Title

Boundaries of the Hindsight Bias

Permalink

https://escholarship.org/uc/item/3q75w2p7

Author

Schatz, Derek

Publication Date

2019

Peer reviewed|Thesis/dissertation

Boundaries of the Hindsight Bias

By

Derek A. Schatz

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Don A. Moore
Professor Leif D. Nelson
Professor Clayton Critcher
Professor Stefano DellaVigna

Spring 2019

Boundaries of the Hindsight Bias

ABSTRACT

Boundaries of the Hindsight Bias

by

Derek A. Schatz

Doctor of Philosophy in

Business Administration

University of California, Berkeley

Professor Don A. Moore, Chair

The hindsight bias may not be as robust as previously believed. Also known as the "knew-it-all-along effect" (Fischhoff, 1975) the hindsight bias refers to the inability for individuals to remember their previous state of knowledge after learning an outcome. Researchers have found people fall victim to hindsight biases in a variety of domains, including general knowledge (Fischoff, 1977; Wood, 1978), medical decisions (Arkes, Wortmann, Saville, & Harkness, 1981), and political outcomes (Fischhoff & Beyth, 1975). Multiple failures to debias, with methods including informing participants about the bias (Fischhoff, 1977) and manipulating participant perspective (Wood, 1978) further grew the reputation for the strength of this bias. However, we must interpret older results in the behavioral sciences with new perspective on the importance of sample size and statistical power (Ioannidis, 2005). This dissertation proposes to investigate when and why hindsight bias fails to appear, and what it means for the psychological processes underlying the hindsight bias, and more broadly, the interpretation of older psychological research.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Thank you to my dissertation committee Don Moore, Leif Nelson, Clayton Critcher, and Stefano DellaVigna. I am humbled that you all generously agreed to lend your time and expertise on behalf of this work, and it is that much stronger of a dissertation due to your collective wisdom and mentorship.

Don, to put it simply, you have changed how I see the world in the four wonderful years during which I had the pleasure to learn from you. Thanks to your tutelage I now see objects and events in the world as points in a distribution, and I am ever-vigilant in monitoring myself to prevent the presence of any bias in my decision making – of course watching out for the most pernicious of all; overconfidence. I could not have wished for a more thoughtful, supportive, and conscientious advisor, and I am proud to say that outside of research you taught me useful skills in communication and collaboration which will serve me just as well in the years to come.

Leif, thank you for your refreshing perspective on research and human behavior, and your strong sense of practicality when making social inferences. Your leadership and guidance helped me hone my ability to critique the published record, and this ability to critically digest research will benefit me for the rest of my life. Clayton, thank you for your patience and your creativity, and for showing me how to expertly analyze the assumptions and mechanisms inherent in studying human behavior. You have always challenged me to think one level deeper and I am all the smarter for it. Stefano, thank you for your perspective and your expertise. Your contributions and feedback on this work has been invaluable. You taught me how to truly think like an economist, and I am forever grateful for the opportunity to learn from you.

Thanks to the faculty at Haas and at Berkeley at large, especially the entire Management of Organizations faculty. I appreciate your generosity of time and advice. Thanks especially to Ellen Evers in Marketing and Serena Chen in Social Psychology. Ellen, you have always expanded my perspective with thought-provoking questions, and you have helped me to critically examine my assumptions. Serena, thank you for your generous time and patience in supervising and advising my undergraduate psychology honors thesis. Frankly without your early support, I would not be writing this today.

Kari Thurman, none of this would be possible without your unwavering support over the years. Any accomplishments of mine are owed also in part to you, as you believed in me and had faith in me at times when even I did not. Kari, you are my rock and I cannot express in words how grateful I am to have you along with me for this journey.

Thanks to my parents, Allen and Kathryn and my sister, Allison. You have all believed in me since the literal Day One, and I am blessed and forever thankful to have such a supportive family that has always pushed me to pursue my goals and dreams. You have always been my loudest and most enthusiastic supporters, and I wouldn’t want it any other way.

PREFACE

My dissertation is organized as follows. In Chapter 1, I review the existing hindsight bias literature, discussing both early findings of the bias as well as later applications and extensions of the research across domains of decision making. I place this early research in the recent context of the open science movement and in this new age of methodological rigor and concern for the replicability of scientific results. This work shows a paucity of supporting evidence for the hindsight bias across multiple types of stimuli and paradigms.

Across ten experiments, I tested the degree to which people report inflated feelings of confidence due to the hindsight bias. These experiments are organized into four categories. First, I employed the original between-subjects outcome feedback paradigm with a variety of stimuli. Across Experiments 1- 4 (Chapters 2 – 5) participants failed to exhibit the hindsight bias. Seeing the outcome of an event or question did not appear to increase reported confidence in the occurrence of that particular outcome. Following these initial four experiments, I began to search for possible moderators that could explain the repeated failures to replicate the basic effect.

The second category of experiments used an alternate, longitudinal paradigm to test the hindsight bias. Experiment 5 (Chapter 6) tested whether a distractor task could elicit the bias between trials within the same person and found the manipulation failed to create the hindsight bias. Experiment 6 (Chapter 7) examined the hindsight bias over a span of months comparing predictions of the 2018 U.S. Midterm Elections reported prior to the elections to those reported ‘in hindsight’ after the elections. Participants responding after the election results were known did not claim any greater confidence in their past predictions compared to participants asked before the election had occurred. These two experiments explored the possible moderators of surprisingness and expertise. However, as there was no hindsight bias present, there was no effect to attenuate with any moderation.

The third category of experiments explored a possible moderator; that of perceived randomness (Wasserman, Lempert, & Hastie, 1991), in which events seen as randomly determined should elicit less hindsight bias. Experiment 7 (Chapter 8) measured the degree to which participants considered the stimuli as random to test if that perception would moderate hindsight bias. Participants did not exhibit the hindsight bias in the outcome feedback (hindsight) condition, therefore there was subsequently no opportunity to measure whether perceived randomness moderates the bias. Experiment 8 (Chapter 9) attempted to manipulate perceptions of randomness with a pretested informational vignette, and measured expertise as a potential moderator. Participants again failed to exhibit the hindsight bias.

In the fourth and final category of experiments, Experiment 9 (Chapter 10) compared the current views of experts in the fields of social psychology and decision-making to my recent empirical findings, in order to contextualize these several failures to replicate the effect. Experts reported believing the hindsight bias is both more widely found as well as stronger than comparative evidence shows. In Experiment 10 (Chapter 11) I administered a direct replication of the original hindsight finding (Fischhoff, 1975) employing the same paradigm and stimuli, while testing for a moderator of question difficulty. Participants in the direct replication condition did exhibit hindsight bias, however the evidence does not implicate question difficulty as a possible moderator.

In Chapter 12, I summarize my research findings regarding the robustness of the hindsight bias. I discuss the implications for this field of research, concerns in utilizing old research in the present day, as well as future directions for this program of research. I close by reconciling the widespread beliefs concerning the ubiquity of the hindsight bias with the evidence at hand.

CHAPTER 1

Theoretical Background and Literature Review

The common saying that “Hindsight is 20/20” reflects the ease with which we make sense of events in hindsight, and also alludes to the fact that people often express greater-than-justifiable confidence when claiming to have known a particular outcome after the outcome is already known. While the hindsight bias is considered one of the most robust and well-documented biases in the study of human behavior (Hawkins & Hastie, 1990, it may not be as robust or as widely applicable as once previously thought. Four preliminary experiments and six additional experiments examine the hindsight bias and seek to identify its limits.

Establishment of Hindsight Bias in the Literature

The hindsight bias, or the “knew-it-all-along” effect (Fischhoff, 1975), is a well-documented bias that impedes individuals from accessing a previous state of ignorance once they are informed of an outcome. Hindsight bias holds a notable place in the bias literature, due not only to its many documented occurrences (as described below), but to its prominent status in colloquial speech. Walster (1967) first studied hindsight bias through a house-buying paradigm. Participants reported how confident they would have been in predicting certain outcomes from the buying process. In this first study of the phenomenon, Walster provided various outcomes of buying a house to different participants and asked for what their confidence would have been retrospectively in anticipating such outcomes. Those that learned of more extreme outcomes such as bankruptcy reported they would have predicted such outcomes with greater confidence than those who learned of more mild outcomes. The bias was named initially by Baruch Fischhoff (1975), who first used obscure historical stimuli along with a paradigm he coined as an outcome feedback design. In this paradigm, participants saw a list of possible outcomes or a true statement of the actual outcome. Participants expressed how confident they would have been had they not known the actual result. It is in this early work that Fischhoff coined the phrase *creeping determinism*, a mechanism of hindsight in which outcome information is automatically processed and thus participants cannot see the result as anything other than inevitable.

Researchers introduced an alternative paradigm early on that examined hindsight longitudinally across time (Fischhoff & Beyth, 1975). In this study they asked participants to predict the likelihood of various outcomes pertaining to President Nixon’s China tour prior to its occurrence. Following the tour, the same participants attempted to recall their initial predictions. In doing do, participants inflated their probabilities of outcomes that actually occurred and likewise decreased their recalled probabilities for outcomes that did not occur. Subsequent research expanded on this paradigm by using topical current events to measure hindsight bias. Evidence for hindsight arose in this period from studying predictions of union strike outcomes (Pennington, 1981) and perceptions of nuclear power plants before and after the Chernobyl disaster (Verplanken & Pieters, 1988).

Early Attempts at Moderating Hindsight Bias

Following these initial investigations, scholars began to look for moderators of the hindsight bias in a variety of decision-making domains. Fischhoff (1977) created a more explicit debiasing paradigm in which participants in the outcome feedback condition read, “respond as you would have had you not been told what the answer was,” as an attempt to attenuate hindsight bias. However, his evidence showed that such a manipulation had no effect in lessening the bias compared to earlier manipulations. Researchers discovered successful methods of attenuation through a variety of means, including providing information that the outcome feedback provided was “wrong” information (Hasher, Attig, & Alba, 1981) as well as asking participants to specifically consider possible alternative outcomes (Slovic & Fischhoff, 1977), and providing participants with their past reasoning in a longitudinal recall paradigm (Davies, 1987).

Apart from the cognitive account proposed by Fischhoff (1975), researchers also explored a motivational account as an alternative cause of hindsight bias. Campbell and Tesser (1983) employed a two-stage between-subjects outcome feedback design over time with the same participants using trivia stimuli. They found that both system-justifying beliefs (i.e. believing in a predictable, certain world) and social desirability motivations were correlated with greater hindsight bias. Monetary incentives for accurate responses were also introduced to attenuate any motivational components of the hindsight bias such as self-esteem or social desirability (Hell, Gigerenzer, Gauggel, Mall, & Müller, 1988). In a complex series of results, they showed that large monetary incentives do attenuate hindsight bias. However when the initial estimate is memorable, incentives have no additional effect.

Applications and Extensions of Hindsight Bias

Following these several well-known findings in the laboratory, researchers began further utilizing the longitudinal paradigm for hindsight in real world settings, in the same vein as the study of Nixon’s China trip in Fischhoff and Beyth (1975). The first papers to follow in this line examined a college football game (Leary, 1981) and the 1980 presidential election (Leary, 1982). Leary found that participants responding after the event in question occurred were more accurate and confident in what they claimed to be their ex-ante event predictions compared to the actual predictions of participants asked prior to the event. Tests for motivational moderators of the bias were conducted with the 1982 Hawaii gubernatorial election, in which Synodinos (1986) measured self-esteem and political involvement as possible attenuating influences on hindsight. His results showed no such attenuation, further bolstering the reputed robustness of the bias. Similar results arose testing hindsight for several elections in 1984 (Powell, 1988).

In search of further generalizability for the bias, researchers began testing for hindsight bias within the domain of professional judgments, such as medical diagnoses. In one experiment, researchers compared diagnoses from doctors who either read a case history with no past diagnosis or one labeled with a history of a specific medical condition. Doctors who read the latter expressed greater confidence that they would have diagnosed that particular condition. This study was the first to provide evidence that the hindsight bias is prevalent even among professionals in their field (Arkes et al., 1981). Neuropsychologists also succumbed to the hindsight bias when considering diagnoses of brain damage (Arkes, Faust, Guilmette, & Hart, 1988). In the field of law, scholars found evidence suggesting that jurors exposed to incriminating evidence for a defendant were more likely to deliver a guilty verdict compared to those who had not seen the evidence. While not surprising by default, this effect held for jurors who were informed after the fact that the evidence was inadmissible (Sue, Smith, & Caldwell, 1973; Thompson, Fong, & Rosenhan, 1981; Werner, Kagehiro, & Strube, 1982).

The early findings and extensions of hindsight bias inspired related work regarding social judgment and cognitive inference. Many of these examples apply some form of 'creeping determinism' with regards to human inability to purposefully forget or ignore information for the sake of impartiality. (Mitchell & Kalb, 1981) found evidence suggesting that outcome knowledge for nurse decisions created supervisor perceptions of greater responsibility following bad outcomes. Outcome information similarly impacted perceptions of decision quality for accounting decisions (Buchman, 1985) as well as monetary gambles (Baron & Hershey, 1988). Related research suggests that outcome information and hindsight can impact impression formation. There are several examples in which participants read trait or behavioral information of a target and are subsequently asked to ignore some of the information. As could be expected in light of the extant literature, this request to ignore information proved difficult for subjects when attempting to form an impression of the target (Ross, Lepper, & Hubbard, 1975; Schul & Burnstein, 1985; Wyer & Unverzagt, 1985).

Proposed Antecedents of the Hindsight Bias

Researchers have theorized two primary categories of inputs which lead to hindsight bias: cognitive inputs and motivational inputs (Roese & Vohs, 2012). Cognitive causes of hindsight relate to bias resulting from the memory process. This is particularly relevant in designs in which participants attempt to recall their initial knowledge or their estimates following outcome feedback. There exists evidence along this line that suggests that more elaborate encoding of initial information can attenuate the bias (Hell et al., 1988). Relatedly, outcome information has the potential to update existing knowledge and strengthen compatible information in one's memory (Blank & Nestler, 2007). In other words, people have the tendency to adjust their memory in light of the novel outcome information to make the novel outcome predictable. Finally, the process of sensemaking also can explain hindsight in some circumstances. This process pertains to the oversimplifying of cause and effect relationship for a particular outcome (Kruglanski, 1989). In this line, outcomes which are preceded by more straightforward causal explanations therefore elicit greater hindsight bias (Trabasso & Bartolone, 2003; Wasserman, Lempert, & Hastie, 1991; Yopchick & Kim, 2012). This process is an alternative wording of Fischhoff's original 'creeping determinism' hypothesis (Fischhoff, 1975).

Motivational inputs to hindsight bias include both the need for closure and for self-esteem. The need for closure leads people to seek meaning in life that creates order and predictability in their world (Jost, Banaji, & Nosek, 2004; King, Hicks, Krull, & Del Gaiso, 2006). There are many examples in the literature that connects this need with hindsight bias, in which those with a greater need for control exhibit more hindsight bias (Campbell & Tesser, 1983; Hirt, Kardes, & Markman, 2004; Musch & Wagner, 2007). This is because hindsight bias leads individuals to view events as more pre-determined and predictable, directly satisfying this need for closure. Self-esteem is related to hindsight bias in regards to good and bad outcomes. That is, individuals could boost their self-esteem by strategically claiming that they would have predicted a good outcome to occur. Conversely, individuals could absolve themselves of blame by claiming prior ignorance in the face of a negative result. This connection between self-esteem in hindsight bias entails is supported by mixed evidence, however. There are several examples in which people exhibit less hindsight after negative outcomes compared to positive ones (for example, (Hölzl, Kirchler, & Rodler, 2002; Louie, Curren, & Harich, 2000; Pezzo & Beckstead, 2008). One should note that there is also evidence of the opposite, in which negative outcomes increase hindsight bias (Tykocinski, 2001; Wann, Grieve, Waddill, & Martin, 2008).

Robustness of the Hindsight Bias

Hindsight has proven stubborn in the face of debiasing efforts. Incentives for ignoring outcome feedback failed to attenuate the bias, while providing mixed evidence for the efficacy of incentives on hindsight overall (Camerer & Loewenstein, 1989) Attempts to debias hindsight have even backfired. Following several attempts to debias hindsight through instructing participants to 'consider the opposite' in terms of how an outcome could have occurred differently, (Arkes et al., 1988; Koriat, Lichtenstein, & Fischhoff, 1980; Slovic & Fischhoff, 1977) – none of which fully eliminated hindsight. Sanna, Schwarz, and Stocker (2002) attempted to create a powerful manipulation in which they asked participants to list many possible alternatives to a given outcome. Participants who did so, however, displayed even greater hindsight bias. In fact, one is hard-pressed to find much critique of the original and robust hindsight bias literature, save for one possibly prescient lament from Hawkins and Hastie (1990): "…virtually all of the experiments explicitly directed at the question of the robustness of hindsight effects have used almanac question materials."

Examining the Evidence for Hindsight Bias

While the research history of hindsight bias seems impressive given the many occasions on which the phenomenon has been replicated in the past, sample sizes abound among much of the foundational hindsight literature. Low sample sizes present a true concern to the validity of statistical effects, as they produce low statistical power which in turn can lead to inflated effect sizes and hard-to-reproduce results (Button et al., 2013; Ioannidis, 2005). Researchers decades ago were not necessarily nefarious in their underpowered studies, as norms of research were simply different then. Indeed, these patterns are not unique to the hindsight bias literature. Armed with more statistical knowledge, however, present-day researchers are equipped with tools and techniques to do better, as well as to re-interpret prior findings.

With this in mind, how does one evaluate the quality of evidence in a body of literature? To start, it is important to remember that the published record does not represent all of the research conducted on a particular topic. The fact that only successfully significant studies are published introduces bias to the results found in journals (Pashler & Harris, 2012; Rosenthal, 1979). Knowing that the individual results found in papers can be inflated, scholars often use the number of published results and replications to evaluate the strength of evidence for a phenomenon. This, too, is not a perfect measure of evidentiary quality.

An important component which contributes to the quality of the published evidence in a field is the set of behaviors the researchers engage in when collecting and analyzing data. Researchers have several options in when and how to collect data (or stop collecting data) in addition to which outliers and covariates to include or exclude. Strategically using these “researcher degrees of freedom” to achieve significant results is known as *p*-hacking and can lead to weak evidence entering the published record (Simmons, Nelson, & Simonsohn, 2011).

With this in mind, one method with which to evaluate the quality of published evidence is with a *P*-curve (Simonsohn, Nelson, & Simmons, 2014). A *P*-curve tests for the quality of evidence within a series of papers by identifying which significant *p* values are more likely and less likely. Looking at a series of p values when an effect is true, there should be more *ps* < .025 than .025 < *p*s < .05, creating a right-skewed distribution (Cumming, 2008; Hung, O’Neill, Bauer, & Kohne, 1997). The degree to which a *p*-curve is right-skewed is caused by the true power of the studies included. These truths therefore conclude that a *p*-curve which shows a *v* shape suggests a series of very underpowered studies, and an entirely left-skewed distribution would suggest a series of spurious, *p*-hacked results, or in other words, very weak evidence.

I created a *P*-curve using all published, peer-reviewed, accessible articles used in both of the two existing hindsight bias meta-analyses (Christensen-Szalanski & Willham, 1991; Guilbault, Bryant, Brockway, & Posavac, 2004). This search created a group of 34 empirical papers studying hindsight bias. Due to the age of the research, 15 of the papers did not include the statistical values required for the *p*-curve calculation algorithm, resulting in a final list of 19 papers (See Exhibit 1). This *P*-curve suggests that while some of the hindsight bias evidence is underpowered, there is little evidence to suggest that the literature is profoundly tainted by *p*-hacking (See Exhibit 2 for a *p*-curve disclosure table).

Overview of Experiments

Following initial failures to replicate the hindsight bias, this dissertation engages in a search for potential moderators and boundary conditions which could explain such null results. The goal of the experiments in this dissertation is to

1. Test if people elicit the hindsight bias for a variety of different stimuli, while measuring numerous potential moderating variables across studies,

2. Test if elicitation of the hindsight bias is specific to the research paradigm used,

3. Explore the proposed mediator of ‘creeping determinism’ and its role in manifesting the bias,

4. Examine the views of current scholars in the field regarding the bias, and

5. Attempt a close replication of an original hindsight bias finding.

I first test the basic hindsight bias effect with a variety of stimuli. Experiments 1 – 4 sought to test whether providing participants with outcome feedback would increase their confidence in having known the correct answer compared with those who did not receive the feedback. The evidence suggests that the outcome feedback consistently failed to elicit the hindsight bias.

Experiments 5 and 6 examined the hindsight bias with an alternative paradigm, using a longitudinal design. Experiment 5 used a within-subjects approach which asked participants to recall initial confidence estimates following a distractor task and the correct outcome information. Experiment 6 compared predictions and confidence levels made prior to the U.S. midterm elections with those reported after election results were known. These experiments also tested the potential moderators of perceived surprisingness and domain expertise. The hindsight appeared weakly in Experiment 5 but bias but did not manifest at all in Experiment 6 with this longitudinal paradigm.

Experiments 7 and 8 explored whether the degree to which people perceive an event as random would lessen their susceptibility to the hindsight bias. Experiment 7 utilized a multiple item measure of perceived randomness to test whether it had a moderating effect on the bias. Participants did not display any hindsight bias to begin with, depriving the moderator of any supportive evidence. Experiment 8 tested an intervention designed to manipulate participant perceptions of randomness. This randomness manipulation failed to exacerbate or attenuate any hindsight bias effect.

Experiment 9 contextualized these failures to replicate the hindsight bias by soliciting the current views of scholars in the field regarding the bias. While present-day researchers believe that hindsight bias is both a powerful bias and one that manifests across many stimulus domains, present evidence collected in parallel provided no support for the hindsight bias. Experiment 10 tested a direct replication using original stimuli alongside the potential moderator of question difficulty. The direct replication condition successfully elicited hindsight bias, however question difficulty did not prove to be a moderator for the effect.

CHAPTER 2

Experiment 1: Debiasing Probability Forecasts

Experiment 1 tests how the hindsight bias affects subsequent forecast accuracy and calibration. Do participants ‘primed’ with hindsight bias make less accurate forecasts? Hindsight bias would suggest that after learning of the outcome of an event, participants would claim excessive confidence and even accuracy, compared to the confidence and accuracy of participants in a control condition. The goal of this experiment was to elicit hindsight bias and measure the impact of the bias on prediction accuracy.

All participants reviewed the same basketball games that had occurred in the prior week. The paradigm for this experiment was borrowed from Kelly and Simmons (2016), as the survey elicited predictions of sports outcomes along with confidence measures. This operationalization is useful for several reasons. Most importantly, the general population is familiar with outcomes and scores of major sports. In addition, the length of time it takes for a forecasted sporting event to occur can be in a matter of days. The outcome feedback manipulation was adapted from Hoch and Loewenstein, 1989.

In this and all other experiments, I report how sample sizes were determined, pre-registered data exclusions, and all conditions. The link to experimental materials, data, analyses, and pre-registrations for all ten studies in this dissertation can be found at https://osf.io/zc3kn/.

Method

Participants

This study recruited 200 participants (70% male, M_{age} = 34.70) from Amazon Mechanical Turk and paid $1.00 to each with a chance for a $20 bonus. The survey instructed participants that well-calibrated responses would increase their chance of winning the $20 bonus. The survey randomly assigned participants to two between-subjects conditions. I decided to collect 200 participants, 100 for each condition. I did not know what effect size to expect, so I chose this number in hope that it would be an adequate sample size.

Design

The experiment had a 2-cell (outcome feedback vs. control) between-subjects design that manipulated whether or not participants provided post-hoc answers and confidence levels during the review portion of the study. Participants first reviewed 20 basketball games and then made forecasts for 20 games yet to occur in the upcoming week. The main dependent variable was forecast confidence, and hit rates were also measured.

Procedure and Materials

Overview. Participants reviewed 20 recent basketball games one at a time. The content of their review was determined by which condition the survey assigned them, discussed below. Following the review portion, participants then submitted forecasts for 20 upcoming games. The instructions asked participants to indicate, for each game, which team they predicted to win, and their level of confidence in their estimate.

To view the online questionnaire as a participant, follow this link: https://berkeley.ca1.qualtrics.com/jfe/preview/SV_9WxWwutCRnyLaa9?Q_SurveyVersionID=current&Q_CHL=preview

Hindsight Manipulation. During the review portion, participants in the control condition, reviewed each of the 20 games one at a time. I presented participants with the teams, each team's record of wins and losses prior to the game, day of the game, and score for each game, with the winner. These participants were told to review these games "in order to get an idea of what recent scores are like."

In the hindsight condition, participants answered two questions for each past game. For each game outcome I asked participants which team they *had* expected to win this game, and also how confident they *would have* been prior to the game. The following is the language participants read:

> In the following section, you will be shown a series of scores from NBA games that occurred in recent days. For each game, imagine that you were asked to predict the winner BEFORE the game happened. Please report who you WOULD HAVE predicted to win the game. Also, report how confident you WOULD HAVE been in your prediction.

For each upcoming game, all participants predicted the winning team, with a confidence level for their prediction. Participants read:

> Now in this next section, you will see a series of games that have NOT yet occurred. For each game, please predict who you think will win the game, and mark how confident you are in that prediction.

Confidence. For all confidence estimates, participants reported they were/would have been in their prediction, "How confident would you have been in your prediction of the winner?" on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*.

Hit Rate. Participant answers were coded for coded for correctness and a mean hit rate was calculated for each participant by averaging correctness across all items.

Incentivization. I incentivized participants using the Quadratic Scoring Rule, or the Q score. The Quadratic Scoring Rule is a common incentive-compatible scoring rule for assessing probabilistic predictions of categorical outcomes (Brier, 1950; Selten, 1998). Participants' Q scores were directly proportional to their chances of winning a $20 bonus. Participants saw the following explanation of the incentivization scheme:

> Please do your best to report exactly how confident you are for each prediction. The more well-calibrated your responses, the better your chance of winning the $20 prize.
>
> For each prediction you can earn between 2 and -2 points. If you are 100% confident and correct, you earn all 2 points. But if you are 100% confident and wrong, you lose 2 points. Reporting confidence less than 100% gets you somewhere between -2 and 2, depending on your confidence and if you are correct.
>
> The higher your point total at the end, the greater your chance of winning the $20 lottery prize, which will be bonused to your account at the end of the study.

Results

Did providing participants with outcome feedback increase the confidence in their forecasts? An independent samples t-test compared forecast confidence between the two conditions. There was no significant difference in mean confidence between the control condition (*M* = 71.15, *SD* = 13.77) and the hindsight condition (*M* = 68.02, *SD* = 12.89), *t*(199) = 1.83, *p* = 0.07, *d* = 0.23. Within the hindsight condition, there was a significant difference between average confidence among the 20 review items (*M* = 66.01, *SD* = 10.79) and the average confidence in the 20 forecasted items (*M* = 68.02, *SD* = 12.89), *t*(99) = 2.75, *p* = .006, *d* = 0.17 (see Figure 1).

Did past confidence impact accuracy for future events? I employed a linear regression to test whether, within the hindsight condition, greater confidence reported with outcome feedback stimuli would predict lower Q scores for game forecasts. A regression analysis failed to provide evidence of such a relationship, $\beta = 0.1$, *t*(99) = 1.10, *p* = 0.28.

Discussion

Participants did not express greater confidence in their forecasts after they had their confidence in hindsight for past events. This experiment and this result do not provide any evidence directly in support of or against the phenomenon of the hindsight bias. Rather, this first experiment primarily examined antecedents of forecast accuracy. In this line, the results suggest previous confidence does not impact future confidence nor accuracy.

Experiment 1 failed to find evidence of an effect of hindsight bias on subsequent forecasting accuracy. When I designed and conducted Study 1, it was intended to assess forecasting calibration. Experiment 2 serves as another attempt to test the connection between hindsight and forecast accuracy. This study also tested for the basic hindsight effect prior to eliciting any forecasts, to ensure that the paradigm could manifest hindsight bias in the first place.

CHAPTER 3

Experiment 2: Hindsight in Forecasting Sports

Experiment 1 did not provide evidence suggesting outcome feedback impacts forecasting accuracy. Experiment 2 serves as a second attempt to find evidence for this association. However, Experiment 2 also focuses separately on the basic hindsight bias relationship. In this experiment, I elicited confidence levels not only for forecasting stimuli, but also for the past stimuli presented with outcome feedback. Measuring confidence at both instances serves as another attempt at the goal of Experiment 1, while also providing a direct test for the hindsight bias.

Experiment 2 only measures confidence and hit rates for forecasts without Q scores. As Q scores are a direct measure of hit rate combined with confidence, and Experiment 1 provided no evidence to suggest the outcome feedback manipulation affects forecasting hit rates, I considered confidence as my primary dependent measure in this study.

Following Experiment 2, the direction of the research program shifted toward examining hindsight more directly at the expense of further studying forecasting accuracy. Finally, Experiment 2 utilized baseball game stimuli in lieu of basketball games due to a change in season, though prediction confidence is not different across sports (Kelly & Simmons, 2016).

Method

Participants

This study included 200 Amazon Mechanical Turk participants (70% male, M_{age} = 34.70) paid $1.50. The survey instructed participants that well-calibrated responses would increase their chance of winning the $20 bonus.   The survey randomly assigned participants to two between-subjects conditions. The survey excluded all participants that had participated in Experiment 1. I again did not know what effect size to expect, so I chose this number in hope that it would be an adequate sample size.

Design

The design and main dependent variables were the same as Experiment 1: a 2-cell (outcome feedback vs. control) between-subjects design that manipulated whether or not participants provided post-hoc answers and confidence levels during the review portion of the study. The survey randomly assigned participants to the two between-subjects conditions, for 100 in each condition. The main dependent variable was confidence (both in review and forecasted), and hit rates were also measured.

Procedure and Materials

Overview. The procedure was similar to Experiment 1 with the addition of a confidence measure during the review items of the study. Participants reported who they thought to be the winning baseball team for 30 past games, and they made forecasts for 10 games yet to occur in the upcoming week. Participants provided their confidence level for each of the 40 items. Participants read the following instructions for the review items:

> In the following section, you will be shown a series of scores from 30 MLB games that have occurred on either April 5th, April 9th, or April 13th. For each game, try your best to identify the winning team. Then, report how confident you are in your answer.

Participants then read similar forecasting instructions as in Experiment 1 prior to submitting their ten forecasts.

To view the online questionnaire as a participant, follow this link: https://berkeley.ca1.qualtrics.com/jfe/preview/SV_djlQh87oiicZz01?Q_SurveyVersionID=current&Q_CHL=preview

Hindsight Manipulation. During the review portion, participants in the control condition, reviewed each of the 30 games one at a time. I presented participants with the same information for each game as in Experiment 1, except without the final score of each game. In this experiment participants guessed the winning team of each of the 30 games and provided a confidence level for each game.

In the hindsight condition, participants saw the same information as those in the control condition with the addition of the final score for each game and the winning team clearly identified. As in Experiment 1, I asked participants which team they had expected to win each game, had they not just seen the final score. Additionally, I asked participants how confident they would have been, had they not just seen the final score. They answered these two questions for each of the 30 review games. Participants read the following introduction:

> In the following section, you will be shown a series of scores from 30 MLB games that have occurred on either on April 5th, April 9th, or April 13th.
>
> For each game, even though you will be shown the final score, please choose which team *you would have predicted* to win, assuming you didn't just see the score.
>
> Then, report how confident *you would have been* for each answer.

Following the review portion, participants then made forecasts for ten games yet to occur in the upcoming week. For each game, participants answered the same questions as they did for Experiment 1.

Confidence. As in Experiment 1, the main dependent variable was confidence level for each game, on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*.

Hit Rate. Participant answers were coded for correctness and a mean hit rate was calculated for each participant by averaging correctness across all items.

Incentivization. I incentivized participants using the Quadratic Scoring Rule as in Experiment 1. Participants' Q scores were directly proportional to their chances of winning a $20 bonus.

Results

Did providing outcome feedback to participants increase what they reported their confidence *would have been* for past games? An independent samples t-test measured whether outcome feedback led to greater average confidence expressed for the 30 games from the prior week. There was no significant difference in mean confidence between the control condition (*M* = 62.49, *SD* = 11.27) and the hindsight condition (*M* = 64.43, *SD* = 10.59), *t*(199) = -1.25, *p* = 0.22, *d* = -0.18 (See Figure 2).

Did outcome feedback for past games impact confidence on future games? I employed another independent samples t-test to measure the effect of condition on participants' average confidence for their baseball forecasts. The test failed to provide evidence of a significant relationship, as the mean forecast confidence between the control condition (*M* = 69.34, *SD* = 16.63) and the hindsight condition (*M* = 72.39, *SD* = 14.69) was not significantly different, *t*(199) = -1.37, *p* = 0.17, *d* = -0.19.

Did outcome feedback impact reported hit rate for either past or future games? An independent samples t-test comparing reported hit rate for the past 30 games between hindsight and control conditions. The test showed that the mean confidence for the control condition (*M* = 0.58, *SD* = 0.10) was significantly lower than the mean confidence of those in the hindsight condition (*M* = 0.62, *SD* = .12), *t*(199) = -2.62, p = 0.01, *d* = -0.36. A similar test explored the effect of condition on forecast hit rate. Replicating Experiment 1, the test showed no evidence of a difference between the control condition (*M* = 0.56, *SD* = 0.17) and the hindsight condition (*M* = 0.60, *SD* = 0.17), *t*(199) = -1.38, p = 0.17, *d* = 0.24.

This is the first experiment which tests the hindsight bias effect in the classic paradigm for past outcomes. The primary dependent measure for hindsight bias in the foundational literature and in this research program is the confidence participants claim (or would have claimed) in knowing the correct outcome. In order to calculate Q scores I also analyzed participant hit rates. While a difference in hit rates between conditions may intuitively seem like a manifestation of hindsight, the classic operationalization of the phenomenon compares confidence as the key dependent measure. While I measured hit rate data in all experiments, I will cease to report any hit rate analyses from here on.

Discussion

Experiment 2 found that participants did not express greater confidence in their forecasts after provided with outcome feedback for past events. This replicates the evidence from Experiment 1. In addition, this experiment found that providing outcome feedback for past events did not increase reported confidence for those events, which is a failure to replicate the hindsight bias effect.

Experiment 2 served to replicate the findings of Experiment 1. Using stimuli from two different sports, I found that not only did outcome feedback not impact confidence in forecasts, but it also did not impact confidence for the same past events which the outcome feedback directly answered. It is here in the research program in which the focus of study shifts from forecasting to an exploration of the hindsight bias. While the outcome feedback paradigm used Experiments 1 and 2 has successfully elicited hindsight in the past (Hoch & Loewenstein, 1989) using sports stimuli is novel, and may present a possible boundary condition for hindsight. Experiment 3 seeks to test which boundary conditions could exist in moderating the hindsight bias effect, resulting in a failure to observe the phenomenon.

CHAPTER 4

Experiment 3: Robustness of Hindsight on Novel Stimuli

Experiment 3 tests a possible boundary condition of hindsight bias; knowledge domain. In other words, perhaps hindsight bias manifests differently (or not at all) depending on the type of information elicited (i.e. the type of stimuli used). This is a significant implication, as a majority of foundational hindsight bias research employed only one domain of knowledge – general knowledge trivia (Hawkins and Hastie, 1990).

Much past research in this area used questions with epistemic uncertainty, in which the answer was knowable. Many of the landmark hindsight studies used trivia (for example, Fischhoff, 1975; Fischhoff, 1977), but in the present day participants can discover the answers to trivia questions with a quick internet search, making them flawed stimuli for an online study. Instead I used a category of epistemic questions in which the answers were not searchable online: guessing individuals' weights from photographs. I was not concerned with responses for weight guessing stimuli being open-ended, as previous research has also replicated the hindsight bias effect with outcome feedback when participants provide their own answers in an open-ended format (compared to choosing from two options) (Hoch & Loewenstein, 1989).

Method

Participants

This study included 200 Amazon Mechanical Turk participants (66% male, M_{age} = 35.54) paid $1.50 each. The survey randomly assigned participants to four between-subjects conditions. The survey excluded any participants who participated in any prior studies. I determined this sample size a priori with the goal of including 200 participants, for 50 participants per cell. I based my power analysis by first aggregating previous outcome feedback results in the literature, which provided an average effect size of f = .35 (Hoch & Loewenstein, 1989; Slovic & Fischhoff, 1977; Wood, 1978). Seeking power of .90 with an alpha level of .05, analyses indicated 35 participants per condition would be sufficient. However, I am cognizant of the low sample sizes used in the studies from which this average effect size was computed, and decided (ex-ante) to collect 50 participants per condition for a more reliable result.

Design

The experiment was designed to present another test of hindsight similar to Experiment 2, with the addition of a new stimulus type. This experiment was a 2 (outcome feedback vs. control) X 2 (stimulus type: baseball vs. weight guessing) between-subjects design, which manipulated how much information participants received, as well as the type of questions they answered. The main dependent variable was reported confidence, and hit rates were also measured.

Procedure and Materials

Overview. The materials differed from the first two experiments in that there was no forecasting and all items were thus epistemic in nature (i.e. the answers were knowable). Participants were randomly assigned to answer either baseball game questions or weight guessing questions, and were also assigned to only see the question, or to see the question and correct answer identified simultaneously. For each item, participants reported what they thought to be the correct answer as well as their confidence level.

To view the online questionnaire as a participant, follow this link: https://berkeley.ca1.qualtrics.com/jfe/preview/SV_0TEsjy23umYLS17?Q_SurveyVersionID=current&Q_CHL=preview

Stimuli Manipulation. In the baseball stimuli condition, participants read instructions identical to Experiment 2 concerning the estimates they would make. Then, they reviewed each game one at a time, providing a team choice and confidence level for each. These stimuli were identical to the stimuli used in Experiment 2.

In the weight guessing condition, participants read the following instructions:

> In the following section, you will be shown a series of pictures of people. For each person, try your best to guess how much they weigh in **pounds**. Your answer for each will be considered **correct** if it is within 10 pounds of their true weight. Then, report how confident you are that your answer is within **10 pounds** of their true weight.

Participants then reviewed ten images of individuals and provided their weight estimate and confidence level for each.

Hindsight Manipulation. The hindsight manipulation was in the form of the outcome feedback paradigm used in the previous experiments. The manipulation as applied to baseball stimuli was identical to that of Experiment 2. The following is an example of the dependent measure language used for weight guessing stimuli:

> This person's actual weight is: 140 pounds.
> A correct answer would be guessing a weight between 130-150 pounds.
>
> How much would you have said this person weights, assuming you had NOT just seen the correct answer?
>
> How confident *would you have been* that your answer would be within 10 pounds of their true weight?

Confidence. As in previous experiments, the main dependent variable was confidence level for each item, on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*.

Results

Did stimulus type have an effect on the relationship between outcome feedback and reported confidence? An omnibus two-way ANOVA showed neither a main effect of stimulus type on confidence, *F*(1,195) = 0.26, *p* = 0.61, nor a main effect of outcome feedback on confidence, *F*(1, 198) = 3.10, *p* = .08 (see Figure 3).

Did outcome feedback increase reported confidence for either type of stimulus? I employed independent samples t-tests for each stimulus type to test the impact of outcome feedback on confidence. For baseball stimuli, there was no significant difference in mean confidence between the control condition (*M* = 62.80, *SD* = 16.80) and the experimental condition (*M* = 65.52, *SD* = 14.38), *t*(99) = -0.86, *p* = 0.39, *d* = -0.17. For weight guessing stimuli, there was no significant difference in mean confidence between the control condition (*M* = 62.54, *SD* = 21.11) and the experimental condition (*M* = 68.37, *SD* = 15.58), *t*(99) = -1.57, *p* = 0.12, *d* = -0.31.

Discussion

Experiment 3 failed to provide evidence suggesting that the type of stimulus used in testing hindsight bias could moderate the effect. Once again, participants did not express greater confidence in their responses after provided with outcome feedback. This replicates the evidence from Experiment 2, which also did not provide supporting evidence of the hindsight bias.

Using two different stimuli domains I found that outcome feedback did not impact confidence for items the outcome feedback directly answered. In order to more thoroughly test the potential moderating effect of stimulus type on hindsight bias elicitation, Experiment 4 systematically varies stimuli across several domains in order to find evidence supporting replication of hindsight.

CHAPTER 5

Experiment 4: The effect of chance on hindsight

Experiment 4 draws on the theorized mechanism in the literature of ‘creeping determinism’ to inform its design. My previous studies of the hindsight bias resulted in curious failures to replicate the basic effect. Participants in the outcome feedback condition did not report more confidence in their responses compared to the control condition in any of the prior three studies. These repeated failures to replicate hindsight bias demanded closer inquiry.

Experiment 4 draws on a mechanism at first only theorized called *creeping determinism* (Fischhoff, 1975). Fischhoff surmised that, due to the hindsight bias process entailing automatic processing of outcome information, individuals would come to perceive an outcome as inevitable. This theory was empirically tested by Wasserman et al. (1991) when they used the stimuli from Fischhoff (1975). They manipulated the outcome feedback of the 19th century wars used in the original study, in which participants received original outcome feedback of the result, with either a ‘deterministic’ explanation (e.g. troop discipline) or a ‘chance’ explanation (e.g. an unexpected monsoon). When provided a chance explanation for an outcome, participants showed no hindsight bias effect.

Contextualized in Fischhoff’s *creeping determinism* theory, hindsight bias only manifests insofar as individuals perceive the outcome as ‘inevitable’, or, pre-determined. Wasserman et al. (1991) posited that the presence of chance in determining an outcome leads to a lack of ‘determinism’ and therefore a lack of hindsight bias. Grounded in my previous work, Experiment 4 tests whether hindsight bias fails not just when there is ‘chance’ present, but when an outcome can vary over time.

This ‘temporal variance’ in outcomes subsumes outcomes determined by chance, and also includes epistemic outcomes that contain little chance (e.g. the weight of someone in an image or baseball games). In other words, the truth of certain domains can change over time. As creeping determinism relates to a belief that an outcome is inevitable, the more unpredictable an outcome is, it logically follows that it would be less deterministic. At the most extreme end of this spectrum, purely chance events such as coin flips are rarely regarded as pre-determined, as they are the most unpredictable. However, the weight of individuals does change slowly over time, and the ‘winner’ of a Yankees-Red Sox match-up changes multiple times a year (i.e. whenever the winning team changes). These examples are more ‘constant’ in time than coin flips, but they still vary greatly relative to the general trivia questions present in most of the classic hindsight bias research (as the correct answer to a trivia question can stay constant through time.

This experiment examined the robustness of hindsight bias using a variety of stimuli which range in their degree of temporal variance. While one empirical finding suggests a boundary condition of hindsight bias when it comes to the presence of chance in an outcome (Wasserman, Lempert, & Hastie, 1991), this logic has not yet been extended to apply to entire domains of knowledge (namely, any outcomes that are not constant over time, i.e. sports, quarterly profits, political events, etc.).I began to wonder whether there exists a range of the degree to which an outcome can vary over time, and seek to explore the effects of hindsight in the context of this range.

Method

Participants

This study included 200 Amazon Mechanical Turk participants (70% male, M_{age} = 34.70) paid $1.50 each. The survey randomly assigned participants to two between-subjects conditions. The survey excluded any participants who participated in any prior studies. This sample size was determined using the same rationale as that of Experiment 3.

Design

The experiment contained two between-subjects conditions (outcome feedback vs. control), and four within subjects conditions (stimulus type: baseball, coin flips, weight guessing, and general trivia). The survey randomly assigned participants to one of the two between-subjects conditions. Each participant saw the four within-subjects conditions in a randomized order to prevent any possible order effects. The main dependent variable was reported confidence, and hit rates were also measured.

Procedure and Materials

I compiled the trivia questions from a variety of sources (e.g. Fischhoff, 1977). The survey randomly assigned half of the participants to the experimental condition, in which they saw the correct outcome for each question. For each outcome I asked participants some variation of “Which outcome would you have believed to be correct, assuming you had not just seen the correct answer?” They also reported their confidence in their expectation on a 1-100 scale.

In the control condition, participants experienced the same portions of the study as those in the experimental condition, but without any outcome information. They responded with which outcome they believed to be true for each question with a confidence measure 1-100.

Overview. The materials and procedure included direct replications of Experiment 3 with the addition of two novel stimulus types. Participants reviewed 40 items in total, ten items for each of four stimulus types. As in Experiment 3, participants were randomly assigned to only see the question, or to see the question and correct answer identified simultaneously. For each item, participants reported what they thought to be the correct answer as well as their confidence level.

The baseball and weight guessing stimuli sections were direct replications of Experiment 3. I compiled the trivia questions from a variety of sources (e.g. Fischhoff, 1977). Participants in the control condition read the following prior to the trivia portion:

> Now you will see 10 general knowledge trivia questions. For each of the following 10 questions, please indicate which choice you think is correct. Then, mark how confident you are in your choice being the correct answer.

For coin flip stimuli, I simulated ten coin flips using a random number generator and recorded the first ten iterations as the 'true' events for the purposes of hit rate calculation and in order to be able to provide outcome feedback. The coin flip results were: Tails, Heads, Heads, Tails, Heads, Tails, Heads, Heads, Tails, Tails – T H H T H T H H T T). Participants read the following:

> Now, you will guess the result of 10 different coin flips simulated in this program, each done with a fair coin. This means that each coin flip is completely random, with an equal chance of coming up "Heads" or "Tails". For each flip, please mark whether you think the result will be "Heads" or "Tails", and then mark how confident you are in your answer.

To view the online questionnaire as a participant, follow this link: https://berkeley.ca1.qualtrics.com/jfe/preview/SV_0jOKoFEXggr5Uwd?Q_SurveyVersionID=current&Q_CHL=preview

Hindsight Manipulation. The hindsight manipulation was in the form of the outcome feedback paradigm used in the previous experiments. Below is an example of the language for the domain of general trivia:

> Now you will see 10 general knowledge trivia questions. For each question, even though you will be shown the correct answer, please mark which choice you would have guessed to be correct, assuming you didn't just see the correct answer. Then, report how confident you would have been for each answer.

Confidence. As in previous experiments, the main dependent variable was confidence level for each item, on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*.

Results

Did comparing stimulus type with outcome feedback impact reported confidence? An omnibus mixed ANOVA showed only a main effect of stimulus type on confidence, $F(3,198) = 52.40$, $p < .001$, and no main effect of outcome knowledge on confidence, $F(1, 198) = 1.72$, $p = .16$ (See Figure 4).

Were some stimuli more susceptible to hindsight than others? I predicted trivia questions would elicit the greatest hindsight bias (that is, the greatest difference in average confidence between conditions). Using planned contrasts out of the earlier mixed ANOVA, results showed no significant interaction between stimulus type and hindsight condition on confidence, *F*'s < 2.29, *p*'s > .13.

Discussion

The results show that the outcome feedback paradigm again failed to elicit hindsight. Experiment 4 systematically measured the effect of outcome feedback in eliciting hindsight bias for a variety of stimulus types. Of note was the failure to replicate hindsight bias with general trivia stimuli, which is the foundational domain in which researchers first examined hindsight.

These analyses of these first four experiments fail to provide any evidence for the existence of the hindsight bias. If stimulus type is not the cause of these failures to replicate, what could be? The following experiments in this research program test several possibilities. One reason could be that the hindsight bias presents more consistently in a longitudinal paradigm (e.g. Fischhoff & Beyth, 1975). Alternatively, perhaps the online environment for the studies impacted the manifestation of the bias in some way. Lastly, researchers have acknowledged in the literature already (Wasserman et al., 1991), perhaps there is some attribute of the original general trivia stimuli used in many of the foundational hindsight papers (Fischhoff, 1977; Hoch & Loewenstein, 1989) which has provided an inflated sense of the bias.

CHAPTER 6

Experiment 5: A within-person paradigm of hindsight

In Experiment 5, I utilized an alternative method of eliciting the hindsight bias also found in the literature (Fischhoff & Beyth, 1975), wherein hindsight is measured within-subject instead of between. In this paradigm, participants make judgments about general trivia, and after some time, attempt to recall their initial judgments after they learn of the event outcome. It is important to consider this alternative paradigm as it may provide evidence that the hindsight bias literature is better supported through a longitudinal elicitation compared to one between-subjects.

Experiment 5 also explored the possible moderator of surprisingness. This moderator, originally proposed by Fischhoff (1975), suggests that events whose outcomes are quite surprising do not elicit hindsight bias as much as events which are considered less surprising, or more predictable. This explanation is derivative of the creeping determinism hypothesis much like the perceived randomness moderator. However, this approach examines the phenomenon of hindsight bias from an emotional lens as opposed to a probabilistic lens. . While this moderator was first proposed by Fischhoff (1975), it is lacking solid experimental evidence (Christensen-Szalanski & Willham, 1991). I expected that this method of elicitation will show the presence of hindsight bias in trivia questions, and that greater ‘surprisingness’ will reduce the hindsight bias. In this study, the time between initial judgment and subsequent recall was relatively short, and participants engaged in a short distractor task in order to prevent memorization of the initial judgment.

Method

Participants

The final sample includes 189 participants (46% male, M_{age} = 20.91) between two between-subjects conditions. Participants completed this experiment in person through the Haas Research Participation Pool (RPP) in the lab for either $1.00 or course credit. The survey was advertised as a survey about people’s beliefs about predicting events. While I pre-determined a sample size of 100 total (50 per cell), per the same rationale used in Experiments 3 and 4, I collected data until the semester ended.

Design

The experiment had a 2-cell (distractor task vs. no distractor task) between-subjects design. The within-subjects factor consisted of two stages. In the initial stage, participants first responded to 25 trivia questions. In the recall stage, they attempted to recall their initial estimates.

Procedure and Materials

Overview. In the initial stage, participants reported their initial judgments on 25 general trivia questions, consisting of their answer choice and confidence level. In the second stage, I informed participants of the true outcome feedback (answer) and asked participants to recall their initial answer choice and initial confidence level. Participants also reported how surprised they felt after learning the answer to each item. I predicted that participants' recall of their initial estimates would systematically adjust closer to the correct answer and their recalled confidence would increase after they learned the correct outcomes.

To view the online questionnaire as a participant, follow this link (same link as Experiment 10): https://berkeley.ca1.qualtrics.com/jfe/preview/SV_25iQkX8PRUoxR2J?Q_SurveyVersionID=current&Q_CHL=preview

Distractor Task Manipulation. The distractor task manipulation contained a three trial memorization task in which participants were instructed to memorize a list of words and then identify those words on a subsequent page within a larger list of words. Participants in the control condition simply progressed immediately from the initial stage to the second recall stage.

Confidence. As in previous experiments, the main dependent variable was confidence level for each item, on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*. This was measured in the first stage.

Recall Confidence. On the same scale, participants also attempted to recall their initial confidence measure. These responses were coded separately for analyses.

Surprisingness. All participants rated their feeling of surprise on a 1-7 scale during the recall stage for each item.

Results

Did the presence of a distractor task impact the recall of initial confidence? A 2 (initial stage vs. recall stage, within-subjects) X 2 (distractor condition vs. control condition) omnibus mixed ANOVA showed only a weak main effect of stage on confidence, *F*(1,188) = 4.06, *p* = .045, and no main effect of distractor task on confidence, *F*(1, 188) = 1.51, *p* = .22. A closer look compared confidence within-subjects between the initial and recall stages with a paired t-test. This test also showed that the mean confidence during the initial stage (*M* = 53.28, *SD* = 14.33) was lower than the mean confidence reported in the recall stage (*M* = 54.46, *SD* = 16.24), *t*(188) = 2.01, *p* = .045, *d* = 0.08 (See Figure 5).

Did participants' perceptions of how surprising each item was impact their reported confidence? A linear regression predicting change in confidence by condition and reported surprise likewise does not show a main effect of surprisingness on confidence change, b = -0.53, *t*(186) = -0.50, *p* = .62.

Discussion

Experiment 5 was a necessary experiment in order to be thorough in testing the replicability of the hindsight bias. While most hindsight bias literature employs the between-subjects outcome feedback manipulation used in the prior studies, researchers have also used this within-subjects paradigm. This alternative paradigm offers very weak evidence in support of the hindsight bias. The lack of an effect for the distractor task could be due to the distractor task not having created a sufficient enough delay in order to facilitate the memory decay necessary for a larger change in confidence due to recall failure.

The design of Experiment 5 allowed me to address a large portion of hindsight bias literature through the use of the within-subjects recall paradigm while looking at the moderator of surprisingness. Experiment 6 continues in this longitudinal trend with a longer timeline, providing a more faithful replication with past research using the recall paradigm, as well as addressing the weakness of Experiment 5 in its insufficient delay.

CHAPTER 7

Experiment 6: A longitudinal paradigm of hindsight

In Experiment 6, I tested whether the hindsight bias would replicate using a longitudinal recall paradigm. The repeated failures to replicate hindsight bias in previous experiments could be due to the use of the outcome feedback between-subjects paradigm. In that event, the bias may be more susceptible to the paradigm used to test the phenomenon than any other moderator. The weak evidence in support of the hindsight bias in Experiment 5 supports this theory. Instead of a distractor task, this study took place in two phases, with Phase 1 occurring before the 2018 midterm elections and Phase 2, occurring afterwards. Political elections have been popular fodder for studying hindsight in the past, though with few successful attempts at moderation (Leary, 1982; Powell, 1988). This study most clearly resembles landmark hindsight bias research (Fischhoff & Beyth, 1975) and tested the robustness of hindsight bias in the domain of the 2018 midterm elections.

This study examined an additional moderator of participant expertise which has been previously operationalized as ‘familiarity’ with the task, but not breadth of subject area knowledge (Christensen-Szalanski & Willham, 1991). However, researchers have studied the effect of expertise on hindsight with differing results. Findings show doctors making decisions in a field in which they are expert still display hindsight bias (Arkes et al., 1981). Additionally, however participants with knowledge of the exact stimuli being tested (through a ‘learning’ portion followed by a ‘testing’ portion for the same set of stimuli show less bias (Hertwig, Fanselow, & Hoffrage, 2003).

In the context of these seemingly inconsistent patterns regarding the effect of expertise on hindsight, I sought to explore how expertise in a subject area affects hindsight bias. I chose this operationalization as it increases the generalizability regarding the power of expertise to the general population, as individuals can be knowledgeable about certain subject matter without making it a career. I predicted that greater expertise in the subject area of U.S. politics would amplify hindsight bias in participants. It is in this paradigm which I tested this moderator. I expected the main results to replicate the hindsight bias and provide further evidence of the longitudinal paradigm successfully eliciting hindsight bias, and for greater expertise to amplify the hindsight bias effect.

Method

Participants

The final sample includes two hundred participants (61% male, M_{age} = 35.37) divided evenly between the two phases of data collection. Participants completed this experiment through Amazon Mechanical Turk for $1.00. The survey was advertised as a survey about people’s beliefs about predicting events. This study included 100 participants before the election and 100 participants after, for 200 participants total. This sample size was determined using the same rationale as in my previous studies.

Design

The experiment consisted of two stages. First, participants provided forecasts for the number of seats won by Democrats and Republicans for the Senate and House of Representatives overall, respectively. Confidence levels were elicited for each of the two predictions. Second, as a test of participant expertise in the field of U.S. politics, participants answered a ten item test on political knowledge. The primary dependent variable will be participant confidence in their predictions, comparing participants prior to the elections to those after the elections. Additionally, predictions in how many seats in the House of Representatives and in the Senate will be/were won by each party will be collected, as will a measure of political expertise.

Procedure and Materials

Overview. Phase One of data collection occurred prior to the midterm elections. Participants submitted two sets of predictions (a Senate prediction and a House prediction), each with a respective confidence level. The survey program presented the elections questions to each participant in a randomized order after providing preliminary information about the election landscape. This is what participants read prior to submitting a Senate prediction before the election:

> In the 2018 Midterm Elections, there are 35 Senate Seats up for election. Currently, the Republican Party has 51 seats and the Democratic Party has 49 seats. To have control of the Senate a party needs 51 seats (or 50 with a Vice President to break ties).
>
> A recent composite forecast states that of the 35 open seats, 8 lean Republican, 25 lean Democrat, and 2 are toss-ups.
>
> How many seats will go to each party out of these 35 seats? (Total must sum to 35).

Participants then submitted a confidence level for their prediction, for example: “How confidence are you in your prediction of the Senate races?” ranging from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*. The composite election forecasts referenced in the survey were created by fivethirtyeight.com.

Phase Two of data collection occurred after the midterm elections results were known. A sample of different participants completed a similar survey which provided similar information to Phase One with the addition of the outcome information for each set of elections, again in a randomized order. Participants reported how many seats they had predicted to be gained and lost prior to the election, with how confident they were in each prediction prior to the election. The following is the House elections questions from Phase Two for context:

> In the 2018 Midterm Elections, all 435 seats in the House of Representatives are up for election. Before the election, the Republican Party had 240 seats and the

> Democratic Party had 195 seats. To have control of the House, a party needs 218 seats.
>
> **The results of the 2018 Midterm Elections saw the Republican Party win 200 seats and the Democratic Party win 235 seats.**
>
> BEFORE the election happened, how many seats DID YOU THINK would go to each party? (Total must sum to 435).

Participants provided their past confidence through a method similar to the previous experiments with outcome feedback: "How confident WERE you in your prediction of the House races, BEFORE the election happened?" ranging from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*.

To view the Phase One online questionnaire as a participant, follow this link: https://berkeley.ca1.qualtrics.com/jfe/preview/SV_cAbICTSWWxNNg2h?Q_SurveyVersionID=current&Q_CHL=preview

To view the Phase Two online questionnaire as a participant, follow this link: https://berkeley.ca1.qualtrics.com/jfe/preview/SV_eWKCNS7QB2yPFYh?Q_SurveyVersionID=current&Q_CHL=preview

Confidence. All participants in both Phase One and Phase Two predicted a quantitative number of seats gained and lost by each party in both the House and the Senate and rated their confidence on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*.

Political Expertise. Participants in both phases completed the same ten-item multiple choice quiz to test their knowledge of the U.S. political system and current events in order to create an expertise score for each participant. This measure is used as a control variable in analyses.

Results

Did hindsight bias manifest when participants reported confidence for the election results after the results were known? Independent samples t-tests showed non-significant differences in confidence between Phase 1 and Phase 2 for both the House and the Senate races, *t*'s < .96, *p*'s > .34, failing to replicate hindsight bias (See Figure 6).

How did political expertise affect confidence levels? I ran linear regression models for both the House and the Senate, predicting confidence by phase group and political expertise. There was no main effect of phase on confidence, meaning that participants after the elections were not more or less confident than those before the election. There was a main effect of expertise on confidence in both the House races, b = 24.57, *t*(199) = 3.09, *p* = .002, and the Senate races, b = 17.33, *t*(199) = 2.11, *p* = .04. In other words, more expert participants reported greater average confidence for both estimates. However, there was no significant interaction between expertise and responding after the results were known, b = -6.66, *t*(199) = -0.57, *p* = .56.

Was this greater confidence on the part of experts warranted? Sort of. Linear models predicted absolute distance from the actual result with confidence and expertise, using the data prior to the election. This test showed that before the election, there was a non-significant effect of expertise for the House, b = -18.23, *t*(199) = -1.60, *p* = .11, and a significant main effect for the Senate, b = -3.50, *t*(199) = -2.29, *p* = .02. These results suggest that participants who were more expert in the field were more accurate for the Senate forecast, but not the House forecast.

Discussion

Experiment 6 complements Experiment 5 in testing the replicability of the hindsight bias using a longitudinal paradigm. However, the hindsight bias did not replicate in this experiment. Participants responding after the midterm elections did not show more confidence, nor did they claim more accurate predictions, compared to participants who responded before the elections occurred. Similarly, subject matter expertise did not moderate any hindsight bias effect. Expertise did increase confidence during both phases, but again with no difference between those who responded before vs. after the elections.

Searching for evidence in support of the hindsight bias, the alternative longitudinal paradigm used in Experiments 5 and 6 showed weak and inconsistent results. The next direction I pursued in searching for a potential moderator of hindsight entailed the theorized mechanism of ‘creeping determinism,’ proposed by Fischhoff himself (1975).

CHAPTER 8

Experiment 7: Hindsight with perceived randomness

Experiment 7 explores Fischhoff's creeping determinism hypothesis through measuring perceptions of randomness in events. In pursuit of identifying a strong moderator that could attenuate the hindsight bias that is similar but not identical to stimulus type, this study examines various outcomes and their degree of perceived "randomness", as I predicted that perceptions of greater randomness would attenuate the hindsight bias, as Wasserman et al. (1991) suggest could be the case.

This experiment uses two different sets of stimuli in follow up to Experiment 4: coin flips, to emulate a completely random event, weight-guessing items, to emulate a more 'deterministic' event. Comparing hindsight bias across different stimulus types while measuring perceptions of randomness could suggest that certain attributes about each stimulus could play a role in whether hindsight manifests.

Participants provided perceptions of how much randomness played a part in the result of the outcomes, in which the more an outcome appears randomly determined, the less susceptible participants should be to hindsight bias, according to the creeping determinism hypothesis (Fischhoff, 1977). This moderator is undergoes a direct test in this study.

Method

Participants

The final sample included 205 participants. I originally planned for two hundred participants determined using the same rationale as in my previous studies. Participants completed this experiment online through Amazon Mechanical Turk for $1.00. The survey was advertised as a survey about decision making ability.

Design

The experiment had a 2 (outcome feedback vs. control) X 2 (coin flips vs. weight guessing) cell design that manipulated whether or not participants receive the actual outcome prior to making their estimate and confidence measure. The primary dependent variable was confidence in their chosen prediction for each item. Participants reported the degree to which they perceived the outcomes as randomly determined. Finally, hit rates were also collected.

Procedure and Materials

Overview. The Qualtrics survey randomly assigned participants to either the outcome feedback condition (*n* = 50) or the control condition (*n* = 50). Additionally, the survey randomly assigned participants to either the coin flips condition (*n* = 50) or the weight guessing condition (*n* = 50). The Qualtrics survey program then presented the series of stimulus items to each participant in a randomized order.

Out of concern that the null result of weight guessing in Experiment 4 was due to participants believing weight guessing as a more randomly determined activity than it is in reality, participants reviewed five images with weight information provided prior to the weight guessing portion.

To view the online questionnaire as a participant, follow this link:
https://berkeley.ca1.qualtrics.com/jfe/preview/SV_4YGzeXMsRJYEWQB?Q_SurveyVersionID=current&Q_CHL=preview

Hindsight Manipulation. Identical to Experiment 4 and others, participants in the outcome feedback condition saw the correct outcome or answer prior to providing their own answer and confidence level.

Perceived Randomness. Participants in both conditions completed a series of four items that measured how 'random' participants perceived the event. For example, the four items for the domain of weight guessing were: 1) "How difficult is it to correctly guess someone's weight from a photograph?" ranging from 1 = *Very Easy* to 7 = *Very Difficult*; 2) How much would additional information help you in correctly guessing someone's weight from a photograph?" ranging from 1 = *It would not help at all* to 7 = *It would help a great deal*; 3) How much randomness is involved in trying to correctly guess someone's weight from a photograph?" ranging from 1 = *Very Little Randomness* to 7 = *A Lot of Randomness*; and 4) "How would an expert at this task perform compared to you?" ranging from 1 = *An expert would be much worse* to 7 = *An expert would be much better*.

Confidence. As in previous experiments, the main dependent variable was confidence level for each item, on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*.

Results

Did the stimulus type and outcome feedback impact reported confidence? A 2-way ANOVA comparing average confidence between feedback and stimulus conditions showed a main effect of stimulus type on confidence, $F(1, 204) = 25.18$, $p < .001$, but no effect of feedback condition, $F(1,204) = 0.29$, $p = .59$. Looking closer at the effect of stimulus type, weight guessing items elicited greater confidence ($M = 61.42$, $SD = 18.88$) than coin flips ($M = 54.98$, $SD = 12.95$), $t(204) = 4.02$, $p < .001$, $d = 0.40$. However, there was no hindsight bias effect for either stimulus type (See Figure 7).

Did perceived randomness moderate hindsight? A mixed linear model tested the prediction that perceived randomness would attenuate the hindsight bias. Without the basic hindsight effect present in the evidence, it was unsurprising that the model showed no such attenuation by perceived randomness, b = -0.3, $t(204) = -0.30$, n.s.

Discussion

Experiment 7 tested the possible moderator of perceived randomness while administering another test of weight and coin stimuli in studying hindsight bias. Only a main effect of stimulus type predicted an increase confidence, and there was no evidence in support of the hindsight bias.

The failure to replicate hindsight bias in these domains could suggest some boundary condition of stimulus type for the phenomenon. Unfortunately, due to the lack of a main effect of outcome feedback on confidence, there is insufficient evidence in order to determine whether or not perceived randomness can attenuate the bias. This moderator is tested once more in Experiment 8, below, by manipulating perceptions of predictability for an entire domain of events.

CHAPTER 9

Experiment 8: Aleatory Perceptions and Hindsight

Experiment 8 tested an overarching theory that would predict which domains of outcomes are susceptible to hindsight bias, and which are not susceptible to such bias. This theory draws on a theorized mechanism known as creeping determinism (Fischhoff, 1975). Fischhoff surmised that, due to the hindsight bias process entailing automatic processing of outcome information, individuals would come to perceive an outcome as inevitable, manifested as having greater confidence.

Contextualized in Fischhoff's creeping determinism theory, hindsight bias arises when individuals perceive the outcome as 'inevitable', or pre-determined. Wasserman, Lempert, and Hastie (1991) posited that the presence of chance in determining an outcome leads to a lack of 'determinism' and therefore a lack of hindsight bias. I build on their explanation, and argue that hindsight bias fails not just when there is 'chance' present, but when an outcome can vary over time.

This 'temporal variance' in outcomes goes beyond the simple categories of chance outcomes vs. deterministic outcomes. Temporal variance pertains to domains in which the truth of certain domains can change over time, making them less predictable as a result. The time scale in question can vary in size. For instance, the answer to the question, "What is the most populous city on the planet?" has changed over time. However, the time scale is so large it is unlikely to impact how people answer the question. On a smaller time scale is the question posed in the preface for Experiment 4: "Who most recently won between the New York Yankees and the Boston Red Sox?" For this question, it is easier for individuals to remember instances in which the Yankees won, and in which the Red Sox won. The 'winner' of this match-up is quite variable and unpredictable across time, and knowing this fact, individuals may exhibit less confidence in their certainty of any one event in time.

The present study similar examines this theory with stimuli from differing ends of this spectrum. At the most extreme end, purely chance events such as coin flips are rarely regarded as pre-determined, as they are the most unpredictable. However, the weight of individuals does change slowly over time. This domain is more 'constant' over time than coin flips, and thus more predictable, but the weight of someone can still can vary relative to the general trivia questions present in most of the classic hindsight bias research (as the correct answer to a trivia question can stay constant through time.

Experiment 8 tests this theory by manipulating participants' perception of how 'predictable' an outcome is over time. I predicted that participants who are made to think of a domain of outcomes as highly unpredictable in nature will not present hindsight bias, while those participants that are influenced to see a category of outcomes as more constant and unchanging over time will in fact display hindsight bias.

Method

Participants

150 participants completed this experiment through Amazon Mechanical Turk for $1.00. The survey was advertised as a test of decision making ability. This study included 50 participants randomly assigned to each of three between-subjects conditions. This sample size was determined using the same rationale as in my previous studies.

Design

The experiment had a 3 (high predictability vs. control vs. low predictability) cell design that manipulated participants' perceptions of the predictability of basketball games. The primary dependent variable was confidence in their choice of the winning team for each game. The survey also collected hit rate data.

Procedure and Materials

Overview. The Qualtrics survey randomly assigned participants to either the high predictability, control, or low predictability condition. Participants first read an authoritative vignette designed to influence their perception regarding the predictability of basketball games. The survey forced participants to wait at least 20 seconds before proceeding past the vignette in order to encourage close reading by the participants.

The Qualtrics survey program then presented the series of 30 basketball games to each participant in a randomized order. All participants received outcome feedback for each game by clearly identifying the winning team for each. These stimuli were unique, but of a style very similar to previous experiments.

To view the online questionnaire as a participant, follow this link:
https://berkeley.ca1.qualtrics.com/jfe/preview/SV_4YGzeXMsRJYEWQB?Q_SurveyVersionID=current&Q_CHL=preview

Predictability Manipulation. In the high predictability condition, participants read a prompt that described basketball game results as being highly determined by the skill of the players on each team, and that research shows that few random events actually have a significant effect on the final score:

> New research released by the National Basketball Association in collaboration with the University of Rochester and Purdue University has determined that **basketball game results are largely determined by player skill, a force very much within players' control.**
>
> This new research program sheds light on NBA results, as the evidence shows that effects of other forces outside of player control, such as luck, player mood, game attendance, and time of game, have less impact on the final score than previously believed.

> This research clears up former controversy around the potential factors that impact the final score of a game, pointing to player skill, within players' control, as a large factor in determining the winner.

In the low predictability condition, participants read a prompt that describes basketball scores as highly influenced by random events:

> New research released by the National Basketball Association in collaboration with the University of Rochester and Purdue University has determined that **basketball game results are largely determined by various forces outside of the control of any player, such as player mood, game attendance, time of game, and luck.**
>
> This new research program sheds light on NBA results, as the evidence shows that effects of other forces within players' control, such as player skill, have less impact on the final score than previously believed.
>
> This research clears up former controversy around the potential factors that impact the final score of a game, pointing to these various forces outside the control of any player as large factors in determining the winner.

In the control condition, participants did not read anything prior to reviewing the 30 basketball games.

Manipulation Pilot. One hundred participants pilot-tested whether the manipulations changed perceptions of randomness of the sports outcomes. The pilot evidence showed that participants with the high predictability stimulus believed basketball to be much more predictable (M = 4.57, SD = 1.62) compared to those with the low predictability stimulus (M = 3.33, SD = 1.51), t(98) = -3.93, p < .001, d = -0.79. Additionally, participants with the high predictability prompt believed basketball was less impacted by random chance (M = 2.96, SD = 1.66), compared to those with the low predictability prompt (M = 4.31, SD = 1.78), t(98) = 3.90, p < .001, d = 0.78.

I included a multi-item basketball expertise measure similar to the expertise measure used in Experiment 6, and planned to control for sports expertise in my analyses. I expected an attenuated effect of this manipulation on well-informed basketball fans.

Confidence. As in previous experiments, the main dependent variable was confidence level for each item, on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*.

Results

Did the predictability manipulation or basketball expertise impact confidence reports? A two-way ANOVA compared average confidence by condition showed no effect of predictability manipulation on confidence, F(2, 149) = .02, p = .98. Conditional means showed that average confidence and therefore hindsight bias was similar between low predictability (M = 65.01, SD =

14.90), control (*M* = 65.27, *SD* = 16.41), and high predictability conditions (*M* = 64.63, *SD* = 16.86). I also employed an Analysis of Covariance (ANCOVA) which used basketball expertise and variability condition as predictors of confidence. This test revealed no significant main effects for condition, *F*(2, 146) = 0.02, *p* = .98, nor for expertise, *F*(1, 146) = 0.21, *p* = .65. There was also a marginal yet nonsignificant interaction effect between the two, *F*(2, 146) = 2.94, *p* = .06. Neither expertise nor the predictability manipulation impacted reported confidence among participants (See Figure 8).

Discussion

Experiment 8 sought to find a possible mediator that would explain why some stimuli succumb to the hindsight bias while others do not. Manipulating participant perceptions of variability failed to elicit systematic differences in reported confidence. While pilot testing showed that the manipulations did impact participant perceptions of predictability and randomness in predicting basketball scores, that was not enough to translate to differences in confidence.

There are multiple possible reasons for this null result. One is that the manipulations were too weak to elicit perceptual changes which were strong enough to impact confidence judgments. Another possibility is that the manipulations could have successfully elicited perceptual changes, but the changes were fleeting did not persist through the task. These results provide no supportive evidence that perceptions of variability mediate the hindsight bias. In light of the many failures to replicate, I next sought to contextualize these null results with what experts in the field believe regarding the bias in Experiment 9, following by a close replication of Fischhoff (1977) using the same trivia stimuli in Experiment 10.

CHAPTER 10

Experiment 9: Lay Beliefs of Hindsight

Experiment 9 examines expert beliefs of the hindsight bias, in order to contextualize the significance of the null results present in Experiments 1 – 4. Those initial experiments showed a failure of hindsight bias for various types of stimuli, from random chance events, to sporting events, and even general knowledge trivia. Is this surprising to current scholars in the field?

Due to the common presence of the bias in common vernacular and the fame of the original findings, I predicted that judgment and decision making scholars would overestimate both the breadth of domains susceptible to the hindsight bias, as well as the degree to which people err due to the hindsight bias. Evidence of such beliefs would highlight the significance of the failures to replicate hindsight bias in multiple domains, as the documented failures to replicate the effect would not only be counterintuitive to established literature, but also to the views of the population at large.

Method

Participants

The final sample included 117 expert participants and 103 ‘actor’ participants for 220 participants total. The actor participants completed this experiment through Amazon Mechanical Turk for $1.00. The survey was advertised as a survey about people’s beliefs about predicting events. I originally planned for 100 actor participants randomly assigned to two between-subjects conditions for 50 participants per cell. This sample size was determined using the same rationale as in my previous studies.

I recruited the expert participants from the Society for Judgment and Decision Making listserve, creating a sample of PhD students and faculty who are experts in social psychology and decision making. I did not know the variance in this population concerning this question, so I collected as many participants as possible in the hopes of obtaining a sufficiently representative sample.

Design

This experiment combined two separate data sets. For the actor data set, the experiment randomly assigned participants to one of two between-subjects conditions. These two conditions were near-replications of Experiment 4. For the expert data set, the survey asked respondents for their beliefs about hindsight in various domains.

Procedure and Materials

Overview. The procedure for the actor data set was nearly identical to that of Experiment 4. Participants answered ten questions while reporting their confidence for 40 items over four different domains: trivia, sporting events, weight guessing, and coin flips. The Qualtrics survey randomly assigned participants in the actor data set to either a control condition or an outcome feedback (hindsight) condition, identical to Experiment 4. Participants in the actor data set responded to the same coin flip and weight guessing stimuli as Experiment 4, and similar but unique basketball game stimuli. The trivia stimuli were a subset of the stimuli used for Experiment 10, being drawn from the same pool of questions used in the original hindsight bias studies (Fischhoff, 1977).

To view the actor data online questionnaire as a participant, follow this link: https://berkeley.ca1.qualtrics.com/jfe/preview/SV_bgfweTmnZZzmpbn?Q_SurveyVersionID=current&Q_CHL=preview

Hindsight Manipulation. The actor data survey manipulated hindsight with the outcome feedback paradigm used in previous experiments. Participants in the hindsight condition answered the same forty questions as those in the control condition, but first saw the correct answer or outcome for each question prior to answering.

Expert Data Set. Participants reported their beliefs as to the effect of hindsight in each of the four tested domains. For each domain, the survey provided participants with a hypothetical example of each domain as an attempt to standardize participant understanding of the task. This is the example for general knowledge trivia:

> *This is an example item for the 'general knowledge trivia' domain:*
>
> What is the capital of France?
>
> A) Paris
> B) Rome
>
> How confident are you in your answer?
> 0 ---100

To view the expert data online questionnaire as a participant, follow this link: https://berkeley.ca1.qualtrics.com/jfe/preview/SV_8x0koPXi5C0nZNH?Q_SurveyVersionID=current&Q_CHL=preview

Actor Confidence. The main dependent variable was confidence level for each item, on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*. I calculated the difference in mean confidence for each stimulus type by subtracting mean confidence in the control condition from the mean confidence in the hindsight condition.

Expert Belief in Hindsight. Participants reported the difference in mean confidence (between -100 and 100) between the hindsight and control conditions within the actor data set. They provided four total belief measures, one for each stimulus type.

Results

How did experts view the strength of hindsight bias in each of the four domains? A one-way ANOVA compared the mean predicted hindsight strength for all domains. To recall, participants predicted what they believed would be the average difference in reported confidence between a control condition and an outcome feedback (hindsight) condition. The test showed a significant difference in believed hindsight strength among the domains, *F*(3, 464) = 8.96, *p* < .001 (See Figure 9).

Post-hoc comparisons using the Tukey HSD test showed that experts believed that weight guessing questions (*M* = 23.12, *SD* = 17.83) and sports questions (*M* = 24.08, *SD* = 23.21) would elicit the greatest difference in confidence between conditions, and trivia questions (*M* = 14.28, *SD* = 18.18) and coin flip questions (*M* = 12.92, *SD* = 24.07) would exhibit less hindsight bias (however, still a sizable effect). Pairwise comparisons showed that weight and sports questions were not significantly different, *adjusted p* = .99, nor were trivia and coin questions different, *adjusted p* = .96. However, experts believed sports and weight questions elicited more hindsight than either coin flip questions or trivia questions, respectively, *adjusted p's < .008, d's* > .49.

How did expert predictions compare to experimental results? Four one-sample t-tests showed that the experts overestimated the impact of the hindsight bias compared to its actual observed effect in the parallel study, *t*'s < -2.70, *p*'s < .01. In other words, expert predictions of the difference in confidence between hindsight and control conditions was much greater than in reality for all four domains.

Discussion

Experiment 9 is important in highlighting the counterintuitive nature of the repeated failures to replicate hindsight bias throughout this research. Not only is the power of the hindsight bias consistently shown in the literature, but it is also widely believed to be true among experts in the field. This result emphasizes the importance of the present work, as these findings not only add to the existing literature, but also could make strides to update a widely held belief within the field. There could be stimulus boundary conditions for hindsight which researchers have not yet directly examined. I will discuss this further in the general discussion.

CHAPTER 11

Experiment 10: Hindsight with General Knowledge Trivia

.

Experiments 1 – 8 all failed to find any evidence in support of the classic hindsight bias effect. Experiment 10 provides one more attempt at finding evidence for the bias with a close replication of one of the original hindsight bias studies. This experiment also tests for the moderator of question difficulty to explain the repeated failures to replicate hindsight bias with general knowledge trivia questions. The lack of evidence for hindsight bias with trivia stimuli constitutes the most significant failure to replicate the effect within this research program, as the majority of the literature establishing the hindsight bias and its robustness uses general knowledge trivia as the stimulus.

In one condition I use the identical stimuli from foundational studies to serve as a close replication, provided by Baruch Fischhoff and Decision Research at the University of Oregon (Fischhoff, 1977). I expect to find hindsight bias through outcome feedback with the original stimuli. A second set of more difficult questions tests an interaction effect between question difficulty and outcome feedback, in which I expect difficult questions to attenuate the hindsight bias.

Method

Participants

The final sample included 207 participants. I originally planned for two hundred participants determined using the same rationale as in my previous studies. Participants completed this experiment through the Haas Research Participation Pool (RPP) in the lab for course credit or for $7.00. The survey was advertised as a survey about decision making ability.

Design

The experiment contains a 2 (outcome feedback vs. control) X 2 (original trivia vs. hard trivia) cell design that manipulated question difficulty and whether or not participants receive the actual trivia answer prior to making their outcome estimate and their confidence measure. The dependent variable was the confidence in their chosen answer out of the two choices provided. Finally, hit rates were also collected.

Procedure and Materials

Overview. The Qualtrics survey randomly assigned participants to either the outcome feedback condition or the control condition. Additionally, the survey randomly assigned participants to either review the original trivia questions from Fischhoff (1977), or a set of difficult trivia questions pre-tested for a low hit-rate. The Qualtrics survey program then presented the series of stimulus items to each participant in a randomized order.

To view the online questionnaire as a participant, follow this link (same link as Experiment 5):

https://berkeley.ca1.qualtrics.com/jfe/preview/SV_25iQkX8PRUoxR2J?Q_SurveyVersionID=current&Q_CHL=preview

Hindsight Manipulation. Identical to previous experiments, participants in the outcome feedback condition saw the correct outcome or answer prior to providing their own answer and confidence level.

Confidence. As in previous experiments, the main dependent variable was confidence level for each item, on a scale from 0 = *Extremely Not Confident* to 100 = *Extremely Confident*.

Results

Concerns regarding data quality led to the decision to conservatively remove ten outlier participants from the data set, as their mean confidence across 25 items was less than ten. This resulted in a final sample size of 192.

Did providing feedback and question set impact confidence reports? A 2-way ANOVA predicting confidence from feedback condition and difficulty condition to showed no main effect for feedback condition *F*(1,191) = 2.97, *p* = .09, a non-significant main effect for question difficulty, *F*(1,191) = 1.15, *p* = .28, and a significant interaction effect, *F*(1,203) = 119, *p* < .001 (See Figure 10).

Exploring the interaction further by subsetting the data on question difficulty, the test showed that for difficult questions, participants reported less confidence in the outcome feedback condition (*M* = 38.02, *SD* = 16.31) compared to the control condition (*M*=55.45, *SD* = 17.01), *t*(95) = 6.42, *p* < .001, *d* = 1.05. For the original Fischhoff questions, participants reported greater confidence in the outcome feedback condition (*M* = 56.24, *SD* = 14.24) compared to the control condition (*M* = 30.68, *SD* = 19.10), *t*(95) = -8.88, *p* < .001, *d* = -1.52.

Puzzlingly, exploratory analyses show that the difficult question condition hit rate *(M* = .52, *SD* = 0.13) was actually not more difficult than the original Fischhoff questions hit rate *(M* = .53, *SD* = 0.13), *t*(202) = 0.16, *p* = 0.56, *d* = 0.08.

Discussion

Experiment 10 tested the moderator of question difficulty that could explain the failure to replicate hindsight bias with general knowledge trivia in earlier experiments. This study did in fact replicate the hindsight bias effect, but the story is not so simple, as the ‘difficult’ questions were not actually more difficult. Using the original trivia questions employed in the landmark hindsight bias experiments I did replicate phenomenon. The ‘difficulty’ stimuli, however, found the opposite effect, in which outcome feedback decreased confidence reports compared to control. There is precedent for this latter result, as Hoch and Loewenstein (1989) found the highly difficult questions can elicit a ‘never would have known it’ effect. The novel contribution from this study, merely by serendipity, is that the ‘difficult’ stimuli were in reality not more difficult, though they perhaps *appeared* to be more difficult in the control condition (conversely, the Fischhoff stimuli could have *appeared* to be easier, leading to the hindsight bias).

The results show that a close replication of the original hindsight bias effect with the original stimuli did succeed in recreating the hindsight bias. The supposedly ‘difficult’ set of questions did show the opposite of the hindsight bias effect. The hit rates of the two question sets were not significantly different, however.

Post-hoc analyses show that the hit rates between the two question-type conditions were similar, despite showing different patterns of hindsight bias. With similar hit rates, there is not sufficient evidence to weigh in on whether actual question difficulty moderates the hindsight bias, though the relationship has been suggested previously (see Hoch & Loewenstein, 1989). There may exist some perceptual or psychological process which leads to hindsight bias, activated by the difficult questions *appearing* more difficult at face value, versus how Fischhoff’s original stimuli appear.

CHAPTER 12

Discussion and Future Directions

Summary of Results

The results of these studies suggest that the hindsight bias is not nearly as ubiquitous as the literature implies, nor as experts in the field believe. These experiments examine various stimulus domains and possible moderators which could have explained early failures to replicate the bias, though all came up short. To be precise, I did not find evidence to suggest my potential moderators failed to attenuate hindsight, but rather hindsight failed to materialize in the first place, leaving nothing to debias or moderate. Experiment 10 is notable in that it is the only study in this package to produce clear evidence of the hindsight bias, and there only in one condition, which was a direct replication of Fischhoff (1977. Participants did not exhibit less accurate forecasts after seeing outcome feedback (Experiments 1 and 2). Experiments 3 and 4 tested for hindsight across a variety of domains and found no evidence of the phenomenon.

Experiment 5 provided weak evidence of hindsight bias using a within-subjects longitudinal design. However, Experiment 6 which tested hindsight in the context of the 2018 U.S. midterm elections failed to elicit the bias. Experiments 7 and 8 explored Fischhoff's original proposed mechanism of creeping determinism, but found that perceptions of randomness did not impact participant susceptibility to the hindsight bias. Participants also did not respond with hindsight to a manipulation which was designed and piloted to elicit feelings of predictability.

Experiment 9 asked scholars for their predictions regarding the power of the hindsight bias, and found that they overestimated the strength of hindsight across several domains. Finally, Experiment 10 successfully replicated hindsight bias using the same trivia stimuli used in one of the original studies on the topic (Fischhoff, 1977).

Theoretical Implications

The hindsight bias is a household phrase not for any historical accident. Rather, it is a household concept precisely because people believe it to be widespread and pernicious. The general populace cannot be blamed too much in this regard, as the initial studies into the bias were quite consistent in their agreement as to the robustness and replicability of hindsight bias (e.g. Fischhoff, 1975; Fischhoff, 1977; Wood, 1978). My initial direction of research for this program was in debiasing probabilistic forecasting. To this end, I had assumed the hindsight bias was one of the solid, time-tested 'truths' of psychology. As I have previously described, my repeated failures to replicate the basic effect with different stimuli led me on a new path in searching for the limits of this bias.

What would it mean to find boundary conditions on the hindsight bias? I would argue that such a discovery does not undermine the legitimacy of the bias overall. Drawing such nuance is a natural consequence as our understanding of human decision making processes develops and refines through continued research.

Such a conclusion offers both positive and possibly negative implications. On the positive side, humans may not be so dumb and feeble-minded after all. As the field of judgment and decision making has matured and gained fame, the popular reaction to the common theme of mistake-prone decision making has been mixed at best. An American philosopher once responded to Nobel laureate Daniel Kahneman, in reference to Kahneman's book *Thinking Fast and Slow*, "I am not really interested in the psychology of stupidity" (Burkeman, 2011). Perhaps researchers will discover more contexts for which well-known biases are not as prevalent as once believed, and therefore do *not* lead people astray nearly as often as one might fear. With such a result, the pessimistic reputation of the field could perhaps recover somewhat.

On the other hand, one must wonder about the extent to which old, foundational papers in the field of decision making are in need of fundamental updating. The fate of hindsight bias could merely be indicative of a larger pattern in the field, in which seemingly arbitrary design decisions lead to an oversimplified or myopic examination of phenomena. Indeed, Hawkins and Hastie (1990) note in their review of hindsight bias literature that almost all stimuli used happened to be 'almanac question materials', without a justification as to why that is A) the correct stimuli to use, or B) the only stimuli used in most studies. The tradition of borrowing stimuli in research, though a practical method of achieving legitimacy in one's design, may prove to be a large factor in future failure to replicate going forward.

Reconciling My Results with the Literature

To what degree am I truly undermining the massive amount of research findings on the hindsight bias? The hindsight bias is a well-known decision making bias not only within the field and among scholars, but among the general populace as well. Nine failures to replicate the bias will not destroy the legitimacy of an entire body of published literature. Further, my earlier *p*-curve suggests that the evidence within this literature is not *p*-hacked. What else could explain my consistent null results, in addition to the strong replication in Experiment 10?

I believe my repeated failures to replicate the hindsight bias is primarily due to stimulus selection. My theory is that the same set of general knowledge trivia stimuli which defined the field for years (Hawkins & Hastie, 1990) was particularly effective in eliciting the hindsight bias. It is this same set which I use in Experiment 10 to also find a significant hindsight bias result. When I utilized the same outcome feedback paradigm present in several early studies (e.g. Fischhoff, 1977; Wood, 1978; Hoch & Loewenstein, 1989) with stimuli *other* than the original set of trivia, I failed to replicate the bias. My failure to replicate in Experiment 6 using a longitudinal paradigm for the midterm elections does not lend itself to such a simple explanation. I surmise that the nature of the answer being a continuous variable (in the number of congressional seats won by each party) attenuated the bias, however I have no evidence to support this conjecture.

To test this theory, I categorized the studies in my original *p*-curve by whether they used the original set of general knowledge trivia or not (again, see Exhibit 2 for the disclosure table). I then created two separate *p*-curves; one for hindsight studies with trivia stimuli, and one with any other type of stimuli (see Exhibits 3 and 4). These *p*-curves provide supporting evidence for my explanation. As Exhibit 3 shows, the hindsight effects present in studies which used the trivia questions are large and thus well-powered. However, the effects present in the *p*-curve in Exhibit 4 are quite underpowered, and contain results of weak evidentiary value. While this latter *p*-curve does not suggest rampant *p*-hacking in the literature without trivia stimuli, it does suggest that the hindsight bias is a weaker and perhaps less prevalent bias than previously believed, as without the original trivia stimuli the evidence is weaker and the effects are smaller or less-powered.

Finally, I must acknowledge that this dissertation only persisted for this many studies due to the recent culture in the field brought about by the open science movement. As recently as ten years ago, failing to replicate a high-profile theory in social psychology would have sent me back to the drawing board a long time ago. With more sophisticated knowledge of statistics and research practices combined with a new humility of the science, a culture has arisen that is more open-minded and accepting of this kind of work.

Conclusion

The subject of biases and decision making is of the utmost importance to both researchers and business leaders. Large amounts of time and resources are spent in the pursuit of debiasing decision making and in leading people to be more accurate in their judgments of the world around them. However, what if biases once thought powerful and robust are not so universal after all? Oft-cited papers in this field have aged without significant challenges to their authority, and even simple checks for replication have the ability to go astray and call into question theories once believed as truth. On a broader level beyond just hindsight bias, the current rising trajectory of open science and the growing acceptance of the importance of replicability in social science will prove to be invaluable in the future as researchers strive to not make the same mistakes of the past.

References

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the Hindsight Bias. *Journal of Applied Psychology*, *73*(2), 305–307. https://doi.org/10.1037/0021-9010.73.2.305

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*(2), 252.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569–579. https://doi.org/10.1037/0022-3514.54.4.569

Blank, H., & Nestler, S. (2007). Cognitive process models of hindsight bias. *Social Cognition*, *25*(1), 132–146.

Brier, G. W. (1950). The statistical theory of turbulence and the problem of diffusion in the atmosphere. *Journal of Meteorology*, *7*(4), 283–290.

Buchman, T. A. (1985). An effect of hindsight on predicting bankruptcy with accounting information. *Accounting, Organizations and Society*, *10*(3), 267–285. https://doi.org/10.1016/0361-3682(85)90020-0

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C., & Loewenstein, G. (1989). The Curse of Knowledge in Economic Settings : An Experimental Analysis Published by : The Universit. *Journal of Political Economy*, *97*(5), 1232–1254.

Campbell, J. D., & Tesser, A. (1983). Motivational interpretations of hindsight bias: An individual difference analysis. *Journal of Personality*, *51*(4), 605–620. https://doi.org/10.1111/j.1467-6494.1983.tb00868.x

Christensen-Szalanski, J. J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *48*(1), 147–168. https://doi.org/10.1016/0749-5978(91)90010-Q

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300.

Davies, M. F. (1987). Reduction of hindsight bias by restoration of foresight perspective: Effectiveness of foresight-encoding and hindsight-retrieval strategies. *Organizational Behavior and Human Decision Processes*, *40*(1), 50–68. https://doi.org/10.1016/0749-

5978(87)90005-7

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Judgment and Decision Making*, *1*(3), 33–49. https://doi.org/10.4324/9780203141939

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(2), 349.

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

Fischoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(2), 349–358. https://doi.org/10.1037/0096-1523.3.2.349

Guilbault, R., Bryant, F., Brockway, J. H., & Posavac, E. (2004). A Meta-Analysis of Research on Hindsight Bias. *Basic and Applied Social Psychology*, *26*(2), 103–117. https://doi.org/10.1207/s15324834basp2602&3_1

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107*(3), 311–327. https://doi.org/10.1037/0033-2909.107.3.311

Hell, W., Gigerenzer, G., Gauggel, S., Mall, M., & Müller, M. (1988). Hindsight bias: An interaction of automatic and motivational factors? *Memory & Cognition*, *16*(6), 533–538. https://doi.org/10.3758/BF03197054

Hertwig, R., Fanselow, C., & Hoffrage, U. (2003). Hindsight bias: How knowledge and heuristics affect our reconstruction of the past. *Memory*, *11*(4–5), 357–377. https://doi.org/10.1080/09658210244000595

Hirt, E. R., Kardes, F. R., & Markman, K. D. (2004). Activating a mental simulation mind-set through generation of alternatives: Implications for debiasing in related and unrelated domains. *Journal of Experimental Social Psychology, 40*(3), 374–383.

Hoch, S. J., & Loewenstein, G. F. (1989). Outcome feedback: Hindsight and information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 605–619. https://doi.org/10.1037/0278-7393.15.4.605

Hölzl, E., Kirchler, E., & Rodler, C. (2002). Hindsight bias in economic expectations: I knew all along what I want to hear. *Journal of Applied Psychology, 87*(3), 437.

Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 11–22.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*(6), 881–919.

Kelly, T. F., & Simmons, J. P. (2016). When does making detailed predictions make predictions worse? *Journal of Experimental Psychology: General*, *145*(10), 1298–1311. https://doi.org/10.1037/xge0000204

King, L. A., Hicks, J. A., Krull, J. L., & Del Gaiso, A. K. (2006). Positive affect and the experience of meaning in life. *Journal of Personality and Social Psychology*, *90*(1), 179.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Kruglanski, A. W. (1989). The psychology of being" right": The problem of accuracy in social perception and cognition. *Psychological Bulletin*, *106*(3), 395.

Leary, M. R. (1981). The distorted nature of hindsight. *The Journal of Social Psychology*.

Leary, M. R. (1982). Hindsight distortion and the 1980 presidential election. *Personality and Social Psychology Bulletin*, *8*(2), 257–263.

Louie, T. A., Curren, M. T., & Harich, K. R. (2000). “ I knew we would win”: Hindsight bias for favorable and unfavorable team decision outcomes. *Journal of Applied Psychology*, *85*(2), 264.

Mitchell, T. R., & Kalb, L. S. (1981). Effects of outcome knowledge and outcome valence on supervisors’ evaluations. *Journal of Applied Psychology*, *66*(5), 604–612. https://doi.org/10.1037/0021-9010.66.5.604

Musch, J., & Wagner, T. (2007). Did everybody know it all along? A review of individual differences in hindsight bias. *Social Cognition*, *25*(1), 64–82.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401

Pennington, D. C. (1981). The British firemen’s strike of 1977/78: An investigation of judgements in foresight and hindsight. *British Journal of Social Psychology*, *20*(2), 89–96. https://doi.org/10.1111/j.2044-8309.1981.tb00479.x

Pezzo, M. V, & Beckstead, J. W. (2008). The effects of disappointment on hindsight bias for real-world outcomes. *Applied Cognitive Psychology: The Official Journal of the Society for*

Applied Research in Memory and Cognition, *22*(4), 491–506.

Powell, J. L. (1988). A Test of the Knew-It-All-Along Effect in the 1984 Presidential and Statewide Elections 1. *Journal of Applied Social Psychology*, *18*(9), 760–773.

Roese, N. J., & Vohs, K. D. (2012). Hindsight Bias. *Perspectives on Psychological Science*, *7*(5), 411–426. https://doi.org/10.1177/1745691612454303

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, *32*(5), 880.

Sanna, L. J., Schwarz, N., & Stocker, S. L. (2002). When Debiasing Backfires: Accessible Content and Accessibility Experiences in Debiasing Hindsight. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(3), 497–502. https://doi.org/10.1037/0278-7393.28.3.497

Schul, Y., & Burnstein, E. (1985). When discounting fails: Conditions under which individuals use discredited information in making a judgment. *Journal of Personality and Social Psychology*, *49*(4), 894.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, *1*(1), 43–61.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, *22*(11), 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 544–551. https://doi.org/10.1037/0096-1523.3.4.544

Sue, S., Smith, R. E., & Caldwell, C. (1973). Effects of inadmissible evidence on the decisions of simulated jurors: A moral dilemma. *Journal of Applied Social Psychology*, *3*(4), 345–353.

Synodinos, N. E. (1986). Hindsight Distortion: “I knew-it-all along and I was sure about it.” *Journal of Applied Social Psychology*, *16*(2), 107–117. https://doi.org/10.1111/j.1559-1816.1986.tb02282.x

Thompson, W. C., Fong, G. T., & Rosenhan, D. L. (1981). Inadmissible evidence and juror verdicts. *Journal of Personality and Social Psychology*, *40*(3), 453.

Trabasso, T., & Bartolone, J. (2003). Story understanding and counterfactual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 904.

Tykocinski, O. E. (2001). I never had a chance: Using hindsight tactics to mitigate disappointments. *Personality and Social Psychology Bulletin*, *27*(3), 376–382.

Verplanken, B., & Pieters, R. G. M. (1988). Individual differences in reverse hindsight bias: I never thought something like Chernobyl would happen. Did I? *Journal of Behavioral Decision Making*, *1*(3), 131–147.

Walster, E. (1967). 'Second Guessing'Important Events. *Human Relations*, *20*(3), 239–249.

Wann, D. L., Grieve, F. G., Waddill, P. J., & Martin, J. (2008). Use of retroactive pessimism as a method of coping with identity threat: The impact of group identification. *Group Processes & Intergroup Relations*, *11*(4), 439–450.

Wasserman, D., Lempert, R., & Hastie, R. (1991). Hindsight and Causality. *Personality & Social Psychology Bulletin*, *17*(1), 30–35.

Werner, C. M., Kagehiro, D. K., & Strube, M. J. (1982). Conviction proneness and the authoritarian juror: Inability to disregard information or attitudinal bias? *Journal of Applied Psychology*, *67*(5), 629.

Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 345–353. https://doi.org/10.1037/0096-1523.4.2.345

Wyer, R. S., & Unverzagt, W. H. (1985). Effects of instructions to disregard information on its subsequent recall and use in making judgments. *Journal of Personality and Social Psychology*, *48*(3), 533.

Yopchick, J. E., & Kim, N. S. (2012). Hindsight bias and causal reasoning: A minimalist approach. *Cognitive Processing*, *13*(1), 63–72.

APPENDICES

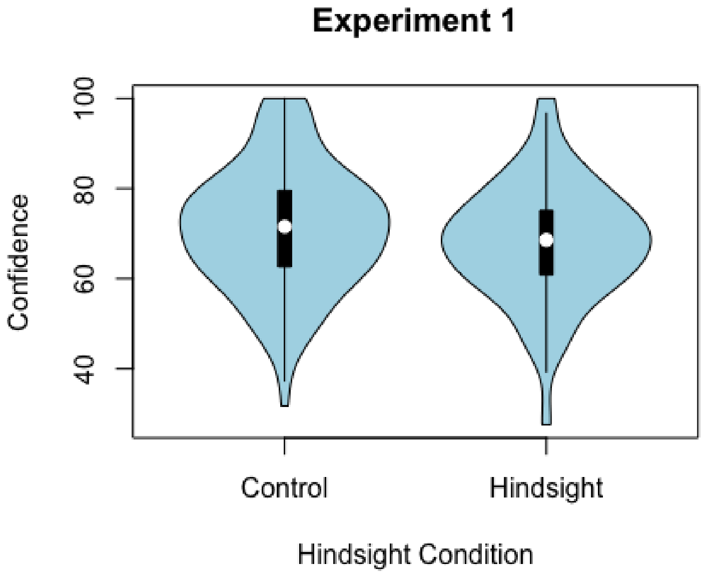


Figure 1. Violin plots, showing confidence as a function of hindsight condition, Experiment 1. The white dot indicates each condition's mean. The dark bar shows the interquartile range.

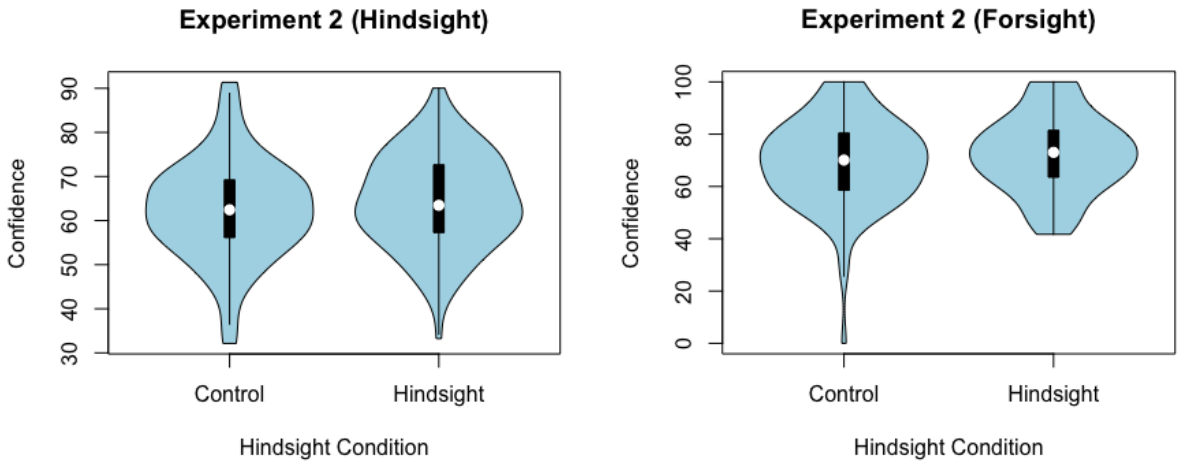


Figure 2. Violin plots, showing confidence as a function of hindsight condition, for review items (left) and forecast items (right), Experiment 2. The white dot indicates each condition's mean. The dark bar shows the interquartile range.

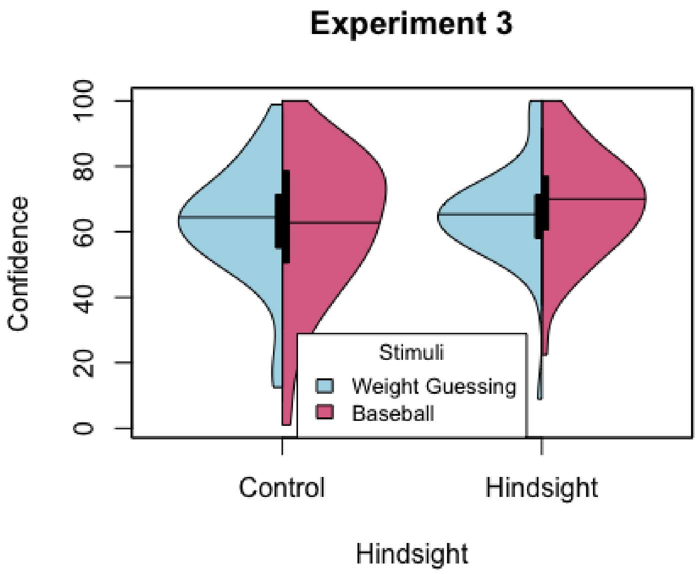


Figure 3. Violin plots, showing confidence as a function of hindsight condition (on the x-axis) and stimulus type (denoted by color), Experiment 3. The dark bars show the interquartile range.

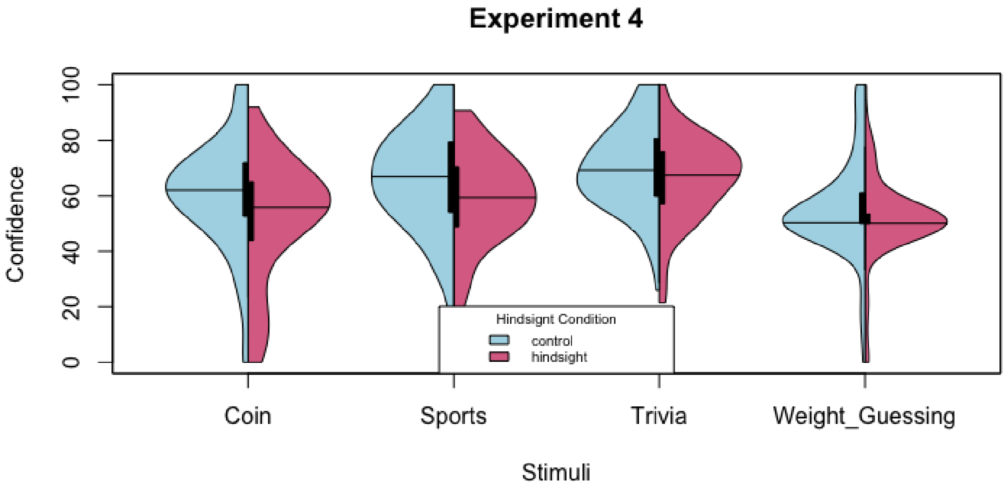


Figure 4. Violin plots, showing confidence as a function of hindsight condition (denoted by color) and stimulus type (on the x-axis), Experiment 4. The dark bars show the interquartile range.

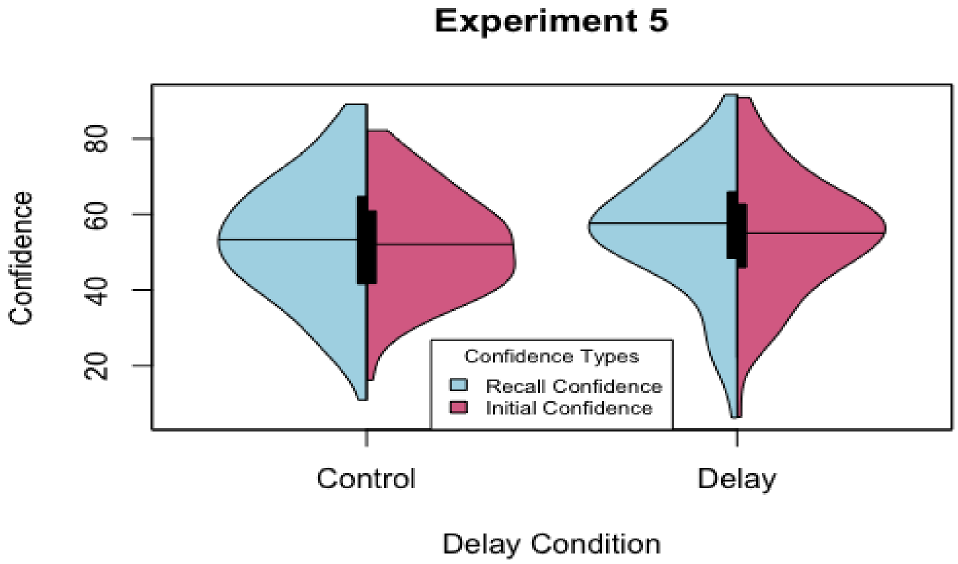


Figure 5. Violin plots, showing confidence as a function of delay condition (on the x-axis) and confidence type (denoted by color), Experiment 5. The dark bars show the interquartile range.

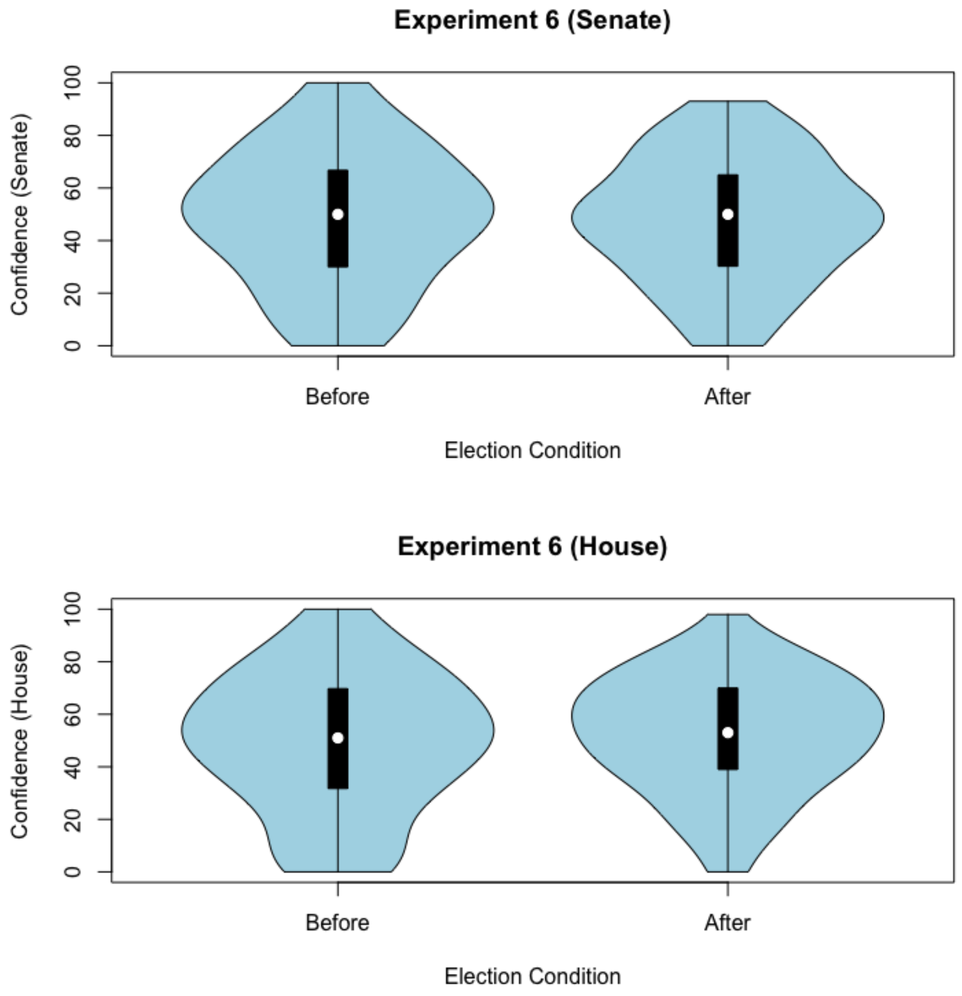


Figure 6. Violin plots, showing confidence as a function of data collection timing (on the x-axis), for the Senate (above) and the House of Representatives (below), Experiment 6. The white dot indicates each condition's mean. The dark bar shows the interquartile range.

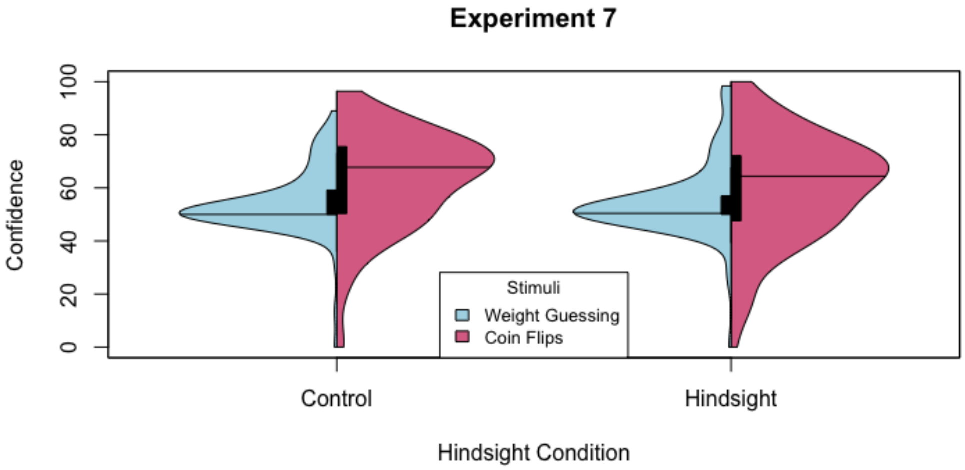


Figure 7. Violin plots, showing confidence as a function of hindsight condition (on the x-axis) and stimulus type (denoted by color), Experiment 7. The dark bars show the interquartile range.

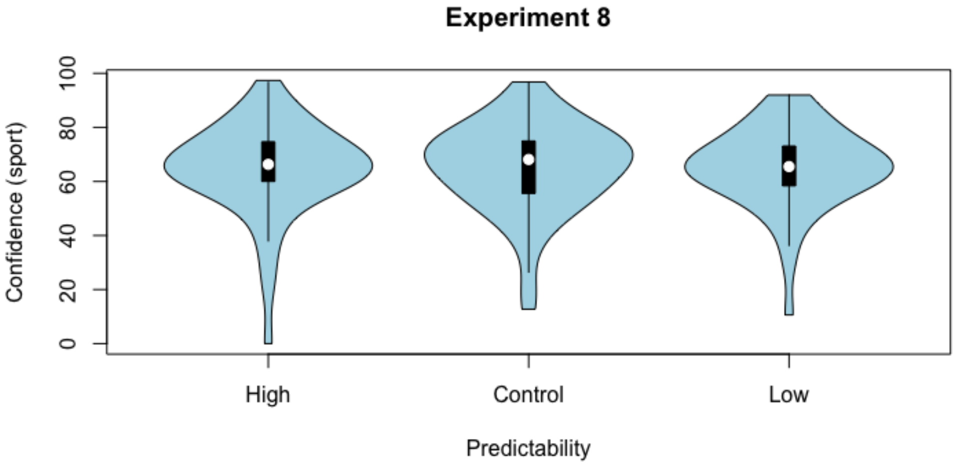


Figure 8. Violin plots, showing confidence as a function of predictability condition, Experiment 8. The white dot indicates each condition's mean. The dark bar shows the interquartile range.

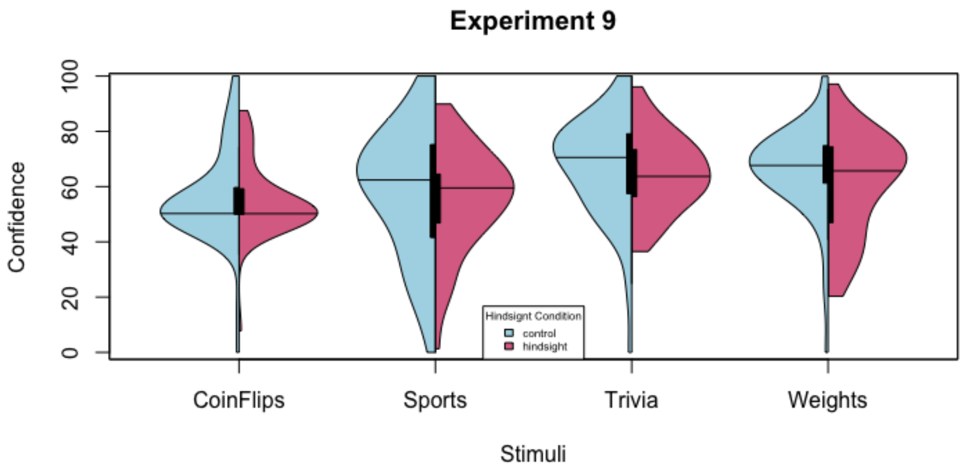


Figure 9. Violin plots, showing confidence as a function of hindsight condition (denoted by color) and stimulus type (on the x-axis), Experiment 9. The dark bars show the interquartile range.

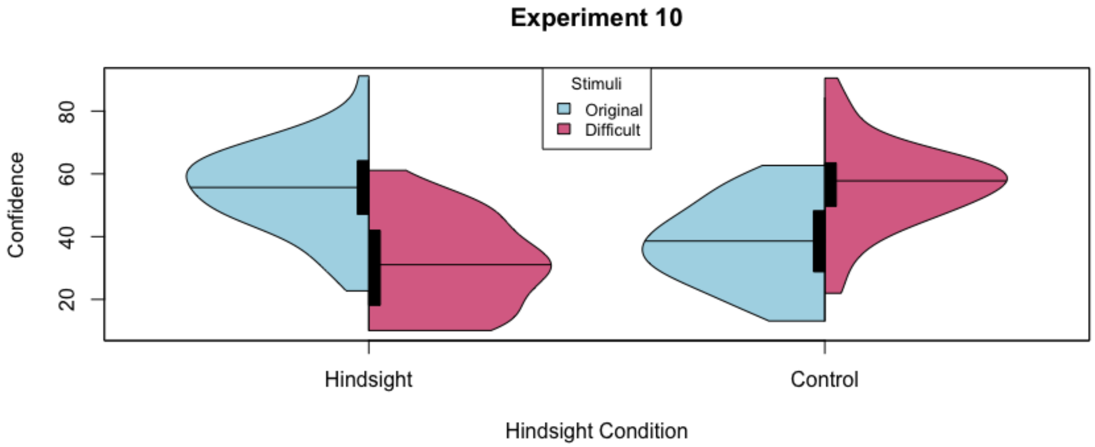
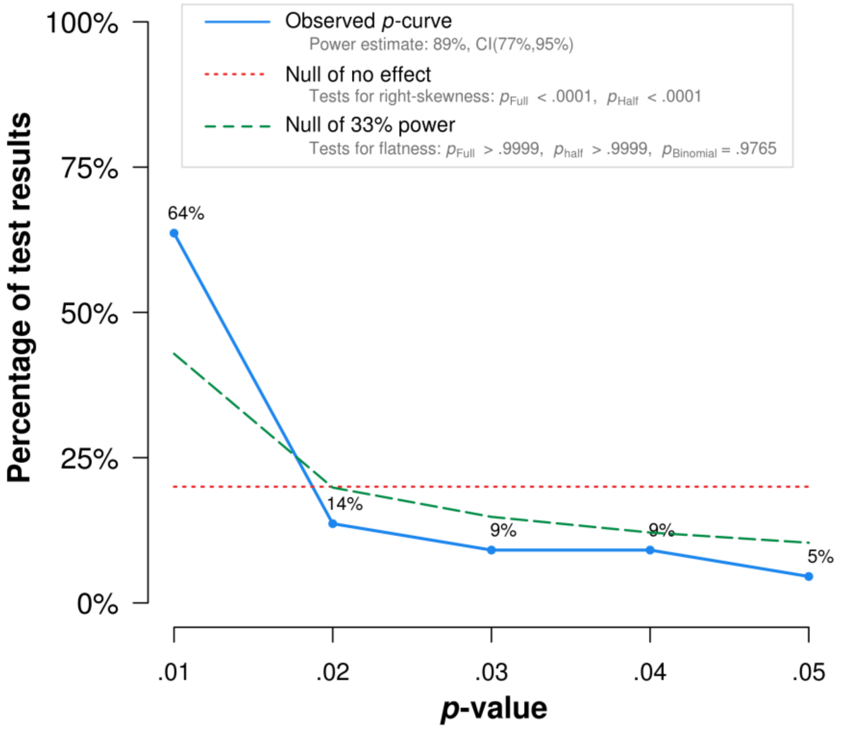


Figure 10. Violin plots, showing confidence as a function of hindsight condition (on the x-axis) and stimulus set (denoted by color), Experiment 10. The dark bars show the interquartile range.



Note: The observed p-curve includes 22 statistically significant (p < .05) results, of which 19 are p < .025.
There were 2 additional results entered but excluded from p-curve because they were p > .05.

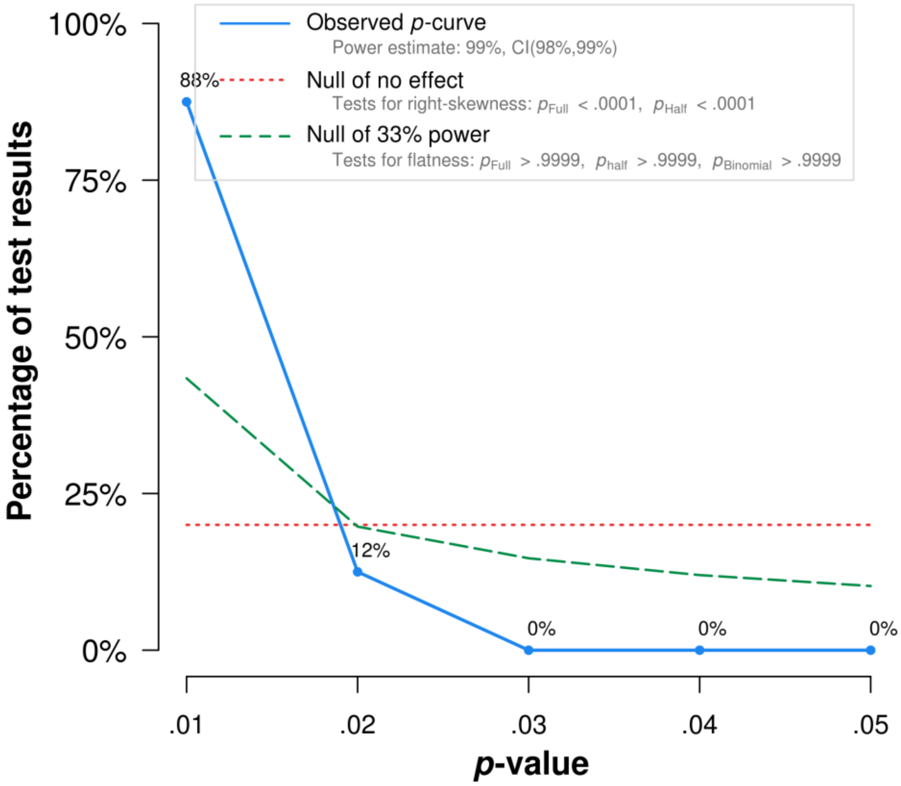
Exhibit 1. A *P*-curve of hindsight bias literature.

Exhibit 2. P-Curve Disclosure Table

Original Paper	(1) Quoted text from original paper indicating prediction of interest to researchers	(2) Study design	(3) Key statistical result	(4) Quoted text from original paper with statistical results	(5) Results	(6) Used trivia stimuli?
Anderson 1986	Unpublished dissertation	N/A	N/A	N/A	N/A	N/A
Arkes et al., 1981	No statistical test value	N/A	N/A	N/A	N/A	N/A
Arkes et al., 1988	Before hindsight subjects indicated the probability they would have assigned to the "correct" diagnosis, we asked them to provide a reason why each of the other diagnoses might have been correct. If the Koriat et al. results generalized, we would have expected the bias to be less in the "reasons" subjects compared with those who were not asked to list such reasons.	**2 (group: reasons vs. no reasons) x 4 (perspective: foresight vs. hindsight-alcohol vs. hindsight-alzheimer's vs. hindsight-brain damage**	**two-way interaction**	A chi-square analysis comparing these two splits (42:30 vs. 28:40) confirms the hypothesis that the frequency of subjects manifesting the hindsight bias is significantly greater among hindsight subjects than among hindsight-reasons subjects, **x2(l, N = 140) - 4.12, p<.05**.	ch2(1) = 4.12	no
Brown & Solomon, 1987	Evaluations of managerial decisions will be significantly affected by outcome information when the evaluator has no prior involvement with the evaluatee's decision process and the reported outcome implies that the evaluatee had relatively hiher responsibility for anticipating the outcome.	**2 (involvement: prior involvement vs. no prior involveent) x 4 (outcome: no report vs. failure-copyright vs. failure-economic vs. success)** *(attenuated interaction)*	**two-way interaction**	Table 1 presents the results together with descriptivestatistics. The results indicate that the outcome effect in both comparisons is significantly different from zero. Thus, hypothesis one is confirmed.	t(95) = -2.71	no
Buchman 1985	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A
Bukszar & Connolly, 1988 - Study 1	The present study attempted to probe that possibility, using subjects with substantial formal training in strategic decision making. We addressed four specific questions: (1) To what extent are such subjects vulnerable to hindsight bias, given high task involvement and rich case materials?	**three-cells**	**Difference of means**	The results (see Table 1. first row for each dependent variable) strongly suggest that subjects were not able to ignore the outcome information. Mean estimates of the probability of first-year success (item a) and mean first-year predicted ROI (item b) were higher for the high-outcome subjects and lower for the low-outcome subjects than for the no-outcome subjects.	F(2,47) = 10.15	no
Bukszar & Connolly, 1988 - Study 2	As suggested earlier, such hindsight effects might influence the way an observer evaluates a decision process and allocates praise and blame to the participants. Such evaluation might, in turn, lead to redesign of the process or replacement of the participants, so that a distorted reading of the past could generate undesirable consequences in the future. To probe this possibility, we added several process-related items to the questionnaire (see the Appendix) and replicated the study with a new group of strategy students.	**three-cells**	**Difference of means**	Substantive measures of hindsight effects (items a, b, and f) reached statistical significance, as they did in study one.	F(1,25) = 11.45	no
Butterworth, 1988	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A
Campbell & Tesser, 1983	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A
Connolly & Bukszar, 1990	The procedure, described below, yielded two groups identical as to case infor- mation and outcome, but differing in opportunity to make self-fiattering connections between known outcomes and analysis of the early situation that led to those outcomes. If 'real outcome' subjects show more hindsight shift than do 'random outcome' subjects it would suggest that self-presentation effects play a significant role in generating hindsight bias. Conversely, if hindsight shift is comparable for the two groups of subjects, motivational explanations that turn on the subject's pleasure in skilful analysis are weakened	**three-cells**	**Difference of means**	One-way analysis of variance showed that, compared to Group 1 (the no-outcome group). Groups 2 and 3 judged the project to have had a significantly higher probability of success, and to have been significantly less risky, at the time of the original investment.	F(2,45) = 12.5	no
Davies, 1987 - Study 1	Armed with their foresight record, hindsight judges should be better equipped to recover their foresight perspective. Using an anchoring and adjustment strategy, judges in receipt of outcome knowledge might start with the reported outcome as the anchor (p = 1 .OO) and adjust downward according to how much they can recall of their foresight state of uncertainty. Having their original foresight records available in hind- sight would then allow judges to adjust their probability estimates further from the initial anchor, thus reducing hindsight bias.	**2 (task: review vs. no review) x 2 (info: outcome knowledge vs. no outcome knowledge)** *(attenuated interaction)*	**two-way interaction**	Although the re- view/no review variable did not reach significance as a main effect (F(1,77) = 2.28, p < .20), it was significant in interaction with the outcome knowledge variable **(F(1,77) = 3.88, p < .06**).	F(1,77) = 3.88	no
Davies, 1987 - Study 2	The production or recording of one's thoughts before the event may result in a more elaborate and durable trace of one's state of knowl- edge or uncertainty, such that this foresight state is more easily accessed in hindsight - in the same way as episodic memory for facts is improved by encoding operations that produce elaboration of stimulus events (Anderson & Reder, 1979; Bradshaw & Anderson, 1982). Experiment 2 investigated this possibility by comparing hindsight judgments of subjects who had previously generated and recorded their foresight cognitions with subjects who had not	**2 (writing: notes vs. no notes) x 2 (info: outcome knowledge vs. no outcome knowledge)** *(attenuated interaction)*	**two-way interaction**	However, there was neither a signifi- cant main effect of notes/no notes nor a significant interaction with out- come knowledge **(Fs < 1)**	F's < 1	no
Davies, 1987 - Study 3	Reviewing one's foresight thoughts and reasons might serve to reduce hindsight bias simply by increasing the amount of attention devoted to the outcome and explanations for the outcome that did not occur, thus increasing the availability of the nonreported outcome without necessarily reviving the original foresight beliefs about it. On this view, the amount of hindsight reduction obtained by reviewing foresight cognitions would be equivalent to the amount of hindsight reduction obtained by hindsight generation of reasons.	**2 (task: review vs. generate) x 2 (info: outcome knowledge vs. no outcome knowledge)** *(attenuated interaction)*	**two-way interaction**	With respect to the main purposes of the experiment, the review/generate x outcome/no outcome interaction was not significant **(F = 0.41)**	F = 0.41	no
Dawson et al., 1988	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A
Detmer et al., 1978	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A
Fischhoff & Beyth 1975	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A
Fischhoff 1975	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A

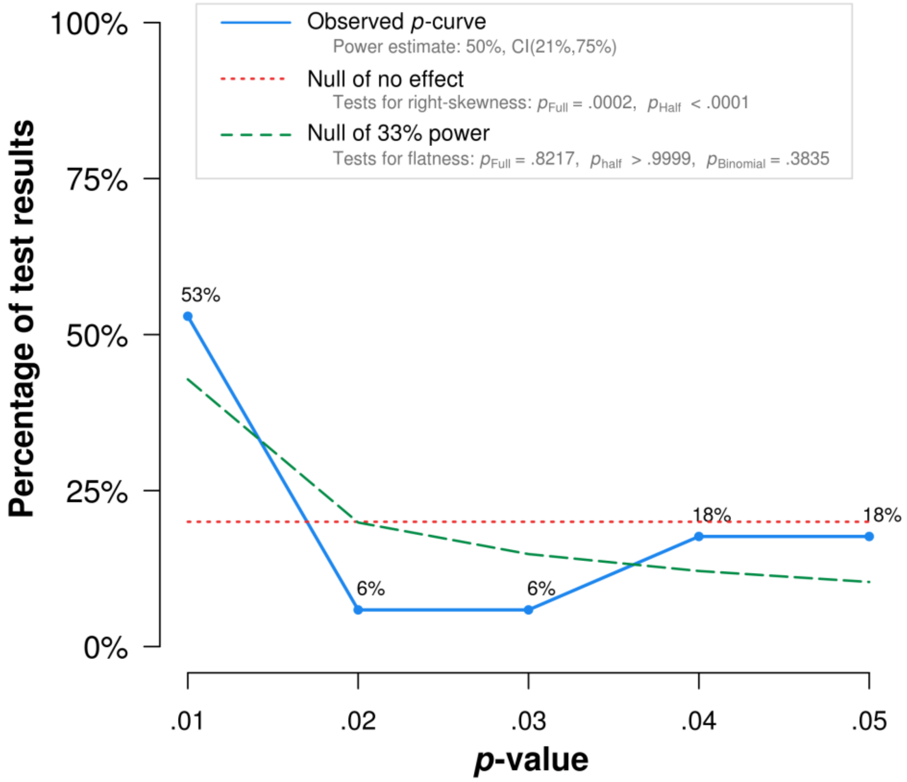
Original Paper	**(1) Quoted text from original paper indicating prediction of interest to researchers**	**(2) Study design**	**(3) Key statistical result**	**(4) Quoted text from original paper with statistical results**	**(5) Results**	**(6) Used trivia stimuli?**
Fischoff 1977	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A
Goggin & Range 1985	Not Accessible	N/A	N/A	N/A	N/A	N/A
Goiten & Rotenberg 1977	Thus, since for Calvinists the damned category is the basic category into which one falls back, once one ceases to provide success signs, even in the absence of signs of failure (Rotenberg, 1974, 1975a) it is possible that Westerners would find negative outcomes in particular moreprobable than would non-Westerners on the basis of the same information.	**2 (nationality: American vs. Israeli) x 3 (group: prediction vs. post-election vs. post-damnation)** *(attenuated interaction)*	**two-way interaction**	While among Israelis there is a drop in probability ratings from pre to post condition (see Table II), this interation, though in the expected direction, is not significant: ***F* (1, 268) = 1.37**, ns.	F(1, 268) = 1.37	no
Greenberg 1982	Unpublished dissertation	N/A	N/A	N/A	N/A	N/A
Hasher et al - 1981 - Study 2	It is possible (in fact, likely, based on the outcome of the second experiment) that the systematic reversal of the true and false categories in a plausible context ("I accidentally reversed them") enabled subjects to continue to believe in th credibility of the feedback and so to continue to use this easily available information during the rerating trial. Thus we made a second attempt to create circumstnaces under which feedback would be disregarded.	**2 (stage: Trial 1 vs. Trial 2 - within) x 3 (condition: true feedback vs. false feedback vs. no feedback)** *(attenuated interaction)*	**two-way interaction**	The next set of analyses was conducted on the change scores shown by subjects in the three xperimental cnditions. A positive value means that subjects' belief in the statements increased across the trails; a negative value, that belief decreased. There ere significant effects for items, *F* (2,174) = 71.60, as well as an interaction between items and instrucitonal conditions, ***F* (4, 174) = 2.996**, MSe = .243.	F(2,174) = 16.70	yes
Hasher et al., 1981 - Study 1	With these findings in mind, we attempted to alter the retrieval plans of subjects in a knew-it-all-along procedure. Our goal was to demonstrate that subjects in such a situationcan indeed remember their original knowledge state. Such a finding would be important because it would establish limits to the knew-it-all-along effect as well as to the extent to which assimilation processes are believed to operate in memory	**2 (stage: Trial 1 vs. Trial 2 - within) x 3 (condition: true feedback vs. false feedback vs. no feedback)** *(attenuated interaction)*	**two-way interaction**	An ANOVA comparing the three item types for both conditions revealed a significant conditions x items interaction , ***F* (2, 140) = 22.04**, MSe = .66.	F(2,140) = 22.04	yes
Hell et al., 1988	In summary, memory trace strength, time of correct information, and motivation to recall correctly were in- dependently varied in an almanac-type questions experi- ment, with numerical estimates as the dependent varia- ble. Hindsight bias should be larger with weaker memory strength for the original response, there are contradictory predictions for the effect of the time of correct informa- tion, and a higher motivation to report correctly should be associated with less hindsight bias if the memory trace for the original response is accessible separately. *(No interaction predicted)*	**2 (request: reason requested vs no reason requested) x 2 (time of correct info: after original response vs. right before recollection) x 2 (motivation: high vs. low)** *(attenuated interaction)*	**Difference of means**	The (no-reason-requested manipulation shows a sig- nificant main effect [**F(1,57) = 19.3, p < .01**]. As can be seen in Figure 1, the effect is in the expected direc- tion: more hindsight bias in the no-response-requested conditions (the upper four data points).	F(1,57) = 19.3	yes
Helleloid 1985	Unpublished dissertation	N/A	N/A	N/A	N/A	N/A
Hoch & Loewenstein, 1989 - Study 1	In sum, it is possible, but by no means obvious, that outcome feedback will help subjects to discriminate between easy and difficult items. Whether it does depends on the degree to which subjects can accurately assess familiarity and can objectively infer the congruence of outcome feedback with existing relevant knowledge. At one extreme it is possible that subjects could extract information from their own reactions to feedback without experiencing any hindsight, though such a finding seems unlikely, given the robustness of the hindsight effect. At the other extreme, subjects could experience hindsight sufficiently powerful to wipe out the potentially beneficial effects of feedback. If subjects always experience strong "I knew it all along" reactions, even on difficult items, then feedback will not aid in item discrimination.	**two-cell**	**Difference of means**	Repeated measures multivariate analyses of variance (MANOVAS) indicated that feedback subjects (M — 69) assigned higher probabilities to the correct answer than did controls (A/= 62), F(l, 106) = 24.8, MS, = 170, p < .0001	F(1, 106) = 24.8	yes
Hoch & Loewenstein, 1989 - Study 2	Same as Study 1	**four-cell**	**Difference of means**	For both studies, own confidence ratings made after receipt of feedback were better aligned with actual target norms [gamma(f, o) in Table 1] than were own confidence ratings without the benefit of feedback, **F(1, 154) = 35.9, MS, = .03,p < .0001**, in E2, and F(3, 77) = 6.6, MSK = .03, p < .001, in E3. The	F(1, 154) = 35.9	yes
Hoch & Loewenstein, 1989 - Study 3	Same as Study 1	**four-cell**	**Difference of means**	For both studies, own confidence ratings made after receipt of feedback were better aligned with actual target norms [gamma(f, o) in Table 1] than were own confidence ratings without the benefit of feedback, F(1, 154) = 35.9, MS, = .03,p < .0001, in E2, and **F(3, 77) = 6.6, MSK = .03, p < .001**, in E3.	F(3, 77) = 6.6	yes
Hoch & Loewenstein, 1989 - Study 4	Our model of outcome feedback proposes that the ability of feedback subjects to experience differential surprise (diag- nostic of underlying base-rates) on receipt of feedback drives the information effect. In this experiment both control and feedback subjects rated their level of surprise to each of the answers.	**2 (task: recall vs. recognition) x 2 (feedback: control vs. outcome feedback)** *(attenuated interaction)*	**Difference of means**	Surprise analyses were conducted at the group and individual level. Five subjects (2 controls and 3 feedback) were excluded from the analyses because they indicated no surprise (s = 1) to all items. The average surprise ratings of the control (M ~ 2.4) and feedback (M = 2.2) groups did not differ, **F(1, 66) = 1.85, MSe — 1.12, p = .185**	F(1,66) = 1.85	yes
Hoch & Loewenstein, 1989 - Study 5	The cues that subjects use to simulate memorability when recall is prevented (either by recall failure in feeling-of-know- ing studies or by preemptive outcome feedback in our exper- iments) seem to be absent for insight problems. Thus, feed- back about insight problems may not provide subjects with much diagnostic information.	**2 (feedback: control vs. outcome feedback) x 2 (problem type: insight vs. incremental)** *(attenuated interaction)*	**two-way interaction**	Target peer predictions were analyzed by using a 2 x 2 x 2 Feedback x Problem Type x Replication repeated measures MANOVA including both sol- vers and nonsolvers.6 Figure 6 shows the significant Feedback x Problem Type interaction, **F{ 1, 62) = 11.68, MS, = 478, p < .001**. On insight problems, exposure to the correct answer caused subjects to estimate that a larger fraction of their peers would successfully solve the problem.	F(1,62) = 11.68	no
Janoff-Bulman & Timko, 1985	If, with the benefit of hindsight, we unjustifiably perceive events as more predictable, then we assume that those whosuffered as a result also should have known about the predicatble sequence of events; thus they should have been able to do something to avoid or prevent the negative outcome.	**2 (outcome: rape vs. no rape) x 4 (likelihood judgment: seducer vs. raped vs. beaten vs. taken home)** *(attenuated interaction)*	**two-way interaction**	A significant outcome X likelihood judgment interaction effect was also found, *F* (3, 312) = 2.74, *p* < .05	F(3, 132) = 2.74	no
Janoff-Bulman & Timko, 1985 - Study 2	It seems liely that the hindsight effec twould particularly influence judgments of behavioral blam by observers. If links between outcomes and prior events become strengthened in hindsight, the in the case of negative events observers are apt to question why the victims did not *behave* differently so as to alter events and avoid the victimization.	**2 (outcome: rape vs. no rape) x 2 (blame: behavioral vs. characterological)** *(attenuated interaction)*	**two-way interaction**	Furthermore, a significant outcoe X type of blame interaction was found, *F* (1, 76) = 8.29, *p* < .01	F(1, 76) = 8.29	no
Janoff-Bulman & Timko, 1985- Study 3	The effect of presenting observers with alternative outcomes depends upon which outcomes they choose to use; that is, which outcomes they feel could reasonably follow from antecedent events. If observers employ more than one outcome, the hindsight effect would essentially be reversed, for they would have ecognized that a particular series of behaviors/events would be minimized.	**2 (outcome: rape vs. no rape) x 2 (blame: behavioral vs. characterological) x 2 (outcome explained: yes vs. no)** *(reversing interaction for the latter 2 x 2)*	**both simple effects**	Test not reported		no

Original Paper	(1) Quoted text from original paper indicating prediction of interest to researchers	(2) Study design	(3) Key statistical result	(4) Quoted text from original paper with statistical results	(5) Results	(6) Used trivia stimuli?
Leary 1981	If self-esteem and/or self-presentation factors mediate hindsight distortion, post facto expectancies would be predicted to be most closely in line with actual game outcome among highly ego-involved Ss who gave their predictions publicly. If the phenomenon arises from nonmotivational, in- formation processing factors, only a main effect of timing (prediction given before vs. after the game) would be expected.	**2 (stage: before game vs after game) x 2 (privacy: public vs. private)** *(attenuated interaction)*	**two-way interaction**	A three-way analysis of variance (timing x response publicness x ego- involvement) performed on the difference scores revealed only a main effect of timing, **F(1, 85) = 6.67, P < .02**.	F(1, 85) = 6.67	no
Mazursky & Ofir 1990 - Study 1	We hypothesized that subjects assessing the high-quality movie (judged by three raters and the previous year's students as representing very high quality) will recall having assigned lower levels of expectations for such movie than subjects exposed to the low-quality movie.	**three-cells**	**Difference of means**	A one-way ANOVA was performed on the three levels of expectations (i.e., preexposure expectations and recalled expectations for the low- and high-quality movies). The analysis revealed significant between-group variation (**F(2,81) = 3.48, p < .03**).	F(2, 81) = 3.48	no
Mazursky & Ofir 1990 - Study 2	Accordingly, we hypothesized that a high degree of surprise will be associated with reconstruction of expectations to reflect the feeling of "I did not expect it to happen."	**three-cells**	**Difference of means**	The omnibus F-tests, however, are significant in both analyses suggesting that both surprise level **(F(2,33) = 11.1, p < .001)** and disconfirmation (F(2,33) = 12.5, p < .001)	F(2,33) = 11.1	no
Mazursky & Ofir 1990 - Study 3	In Experiment 2 as in Experiment 1, the response to unexpected performance was manifested both by assigning higher ratings to the performance measure and by derogating past expectations. Experiment 3 provides a replication of these findings in a different domain.	**two-cell**	**Difference of means**	A significant main effect for the judgment reference factor was obtained as well **(F(1,48) = 5.41, p < .03)**	F(1,48) = 5.41	no
Mitchell & Kalb, 1981	More specifically, we hypothesized that knowledge of the outcome of a subordinate's poor performance will (a) increase the supervisor's estimate of the **future probability of the same outcome recurring**, (b) cause the supervisor to see the subordinate as more responsible for the behavior and for the outcome, and (c) result in more internal attributions for the subordinate's behavior.	**2 (outcome: outcome feedback vs. no feedback) x 2 (valence: negative vs. benign)** *(attenuated interaction)*	**two-way interaction**	The interactions for the two questions were also significant: **F(1, 51) = 4.83, p < .03**, for the probability questions, and F(1, 46) = 3.84, p < .06, for the percentage question.	F(1,51) = 4.83	no
Pennington 1981a - Study 1	In view of this research, and that on hindsight judgements, it was hypothesized that outcomes of the General Election which were favourable to a particular party would be seen as more likely than unfavourable outcomes, in both foresight and hindsight, by members of that party	**two-cell**	**Difference of means**	For outcomes which occurred in Questions tz to 7 there was only one main effect of foresight/hindsight. This was for Question 5b; the swing away from Liberal of o.1 to 3.0 per cent (**F = 4.49; d.f. = 1,36; P < 0.05**)	F(1,36) = 4.49	no
Pennington 1981a - Study 2	With an objective criterion by which to compare estimates made in hindsight and foresight it was predicted that those who were told what the actual figures were (hindsight) would produce estimates closer to those actual figures than those who were not told the figures (foresight).	**two-cell**	**Difference of means**	The ten questions concerned with the number or percentages of women in certain types of employment at the University of \Varwick (Questions 1 to 10 in Table 3) yielded six significant differences between foresight and hindsight estimates. These were all in the predicted direction, and as follows: Question 1, Number of Professors (**F = 7.29; d.f. = 1,82; P < 0.01**).	F(1,82) = 7.29	no
Pennington 1981b	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A
Pennington et al., 1980	It was predicted that those who knew the result of the pregnancy test (hindsight) would perceive that outcome as more likely, i.e: those who received a positive result should see their chances of being pregnant as more likely in hindsight than in foresight; conversely those receiving a negative result should, in hindsight, see their chances of being pregnant as less likely than in foresigh	**2 (pregnant: yes vs. no) x 2 (timing: before result vs. after result)** *(attenuated interaction)*	**two-way interaction**	There was no difference between the before and after (foresight and hindsight) estimates (F = 3.6, d.f. = 1, 18, P > 0.05), and no interaction with positive/negative (**F= 0.5, d.f. = 1, 18, P > 0.05**)	F(1, 18) = 0.5	no
Ross et al., 1977	The present research was designed to test directly the contention that the process of explaining an event increases its subjective likelihood for the perceiver. Three experiments are reported, each dealing with explanations and predictions of behavior in a clinical judgment context. In Experiment 1, experimental subjects attempted to explain events that they believed to be authentic at the time of their explanation task; however, before making predictions, subjects learned that the events were fictitious.	**2 (case: suicide vs. peace corps) x 2 (condition: explanation vs. control)** *(attenuated interaction)*	**two-way interaction**	Test not reported		no
Ross et al., 1977 - Study 2	By examining the effects of explicitly hypothetical explanations and by comparing the relative impact of hypothetical and nonhypothetical explanations, we therefore hoped to explore the extent to which our initial findings were dependent upon such processes.	**2 (event: explained vs. not explained) x 2 (explanation: hypothetical vs. non-hypothetical) x 2 (case: suicide vs. peace corps)** (attenuation of attenuated interaction)	**three-way interaction**	Finally, on the difference-in-likelihood-estimates measure, there was a significant second- order (Event Explained X Hypothetical/Non- hypothetical Explanation X Case) interaction, F(1, 24) = 5.07, p < .05, suggesting that the relative effects of the two types of explanation may vary somewhat as a function of specific contextual factors and circumstances.	F(1,24) = 5.07	no
Ross et al., 1977 - Study 3	Experiment 3 again compared the effects of hypothetical and nonhypothetical explana- tion upon estimates of the subjective likelihood that the explained event had indeed occurred.	**2 (event: explained vs. not explained) x 2 (explanation: hypothetical vs. non-hypothetical) x 2 (case: suicide vs. peace corps)** *(attenuation of attenuated interaction)*	**three-way interaction**	Test not reported		no
Slovic & Fischhoff 1977	No statistical test value (only p value inequalities)	N/A	N/A	N/A	N/A	N/A
Synodios 1986	In line with previous findings, it was hypothesized that subjects will be more confident and estimate the probabilities of occurrence and the percentage of votes received by each candidate more accurately in hindsight than in foresight.	**two-cell**	**Difference of means**	Confidence of subjects in their predictions was significantly affected by timing (**F(1,449) = 18.73, *p* < .0001**): Post-election subjects were more confident than pre-election subjects.	F(1,499) = 18.73	no
Verplanken & Pieters 1988	Not Accessible	N/A	N/A	N/A	N/A	N/A
Wood, 1978 - Study 1	In the present experiments subjects received outcome knowledge for an entire set of items, and then sometime later they were asked to rate the items. The procedure used by Fischhoff is probably optimal for obtaining a knew-it-all-along effect. If you want to demonstrate that subjects are unable to ignore outcome knowledge, the best time to present feedback would seem to be just prior to the time they perform the rating task. The use of a time delay should provide a more robust test of the phenomenon.	**two-cell**	**Difference of means**	The change from Stage 1 to Stage 3 is statistically significant, **F(1, 58) = 16.85, p < .001**, for the feedback conditions.	F(1,58) = 16.85	yes
Wood, 1978 - Study 2	along effect more with memory than with peer instructions because subjects in the memory conditions are explicitly asked to remember their previous knowledge state whereas subjects in the peer-instruction conditions are not.	**two-cell**	**Difference of means**	the two instruction conditions reveals a significantly larger effect for subjects in the memory-instruction conditions than in the peer-instruction conditions, **F(1, 112) =**	F(1,112) = 6.68	yes



Note: The observed *p*-curve includes 8 statistically significant (*p* < .05) results, of which 8 are *p* < .025. There was one additional result entered but excluded from *p*-curve because it was *p* > .05.

Exhibit 3. A *P*-curve of hindsight bias literature only using trivia stimuli.



Note: The observed p-curve includes 17 statistically significant ($p < .05$) results, of which 11 are $p < .025$. There were 3 additional results entered but excluded from p-curve because they were $p > .05$.

Exhibit 4. A *P*-curve of hindsight bias literature without trivia stimuli.