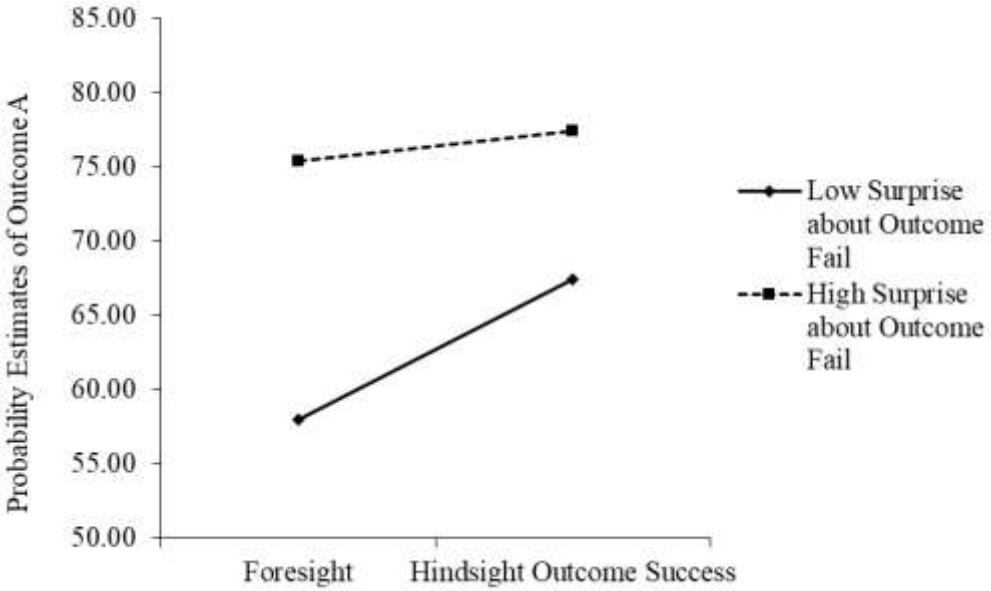
*Study 3: Moderation Analyses*

| **Outcome A** | | | | | **Outcome B** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **DV: Probability – Outcome A** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **DV: Probability – Outcome B** | ***B*** | ***S.E.*** | ***t*** | ***p*** |
| Constant | 69.53 | .88 | 79.06 | .000 | Constant | 41.87 | 1.02 | 41.13 | .000 |
| Hindsight A (vs. Foresight) | 7.35 | 1.76 | 4.17 | .000 | Hindsight B (vs. Foresight) | 11.75 | 2.05 | 5.74 | .000 |
| Surprise A | -6.03 | .70 | -8.62 | .000 | Surprise A | 6.97 | .80 | 8.68 | .000 |
| Hindsight A x Surprise A | 2.01 | 1.40 | 1.43 | .153 | Hindsight B x Surprise A | -.24 | 1.61 | -.15 | .883 |
| **DV: Probability – Outcome A** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **DV: Probability – Outcome B** | ***B*** | ***S.E.*** | ***t*** | ***p*** |
| Constant | 69.75 | .91 | 76.76 | .000 | Constant | 41.83 | 1.01 | 41.52 | .000 |
| Hindsight A (vs. Foresight) | 5.76 | 1.82 | 3.16 | .002 | Hindsight B (vs. Foresight) | 11.82 | 2.02 | 5.84 | .000 |
| Surprise B | 5.89 | .80 | 7.37 | .000 | Surprise B | -8.10 | .89 | -9.15 | .000 |
| Hindsight A x Surprise B | -3.26 | 1.60 | -2.03 | .043 | Hindsight B x Surprise B | -.84 | 1.78 | -.47 | .636 |
| **DV: Probability – Outcome A** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **DV: Probability – Outcome B** | ***B*** | ***S.E.*** | ***t*** | ***p*** |
| Constant | 69.53 | .93 | 74.80 | .000 | Constant | 42.06 | 1.10 | 38.18 | .000 |
| Hindsight A (vs. Foresight) | 6.90 | 1.86 | 3.70 | .000 | Hindsight B (vs. Foresight) | 11.85 | 2.22 | 5.35 | .000 |
| Confidence | 4.31 | .72 | 6.00 | .000 | Confidence | -3.47 | .82 | -4.25 | .000 |
| Hindsight A x Confidence | -.77 | 1.44 | -.53 | .594 | Hindsight B x Confidence | 2.27 | 1.66 | 1.37 | .172 |
| **DV: Probability – Outcome A** | ***B*** | ***S.E.*** | ***t*** | ***p*** | **DV: Probability – Outcome B** | ***B*** | ***S.E.*** | ***t*** | ***p*** |
| Constant | 69.49 | .94 | 74.06 | .000 | Constant | 41.96 | 1.11 | 37.75 | .000 |
| Hindsight A (vs. Foresight) | 7.44 | 1.88 | 3.96 | .000 | Hindsight B (vs. Foresight) | 12.77 | 2.23 | 5.71 | .000 |
| Task Difficulty | -2.91 | .55 | -5.25 | .000 | Task Difficulty | 1.63 | .69 | 2.36 | .019 |
| Hindsight A x Task Difficulty | -.35 | 1.11 | -.32 | .752 | Hindsight B x Task Difficulty | -1.99 | 1.38 | -1.45 | .149 |

*Note.* The left panel shows results of moderation for the comparison between Hindsight Outcome Success condition and Foresight condition on the probability estimates of Outcome A. The right panel shows results of moderation for the comparison between Hindsight Outcome Fail condition and Foresight condition on the probability estimates of Outcome B. A positive relationship between hindsight condition and probability estimates indicates hindsight bias.

Figure S16

*Study 3: Interaction Between Hindsight Outcome Success (vs. Foresight) Condition and Surprise about Outcome B on Probability Estimates of Outcome A.*




## Study 3: Age-related analyses

We examined whether hindsight bias is contingent on age by two sets of analyses. In the first set of analyses, we conducted independent-samples *t* tests in smaller samples consisting of younger participants and older participants, respectively. In the second set of analyses, we tested whether age moderates the relationship between experimental condition and outcomes such as probability estimates and surprise ratings.

As in Studies 1 and 2, we chose 18-31 years old as the age range of younger adults ($n$= 171), and 50 years old and above as the age range of older adults ($n$ = 106). Table S38 showed the means and standard deviations of probability estimates, surprise, confidence, and task difficulty of younger adults. As shown in Table S39, participants in the hindsight (outcome success) condition estimated the probability of a successful replication to be higher than that in the foresight condition, $t$ = 3.17, $df$ = 108, $p$ = .002, $d$ = 0.60. Also, participants in the hindsight

(outcome fail) condition estimated the probability of a successful replication to be lower than that in the foresight condition, $t$ = -2.68, $df$ = 111, $p$ = .008, $d$ = -0.51. These findings provided strong support for Hypothesis 9 among younger participants. We did not find any significant difference between participants in the hindsight conditions and foresight condition on all other dependent variables (i.e., surprise about successful replication, surprise about failed replication, confidence, task difficulty) among younger participants.

Table S40 showed the means and standard deviations of probability estimates, surprise, confidence, and task difficulty of older adults. As shown in Table S41, probability estimates of a successful replication were not significantly different among participants in the hindsight (outcome success) condition and those in the foresight condition, $t$ = 1.60, $df$ = 66, $p$ = .114, $d$ = 0.39, power (1 – $\beta$ err prob) = .35. Also, probability estimates of a successful replication were marginally significantly lower among participants in the hindsight (outcome fail) condition than those in the foresight condition, $t$ = -1.87, $df$ = 68, $p$ = .066, $d$ = -0.45, power (1-$\beta$ err prob) = 0.46. These findings provided little support for Hypothesis 9 among older participants. However, because the statistical power of these two t-tests were smaller than 0.50, we also need to be cautious when interpreting the findings.

Also, as shown in Table S41, for the older adults subsample, confidence ratings were significantly lower among participants in the hindsight (outcome fail) condition than those in the foresight condition, $t$ = -2.50, $df$ = 68, $p$ = .015, $d$ = -0.60. We did not find any other significant differences between participants in the hindsight conditions and foresight condition on all other dependent variables (i.e., surprise about successful replication, surprise about failed replication, confidence, task difficulty) among older participants.

Table S38

*Study 3: Mean Estimations of Outcomes of a Replication of Fischhoff (1975) (in percentage %) (Younger Participants Aged Between 18 and 31 Years Old)*

| | **Foresight (n = 52)** | | **Hindsight Outcome Success: Successful Replication (*n* = 58)** | | **Hindsight Outcome Fail: Failed Replication (*n* = 61)** | |
|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| Estimated probabilities | | | | | | |
| g. Successful replication | 61.56 | 15.60 | 71.91 | 18.38 | 51.10 | 24.11 |
| h. Failed replication | 38.44 | 15.60 | 28.09 | 18.38 | 48.90 | 24.11 |
| Surprise | | | | | | |
| a. Successful replication | 2.63 | 1.31 | 2.21 | 1.25 | 2.51 | 1.22 |
| b. Failed replication | 3.12 | 1.02 | 3.26 | 1.18 | 3.08 | 1.02 |
| Confidence | 3.88 | 1.10 | 4.19 | 1.23 | 3.57 | 1.44 |
| Task difficulty | 3.98 | 1.71 | 4.09 | 1.61 | 4.18 | 1.59 |

Table S39

*Study 3: Independent Samples Student's T-Tests of Estimations of Outcomes of a Replication of Fischhoff (1975) (Younger Participants Aged Between 18 and 31 Years Old)*

| **Hindsight vs. Foresight** | ***Mean Difference*** | ***t*** | ***df*** | ***p*** | **Cohen's *d*** |
|---|---|---|---|---|---|
| **Estimated probabilities of successful replication** | | | | | |
| Hindsight Outcome Success vs. Foresight | 10.36 | 3.17 | 108 | .002 | 0.60 |
| Hindsight Outcome Fail vs. Foresight a | -10.46 | -2.68 | 111 | .008 | -0.51 |
| **Surprise about successful replication** | | | | | |
| Hindsight Outcome Success vs. Foresight | -0.43 | -1.75 | 108 | .084 | -0.33 |
| Hindsight Outcome Fail vs. Foresight a | -0.13 | -0.53 | 111 | .597 | -0.10 |
| **Surprise about failed replication** | | | | | |
| Hindsight Outcome Success vs. Foresight | 0.14 | 0.68 | 108 | .500 | 0.13 |
| Hindsight Outcome Fail vs. Foresight | -0.03 | -0.17 | 111 | .863 | -0.03 |
| **Confidence** | | | | | |
| Hindsight Outcome Success vs. Foresight | 0.31 | 1.36 | 108 | .176 | 0.26 |
| Hindsight Outcome Fail vs. Foresight a | -0.31 | -1.27 | 111 | .206 | -0.24 |
| **Task difficulty** | | | | | |
| Hindsight Outcome Success vs. Foresight | 0.11 | 0.33 | 108 | .740 | 0.06 |
| Hindsight Outcome Fail vs. Foresight | 0.20 | 0.64 | 111 | .521 | 0.12 |

*Note.* a. Levene's test was significant.

Table S40

*Study 3: Mean Estimations of Outcomes of a Replication of Fischhoff (1975) (in percentage %) (Older Participants >= 50 Years Old)*

| | **Foresight (n = 32)** | | **Hindsight Outcome Success: Successful Replication (*n* = 36)** | | **Hindsight Outcome Fail: Failed Replication (*n* = 38)** | |
|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| Estimated probabilities | | | | | | |
| i. Successful replication | 66.84 | 22.01 | 74.81 | 18.96 | 58.18 | 16.79 |
| j. Failed replication | 33.16 | 22.01 | 25.19 | 18.96 | 41.82 | 16.79 |
| Surprise | | | | | | |
| c. Successful replication | 1.97 | 1.12 | 1.83 | 1.16 | 2.08 | 1.19 |
| d. Failed replication | 3.06 | 1.27 | 3.39 | 1.15 | 3.00 | 1.12 |
| Confidence | 4.22 | 1.16 | 4.47 | 1.21 | 3.45 | 1.39 |
| Task difficulty | 3.72 | 1.71 | 3.44 | 1.89 | 4.45 | 1.66 |

Table S41

*Study 3: Independent Samples Student's T-Tests of Estimations of Outcomes of a Replication of Fischhoff (1975) (Older Participants >= 50 Years Old)*

| **Hindsight vs. Foresight** | ***Mean Difference*** | ***t*** | ***df*** | ***p*** | **Cohen's *d*** |
|---|---|---|---|---|---|
| **Estimated probabilities of successful replication** | | | | | |
| Hindsight Outcome Success vs. Foresight | 7.96 | 1.60 | 66 | .114 | 0.39 |
| Hindsight Outcome Fail vs. Foresight | -8.66 | -1.87 | 68 | .066 | -0.45 |
| **Surprise about successful replication** | | | | | |
| Hindsight Outcome Success vs. Foresight | -0.14 | -0.49 | 66 | .627 | -0.12 |
| Hindsight Outcome Fail vs. Foresight | 0.11 | 0.40 | 68 | .694 | 0.09 |
| **Surprise about failed replication** | | | | | |
| Hindsight Outcome Success vs. Foresight | 0.33 | 1.11 | 66 | .270 | 0.27 |
| Hindsight Outcome Fail vs. Foresight | -0.06 | -0.22 | 68 | .827 | -0.05 |
| **Confidence** | | | | | |
| Hindsight Outcome Success vs. Foresight | 0.25 | 0.88 | 66 | .381 | 0.21 |
| Hindsight Outcome Fail vs. Foresight | -0.77 | -2.50 | 68 | .015 | -0.60 |
| **Task difficulty** | | | | | |
| Hindsight Outcome Success vs. Foresight | -0.27 | -0.63 | 66 | .534 | -0.15 |
| Hindsight Outcome Fail vs. Foresight | 0.73 | 1.81 | 68 | .075 | 0.43 |

*Note.* Levene's test was nonsignificant for all comparisons.

Table S42 showed a series of moderation analyses using experimental condition as the independent variable, probability estimates, surprise, confidence, and task difficulty as dependent variables, and age as the moderator. None of the moderation analyses were significant (even before adjusting the *p* values for multiple testing, $n$ = 332~342 for each moderation analysis). The findings therefore suggest that the hindsight bias (based on probability estimates) and the findings related to the other dependent variables in Study 3 were not contingent on participants' age.

Table S42

*Study 3: Age as Moderator*

| | | Hindsight Outcome Success vs. Foresight ($n$ = 332) | | | Hindsight Outcome Fail vs. Foresight ($n$ = 342) | | |
|---|---|---|---|---|---|---|---|
| | | *B* | *SD* | *p* | *B* | *SD* | *p* |
| Estimated probabilities of successful replication | Constant | 65.40 | 1.43 | .000 | 65.40 | 1.66 | .000 |
| | Condition | 7.64 | 1.95 | .000 | -13.11 | 2.24 | .000 |
| | Age | 0.22 | 0.12 | .069 | 0.22 | 0.14 | .118 |
| | Condition * Age | -0.17 | 0.16 | .290 | -0.05 | 0.19 | .784 |
| Surprise about successful replication | Constant | 2.22 | 0.10 | .000 | 2.22 | 0.10 | .000 |
| | Condition | -0.05 | 0.14 | .709 | 0.20 | 0.14 | .153 |
| | Age | -0.02 | 0.01 | .050 | -0.02 | 0.01 | .049 |
| | Condition * Age | 0.01 | 0.01 | .295 | 0.00 | 0.01 | .901 |
| Surprise about failed replication | Constant | 3.06 | 0.09 | .000 | 3.06 | 0.09 | .000 |
| | Condition | 0.32 | 0.13 | .012 | -0.17 | 0.12 | .181 |
| | Age | 0.00 | 0.01 | .676 | 0.00 | 0.01 | .679 |
| | Condition * Age | 0.00 | 0.01 | .955 | -0.01 | 0.01 | .627 |
| Confidence | Constant | 4.00 | 0.10 | .000 | 4.00 | 0.11 | .000 |
| | Condition | 0.18 | 0.14 | .212 | -0.36 | 0.15 | .016 |
| | Age | 0.02 | 0.01 | .082 | 0.02 | 0.01 | .095 |
| | Condition * Age | -0.01 | 0.01 | .617 | -0.02 | 0.01 | .110 |
| Task difficulty | Constant | 3.98 | 0.14 | .000 | 3.98 | 0.13 | .000 |
| | Condition | -0.09 | 0.19 | .640 | 0.21 | 0.18 | .241 |
| | Age | -0.01 | 0.01 | .577 | -0.01 | 0.01 | .558 |
| | Condition * Age | 0.00 | 0.02 | .987 | 0.01 | 0.02 | .633 |

# Discussion regarding impact of demographic variables

### Age differences

Because our samples have a wider age range than those used in the original studies, we also examined whether age played a role in our findings. Past findings about age differences of hindsight bias are inconclusive. Some research suggests that hindsight bias is stronger among children and older adults than among younger adults, because children and older adults are more susceptible to accessibility bias (i.e., encoding irrelevant information presented after the original information) and/or inhibitory deficit (i.e., incapability to suppress the retrieval of interfering information presented after the original information) (Bayen et al., 2007; Bernstein et al., 2011). However, other studies found no age difference between younger and older adults in hindsight bias when the confounding impact of recall ability was removed (Groß & Bayen, 2015) or when the new information was presented in a weak situation (Pohl et al., 2018). We conducted moderation analyses to examine whether age moderated the relationship between experimental condition and outcomes such as probability estimates, surprise, confidence, and task difficulty. None of the moderation analyses were significant in all three studies. These findings therefore suggest that our findings were not contingent on participants' age.

### Cross-cultural differences

Our three studies relied on participants based in the United States. In comparison, the original studies of Studies 1 and 2 recruited participants from Israel.

There is an ongoing debate about whether hindsight bias holds or varies across culture, which has not reached a firm conclusion (Choi & Nisbett, 2000; Heine & Lehman, 1996; Ma-Kellams, 2020; Pohl et al., 2002). Heine and Lehman (1996) compared hindsight bias across Canadian and Japanese cultures. They found no difference between Canadians and Japanese using the memory design, and a marginal difference between Canadians and Japanese, such that

Canadians exhibited greater hindsight bias than Japanese. Choi and Nisbett (2000) conducted another test of hindsight bias across cultures using the hypothetical design. They found that Koreans exhibited greater hindsight bias than Americans, a pattern that is opposite to that found in Heine and Lehman (1996). Pohl et al. (2002) examined hindsight bias using the hypothetical design in a sample containing participants all over the world. They found large and stable hindsight bias among Asian, Australian, and North American participants, and there was no significant difference among these groups. European participants exhibited smaller hindsight bias than participants from the other three continents, yet this difference disappeared after removing participants from Germany and the Netherlands due to their familiarity with the study materials. We cannot directly test cross-cultural differences using our samples collected in the United States, yet we tried to evaluate the impact of cross-cultural differences using indirect means.

First, based on our literature review, Davis and Fischhoff's (2014) replication study of Slovic and Fischhoff (1977) used a U.S. sample, and they found support for hindsight bias in that sample. Second, we examined whether participants made better estimations on cultural-specific questions than would be expected by mere chance. Specifically, we focused on Event B (near riot in Atlanta in 1967) in Study 1, as it occurred within the United States. If our American participants did have knowledge about the near riot in Atlanta, then we can expect that the percentage of participants who predicted the correct historical outcome of Event B would be higher than chance. However, this was not what we found in the data. The mean probability estimates for the correct historic outcome (i.e., dispersion and no outbreak of violence) was the lowest among those for all four outcomes, and it was lower than chance (one-sample t-test: $t$ = -12.87, $df$ = 45, $p$ = .000, $d$ = -1.90).

## Discussion regarding use of Events C and D in Fischhoff (1975)

We noted in the general discussion the challenge regarding the use of Events C and D from Fischhoff (1975) and concluded that "In correspondence with the original author and the editor we felt it needed to include a warning note that that these stimuli should no longer be used in follow-up research. We removed the reporting of these materials and analyses of these events from the manuscript and the supplementary. ".

We note that the stimuli is still included in the frozen pre-registration and the data from these events are still provided on the OSF. We further note that not reporting these in our manuscript is a deviation from the pre-registration, done in consultation with the editor and in correspondence with the original author. We feel this deviation is warranted given the circumstances.

## References

Arkes, H. R., & Gaissmaier, W. (2012). Psychological research and the PSA test controversy. *Psychological Science, 23,* 547–553.

Bayen, U. J., Erdfelder, E., Bearden, J. N., & Lozito, J. P. (2006). The interplay of memory and judgment processes in effects of aging on hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1003–1018.

Bayen, U. J., Pohl, R. F., Erdfelder, E., & Auer, T. S. (2007). Hindsight bias across the life span. *Social Cognition*, *25*, 83-97.

Bradfield, A., & Wells, G. L. (2005). Not the same old hindsight bias: Outcome information distorts a broad range of retrospective judgments. *Memory & Cognition*, *33*, 120-130.

Choi, I., & Nisbett, R. E. (2000). Cultural psychology of surprise: Holistic theories and recognition of contradiction. *Journal of Personality and Social Psychology, 79*, 890–905.

Connor-Smith, J. K., & Compas, B. E. (2002). Vulnerability to social stress: Coping as a mediator or moderator of sociotropy and symptoms of anxiety and depression. *Cognitive Therapy and Research, 26,* 39-55.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Fay, M. P. (2017). R Package asht [computer software]. Retrieved from https://CRAN.R-project.org/package=asht

Fay, M. P., & Malinovsky, Y. (2018). Confidence intervals of the Mann-Whitney parameter that are compatible with the Wilcoxon-Mann-Whitney test. *Statistics in Medicine*, *37*, 3991-4006.

Hamilton, W. K., Aydin, B., & Mizumoto, A. (2016). MAVIS [online software]. Retrieved from http://kylehamilton.net/shiny/MAVIS

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. New York, NY: Routledge.

Heine, S. J., & Lehman, D. R. (1996). Hindsight bias: A cross-cultural analysis. *Japanese Journal of Experimental Social Psychology*, *35*, 317-323.

James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, *69*, 307–321.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, *6*, 115–134.

Karazsia, B. T., & Berlin, K. S. (2018). Can a mediator moderate? Considering the role of time and change in the mediator-moderator distinction. *Behavior Therapy*, *49*, 12-20.

Ma-Kellams, C. (2020). Cultural Variation and Similarities in Cognitive Thinking Styles Versus Judgment Biases: A Review of Environmental Factors and Evolutionary Forces. *Review of General Psychology*.

Muthén, L., & Muthén, B. (1998–2015). Mplus user's guide. Los Angeles, CA: Muthén & Muthén.

Newcombe, R. G. (2006). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 2: asymptotic methods and evaluation. *Statistics in Medicine, 25,* 559-573.

Pohl, R. F., Bayen, U. J., Arnold, N., Auer, T. S., & Martin, C. (2018). Age differences in processes underlying hindsight bias: A life-span study. *Journal of Cognition and Development*, *1*, 278-300.

Revelle W (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.9.12, Retrieved from https://CRAN.R-project.org/package=psych

Wei, M., Mallinckrodt, B., Russell, D. W., & Abraham, W. T. (2004). Maladaptive Perfectionism as a Mediator and Moderator Between Adult Attachment and Depressive Mood. *Journal of Counseling Psychology*, *51*, 201–212.

Zhou, L., Wang, M., Chen, G., & Shi, J. (2012). Supervisors' upward exchange relationships and subordinate outcomes: Testing the multilevel mediation role of empowerment. *Journal of Applied Psychology*, *97*, 668–680.