Reducing Hindsight Bias: Tests of a Retrieval-Based Theory

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Martin Van Boekel

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Sashank Varma, Geoffrey Maruyama

July 2016

Acknowledgements

Without the support and guidance of many people this dissertation would not have been possible. First, I would like to thank my advisors, Sashank Varma and Geoff Maruyama. From the very beginning, they both encouraged me to pursue my interests, pushing me to become a better researcher. Thank you for the excellent mentorship that you have both provided over the years. My journey to becoming a somewhat coherent writer has been long, and torturous for those who have had to read early drafts of my work. So, I must extend a special thanks to Sashank for all of his helpful feedback in this area of my development.

I have worked with many fantastic people over the past five years. Keisha Varma and Panayiota Kendeou, despite not officially being counted as advisors, have played critical roles throughout my graduate career. Thank you for the countless hours spent supporting me, and for the guidance (both professional and academic). I would also like to extend my thanks to my other dissertation committee member, Wilma Koustaaal, for her valuable comments and guidance in the development of this research.

There are several people who also deserve thanks, they haven't necessarily helped with the development of this research project, and are unlikely to ever read this dissertation, but deserve the acknowledgements all the same. Thank you Kristen McMaster, for the support through my first three years of my graduate school career. I appreciate the introduction into the world of large grant funded projects and classroom-based intervention work. I would also like to thank Michael Rodriguez for all of his support and leadership with the research projects coming from MyDrG. I have learned a lot from these research projects, and it has been nice seeing some of this work being useful to the schools. I would like to extend my thanks to my fellow students. I was fortunate to be surrounded by brilliant peers who consistently pushed me to become a better student and researcher.

Finally, I would like to thank the Graduate School for their financial support. My final year has been funded by the University of Minnesota's Doctoral Dissertation Fellowship.

Thank you.

Abstract

Individuals exhibit *hindsight bias* when they are unable to recall their original responses to novel questions after correct answers are provided to them. Here, I present a series of studies that investigates factors that reduce or eliminate hindsight bias. Specifically, I test the *retrieval-based theory*, which proposes that hindsight bias can be avoided under task conditions that support the generation of *sufficiently discriminative retrieval cues*. These discriminative retrieval cues allow participants to selectively retrieve their original judgments, even after being provided with the correct answers. Experiment 1 used the standard memory-design hindsight bias task and a modified design where participants were asked to recall their original judgments and the correct answers. Unexpectedly, participants in the standard memory-design avoided hindsight bias. As predicted, and consistent with the retrieval based theory, participants who engaged in the modified task were also able avoid hindsight bias and were able to recall the correct answers. In order to better understand Experiment 1's surprising results, Experiment 2 used a think-aloud methodology to determine which retrieval strategies participants were using when they correctly recalled their original judgment. Participants were observed successfully using the discrimination strategy when they were not prompted to do so (standard design), and when they were prompted to do so (modified design), providing support for the discriminative retrieval cue mechanism. Experiment 3 investigated whether sufficiently discriminative retrieval cues continued to reduce hindsight bias after a one-month delay, and whether repeated retrieval contributes to this reduction. Participants were observed engaging in hindsight bias after a one-month delay, suggesting the utility of the

discriminative retrieval cues deteriorates over time. Further, repeated retrieval did not mitigate the presence of hindsight bias following this time delay. Understanding the factors that reduce or even eliminate hindsight bias is important because it informs competing cognitive theories of this effect, and because it potentially informs the design of science instruction that minimizes hindsight bias and supports more normative reasoning.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Hindsight bias is the tendency for individuals' recollections of earlier predictions to shift towards the actual outcomes of events once this outcome knowledge is provided (Pohl, 2007). Hindsight bias has been studied extensively in laboratory settings (Christensen-Szalanski & Willham, 1991; Guilbault, Bryant, Brockway, & Posavac, 2004). However, implications of this bias extend beyond the laboratory. Researchers have argued that engaging in hindsight bias may have detrimental effects on learning (Roese & Vohs, 2012) and on scientific reasoning (Slovic & Fischhoff, 1977) because engaging in hindsight bias may lead a person to be overestimate the certainty of their knowledge, and may result in a reduced search for alternative explanations to unexpected outcomes. It is therefore important to understand the implications of hindsight bias on learning, specifically for science education (Bernstein, Aßfalg, Kumar, & Ackerman, in press).

Scientific reasoning, whether by scientists or science students, is subject to cognitive biases. Researchers in educational psychology and science education have examined these biases when learners respond to information provided by scientific experiments that proves to be surprising or contradictory. The results consistently show that both adults and children alike engage in hindsight bias when confronted with outcome knowledge such as unexpected results. More specifically, once people learn the results of an experiment, they tend to adjust their original beliefs and report that the unexpected results were foreseeable, and that they knew the results would happen all along (Davies, 1987; Hom & Kaiser, 2016; Slovic & Fischhoff, 1977). In terms of scientific reasoning, this can be thought of as the tendency to shift one's *a priori* hypothesis after it has been falsified by an experiment and to pretend to have believed an

a posteriori hypothesis all along – "Hypothesizing After the Results are Known," or "HARKing" (Kerr, 1998).

Hindsight bias is problematic for scientists because *a posteriori* hypotheses are over-fitted to the data at hand, and therefore less likely to hold under replication. This bias is problematic for science students because failure to recognize the discrepancy between their preconceptions and scientific knowledge can short-circuit subsequent learning. Some level of cognitive dissonance is necessary for developing reasoning skills and experiencing conceptual change (Bjork, 1994; diSessa & Sherin, 1998). Understanding the processes that give rise to hindsight bias is important for identifying factors that reduce or eliminate hindsight bias. This is a necessary first step in developing instructional strategies to mitigate hindsight bias when it occurs in science classrooms and other applied settings.

The present set of studies investigates factors that reduce or eliminate hindsight bias. Specifically, I evaluate the hypothesis that under the appropriate task conditions, participants will be able to generate *sufficiently discriminative retrieval cues* that allow them to *selectively retrieve* their original judgments (i.e., hypotheses) even after being provided with the correct answers (i.e., experimental outcomes), and to not display hindsight bias by mistakenly retrieving the correct answers. This hypothesis is taken from the retrieval-based theory of hindsight bias and is based on the recent finding that when adolescents were asked to recall *both* their original judgments and the provided correct answers, they were able to do so, and in turn avoid hindsight bias (Van Boekel, Varma, & Varma, 2016). By contrast, adolescents who were asked to simply recall their original judgments, as is typical in hindsight bias studies, engaged in hindsight bias. In an effort to

better understand the factors that facilitate the elimination of hindsight bias during retrieval, the current research examined the discriminant retrieval cue mechanism proposed by the retrieval-based theory – research goal (1) explored in Experiments 1 and 2. It also investigated whether these retrieval cues remain sufficiently discriminative over a delay of one month – research goal (2) explored in Experiment 3. Finally, it investigated the role of repeated retrieval in mitigating the onset of hindsight bias over time – research goal (3) explored in Experiment 3.

Experiment 1 attempted to replicate and extend to young adults the findings of Van Boekel et al.'s (2016) Experiments 1 and 2, which used middle-school participants. In contrast to past research, participants from the present study's Experiment 1 asked only to recall their original judgments were able to avoid hindsight bias and recalled their original judgments accurately. As expected, when participants were asked to recall both their original judgments and the correct answers they were able to do so accurately. The unexpected results from Experiment 1 motivated Experiment 2, which replicated the first experiment under the tight confines of a laboratory setting. In addition, a think-aloud methodology was used to document the recall strategies participants used when they were able to successfully avoid hindsight bias, with the goal of directly testing the retrieval-based theory. Experiment 3 further tested the retrieval-based theory by investigating whether the passage of time may constrain people's ability to selectively discriminate between retrieval cues for their original judgments and the correct answers. In addition, Experiment 3 linked the hindsight bias literature to a widely studied memory phenomenon, *the testing effect.* It investigated whether engaging in repeated recall of both the original judgments and the correct answers (immediately after learning the

correct answers, and again one month later) leads to better performance (i.e., less hindsight bias) one month later when performance is compared to a second group of participants who did not engage in the immediate recall task, and who only recalled their original answers and the correct answers after the one-month delay.

Taken together, these three studies illuminate the phenomenon of hindsight bias and provide partial support for the retrieval-based theory. The General Discussion discusses the importance of understanding the factors that reduce or even eliminate hindsight bias in the context of both cognitive theorizing and instructional design. It also delineates the limitations of the current research and identifies directions for future research.

Chapter 2: Literature Review

In the first empirical study of hindsight bias, Fischhoff (1975) presented participants with a narrative describing one of four actual events: the British-Gurka war, a near riot in Atlanta, and two clinical cases. The narrative provided enough description so that any one of four potential mutually exclusive and exhaustive outcomes was possible. For the British-Gurka war, for example, the possible outcomes were a) British victory, b) Gurka victory, c) military stalemate and d) military stalemate with a peace settlement. Participants in the control condition were asked to rate the probability of the four possible outcomes, whereas those in the hindsight condition were provided with one additional line describing the actual outcome before they made their probability ratings. Participants in the hindsight condition rated the probability of the actual outcome as much higher than participants in the control condition, a result Fischhoff termed *creeping determinism*.

Since the publication of this seminal study, hindsight bias has been observed with children as young as 3 and adults as old as 95 (Bârliba & Dafinoiu, 2015; Bernstein, Atance, Loftus, & Meltzoff, 2004; Bernstein, Erdfelder, Meltzoff, Peria & Loftus, 2011; Birch & Bernstein, 2007; Coolin, Bernstein, Thornton & Loken Thornton, 2014; Coolin, Erdfelder, Bernstein, Thornton, & Loken Thornton, 2015; Ghrear, Birch, & Bernstein, 2016; Groß & Bayen, 2015; Pohl, Bayen, & Martin, 2010), with effects typically strongest for young children and older adults (Bayen, Erdfelder, Bearden, & Lozito, 2006; Bayen, Pohl, Erdfelder, & Auer, 2007; Bernstein et al., 2011, Massaro, Castelli, Sanvito, & Marchetti, 2014). Further, hindsight bias has been observed in a number of different contexts including: economics (Anderson, 2014; Anderson, Lowe & Reckers,

1993; Hussain, Shah, Latif, Bashir, & Yasir, 2013), law (Evelo & Greene, 2013; Harley, 2007), mathematics (Ash & Wiley, 2008), medicine (Arkes, 2013; Littlefair et al., 2016; Motavalli & Nestel, 2016; Reece Jones, 1995), politics (Blank, Fischer, & Erdfelder, 2003; Calvillo & Rutchick, 2014b; Fischhoff & Beyth, 1975; Nestler, Blank, & von Collani, 2008), science (Choi & Choi, 2010; Davies, 1987; Hom & Kaiser, 2016; Slovic & Fischhoff, 1977), and sports/games (Calvillo & Rutchick, 2014a; Gray, Beilock & Carr, 2007; Roese & Maniar, 1997). Researchers have also observed hindsight bias in cross-cultural studies, finding that participants from Asian countries (e.g., Japan and Korea) show greater hindsight bias than do their Western counterparts (e.g., England, France, United States) (Choi & Nisbett, 2000; Yama et al., 2010). However, the findings from the various cross-cultural studies have not always been consistent, and are proposed to vary based on the type of materials selected to study this phenomenon (Heine & Lehman, 1996; Louie, Chandrasekar, & Wu, 2014; Pohl, Bender, & Lachmann, 2002).

In this chapter I describe how hindsight bias is typically studied, and the theories used to explain this phenomenon in order to highlight the need for a theory that mechanistically describes situations where hindsight bias is not observed. First, I review the standard methods used to study hindsight bias, in order to set the stage for my experimental design decisions. Second, I review a number of studies that have successfully reduced or eliminated hindsight bias, in order to identify the factors at encoding and retrieval that support the elimination of hindsight bias. Third, I discuss the dominant memory-based models of hindsight bias, exploring limitations of these theories. Specifically, I identify their limited focus on the retrieval processes involved in hindsight bias as the source of their inability to predict and mechanistically describe situations

when hindsight bias is eliminated. Fourth, I discuss the retrieval-based theory, a theory proposed by Van Boekel et al. (2016) in order to address this limitation. I end by presenting two areas of research from the memory domain: source monitoring and the testing effect. I discuss the source monitoring literature in order to identify the similarities with hindsight bias. Based on these similarities, I argue that methodologies used to study source monitoring may be applied to the study of hindsight bias, notably the use of the think-aloud protocol. When investigating the persistence of hindsight bias across time, the task design requires participants to engage in repeated retrieval. Therefore, the testing effect literature is explored in order to provide insights into the role of repeated retrieval on the presence of hindsight bias over time.

Measuring Hindsight Bias

Hindsight bias is typically studied using either a *hypothetical-design* or a *memory-design* paradigm. In hypothetical-design studies, participants are presented with an event and its actual outcome. They then estimate the probability of the event outcome had they not been made aware of the actual outcome. This is the design used by Fischhoff (1975). Hindsight bias is exhibited when participants with knowledge of the actual outcome rate the probability of this outcome occurring higher than do participants without outcome knowledge.

The standard way of presenting the event using this design has been through the use of text (for recent examples see Blank, Diedenhofen, & Musch, 2015; Hom & Kaiser, 2016; Oeberst, von der Beck, & Nestler, 2014). Recently, researchers have begun to use other methods of observing hindsight bias within the hypothetical-design paradigm. Specifically, researchers have begun using auditory (Bernstein, Wilson, Pernat, &

Meilleur, 2012) and visual (Bernstein et al., 2004, 2011; Bernstein & Harley, 2007; Harley, Carlsen, & Loftus, 2004) stimuli that are degraded in order to reduce the clarity of the words or images being presented. Though the stimuli across these manipulations differ, the procedure used to measure hindsight bias is similar. For example, when studying visual hindsight bias participants are presented with a blurry image. The image's clarity slowly increases, and the participant is asked to notify the experimenter when they can accurately identify the image. For some of the trials participants are given the image's identity prior to completing the task. Researchers then ask the participants to estimate when a participant without outcome knowledge would be able to identify the image. Like the traditional, text-based hypothetical-design studies, hindsight bias is exhibited when participants with outcome knowledge rate the moment when naïve participants could identify the image as sooner than what is actually observed when no outcome knowledge is provided. Manipulating the format of traditional text-based hypothetical-design studies to include auditory and visual stimuli has important implications for the study of hindsight bias. By removing the reading requirement and reducing the potential for scenarios to be culturally biased (Pohl et al., 2002), these new stimuli can be used without the need for modification across development across the lifespan and across cultures (Bernstein et al., 2004, 2011).

The *memory-design* paradigm was first used by Fischhoff and Beyth (1975) and investigated participants' predictions of various outcomes of President Nixon's trips to China and the USSR. For example, for Nixon's 1972 trip to China one of the event outcomes was: *The U.S.A will establish a permanent diplomatic mission in Peking, but not grant diplomatic recognition*. Participants were asked to rate the likelihood of this,

and the other events on a scale from 0% to 100%. After the trips were completed, participants were asked to recall their original predictions and report whether they knew if each event happened or not. Participants' recall shifted away from their original predictions towards the outcome knowledge, becoming higher for events that participants were aware had occurred, and lower for events that did not happen.

Since their inception, the experimental procedure of memory-design experiments has been refined. For example, Fischhoff and Beyth (1975) did not provide participants with the outcome knowledge, but instead asked participants whether they believed the event had occurred or not. Relatedly, because participants were not given the outcome information within the confines of the experiment, the retention interval (the time between learning the outcome information and being asked to recall their original predictions) varied. The length of the retention interval plays an important role in the presence of hindsight bias, a factor that will be discussed later (Nestler, Blank, & Egloff, 2010).

A standard memory-design study of hindsight bias consists of three phases (see Figure 1a). During Phase 1, participants answer a set of questions, typically by providing numerical answers.[1] Their answers are referred to as their original judgments (OJs) (see Appendix A). Phase 1 is followed by a filler task which is used to clear participants' original judgments from working memory. During Phase 2, participants are provided with the correct answers (CAs) to a subset of the questions, this is followed immediately by Phase 3, where participants are given a "surprise memory test" which asks them to

[1] The memory-design is not restricted to questions with numerical responses. Researchers have used paired comparison questions, *What do you call a baby echidna? A puggle or a chuttle* (for example Arnold & Lindsay, 2007). The responses may differ, but the procedure is consistent with the more traditional numerical response items.

recall their original judgments for the entire set of questions; these are called their recall original judgments (ROJs). Hindsight bias is exhibited when participants' recall of their original judgments shifts away from their original judgments and toward the correct answers for the questions for which the correct answers were provided during Phase 2, but remain accurate for the questions for which correct answers were not provided (Hell, Gigerenzer, Gauggel, Mall, & Müller, 1988).

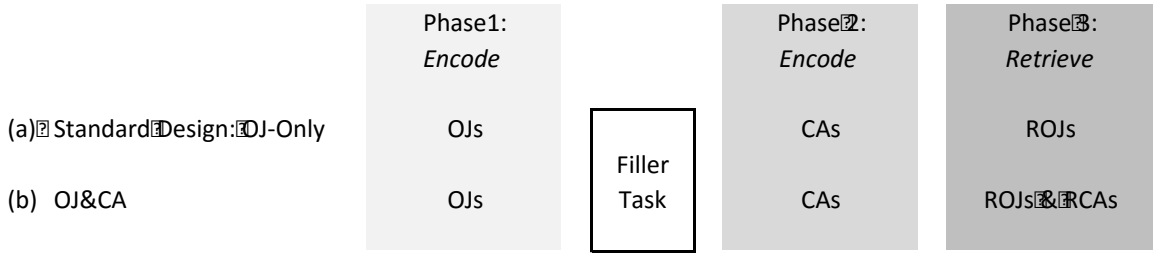

Figure 1. Design and procedure for: (a) the standard memory design, and (b) the modified memory design from Van Boekel et al. (2016). OJ = original judgments; CA = correct answers; ROJ = recall original judgments; RCA = recall correct answers.

Memory-design studies of hindsight bias have typically used almanac-type questions across development (for example Bernstein et al., 2011) and across cultures (for example Heine & Lehman, 1996). There are two important points to consider when using this design within a developmental context. First, because of the method's reliance on numerical estimates, it is important that the age groups under investigation be able to count up to the highest value in the answer set. Second, a memory-design study requires a filler task in order to clear participants' working memory. Extending the time it takes to complete the task may differentially impact participants across development, for example it may lead to greater fatigue for younger and older participants (Bernstein et al., 2011). There is an additional point to consider when using this design across cultural contexts: It

is important to ensure that the almanac-type items selected are of interest across all groups within the study; failing to do so may unduly influence the observation of hindsight bias (Heine & Lehman, 1996; Pohl et al., 2002).

Hertwig, Gigerenzer and Hoffrage (1997) argued that hypothetical-design studies and memory-design studies are not interchangeable, but in fact reveal qualitatively different phenomena. They proposed that the effect revealed by hypothetical-design studies is the "knew it all along" effect (Wood, 1978), whereas the effect revealed by memory-design studies is hindsight bias proper. However, most researchers define hindsight bias as a multi-faceted phenomenon comprised of three distinct components: *impressions of foreseeability*, *impressions of necessity,* and *memory distortions* (Blank, Nestler, von Collani, & Fischer, 2008; Nestler et al., 2010). Impressions of foreseeability correspond to the "knew it all along" feeling associated with hindsight bias; this is indicated when outcome knowledge is obtained, and participants claim that they would have been more likely to predict this outcome than participants who are not told the actual outcome. Impressions of necessity correspond to creeping determinism; this is indicated when outcome knowledge is obtained, and participants rate the event outcome as more inevitable than they would have prior to learning the actual outcome. Memory distortions are explained as biased recollection of original judgments arising from learning the correct answers.

Thus, which research design is needed is determined by which component is being studied, and vice versa. The hypothetical design is typically used to investigate the impressions of foreseeability and necessity components (Blank et al., 2008; Nestler et al., 2010), whereas the memory design is typically used to investigate the memory distortion

component. Nestler et al. (2010) note that because only a handful of studies have explored the relationships between the three components, it is premature to discontinue using hindsight bias as an umbrella term and to refer only to the three distinct components, as Hertwig et al. (1997) advocated.

Partitioning hindsight bias into distinct components enables researchers to focus on specific processes and their associated phenomena, and to propose and test theories of each one (Nestler et al., 2010). The experiments presented in this dissertation evaluate a memory-based theory of hindsight bias, one that focuses on the role of sufficiently discriminative retrieval cues in order to explore the memory distortion component of hindsight bias. Therefore, the experiments that follow adopt the memory-design paradigm.

Eliminating Hindsight Bias

Given the knowledge that hindsight bias is a robust phenomenon (Christensen-Szalanski & Willham, 1991; Guilbault et al., 2004), and the conventional view that hindsight bias has detrimental impacts on learning (Kerr, 1998; Roese & Vohs, 2012; Slovic & Fischhoff, 1977), a natural question is whether there are procedures for reducing hindsight bias. In one of the earliest attempts to reduce hindsight bias, Fischhoff (1977) warned participants about the effects of hindsight bias, and told them to do everything they could to avoid letting their recollections be biased by the outcome knowledge. Participants were also given instructions on how to attempt to recall their original judgment. Nevertheless, hindsight bias was still observed.

Building on this work, Pohl and Hell (1996) also provided participants with information about hindsight bias and the design of the experiment they were participating

in at three different time points in the experiment. Regardless of when the information was given – before the experiment started, after original judgments were made, or after the experiment had concluded – participants across the three groups demonstrated the same magnitude of hindsight bias when recalling their original judgments. In a second experiment, Pohl and Hell only used participants who were knowledgeable about the hindsight bias phenomena. In an effort to reduce the presence of hindsight bias, Pohl and Hell gave participants feedback on their recall attempts so that they could adjust their recalls in follow-up sessions. Even with performance feedback, participants' hindsight bias was constant across recall sessions. These findings demonstrated that hindsight bias persisted not just in the face of feedback aimed at reduction, but also across a time delay of two weeks.

While these studies serve to support the persistence of hindsight bias, there have been a handful of studies that have identified conditions under which hindsight bias can be eliminated. For example, an initial meta-analysis of 128 studies identified only 6 studies where hindsight bias did not occur (Christensen-Szalanski & Willham, 1991). A more recent meta-analysis replicated these findings, and also found that efforts to reduce hindsight bias have largely been unsuccessful (Guilbault et al., 2004). Given the scarcity of studies that have been able to successfully eliminate hindsight bias, it is important to take a closer look at the factors that have led to reductions in hindsight bias in the rare cases where this has been observed. I have organized these factors into four overarching categories: task factors, temporal factors, group factors, and cognitive factors. In this review I focus on memory-design studies. A few of these factors (see the expertise effect and the group effect) have yielded inconsistent effects across the two experimental

paradigms. In order to resolve these inconsistencies, I examined the results from the relevant memory-design studies and hypothetical-design studies.

Task factors. Task factors are those that manipulate the tasks participants perform within the memory-design paradigm. These task manipulations can occur in any of the three phases of the standard memory-design task.

Hindsight bias can be eliminated by improving the encoding of original judgments during Phase 1. For example, Hell et al. (1988) improved encoding by capitalizing on the generation effect – the finding of improved memory outcomes when information is actively generated versus passively comprehended (Slamecka & Graf, 1978). During Phase 1, participants generated their own explanations/reasons for half of their original judgments. During Phase 3, their recall judgments were much more accurate for just the original judgments for which they actively generated explanations, and consequently hindsight bias was greatly reduced.

Hindsight bias can also be eliminated by targeting the encoding of correct answers during Phase 2. Hasher, Attig, and Alba (1981) modified the memory-design task so that after the correct answers were given, participants in the experimental condition were immediately told that a mistake had been made and that the answers were, in fact, incorrect. During Phase 3, these participants made accurate recall judgments, avoiding hindsight bias. This finding has been replicated (Erdfelder & Buchner, 1998).

Manipulating the recall task during Phase 3 can also eliminate the effects of hindsight bias. Van Boekel et al. (2016) modified this phase by informing participants at the beginning of Phase 3 that they would be recalling *both* their original judgments and the correct answers, as opposed to the standard retrieval task of only recalling the original

judgments (see Figure 2a). These participants did not demonstrate hindsight bias, and additionally demonstrated accurate recall of the correct answers. These results held when both retrieval tasks were performed simultaneously (Experiment 2, see Figure 2b) or successively (Experiments 3 & 4, see Figure 2c).

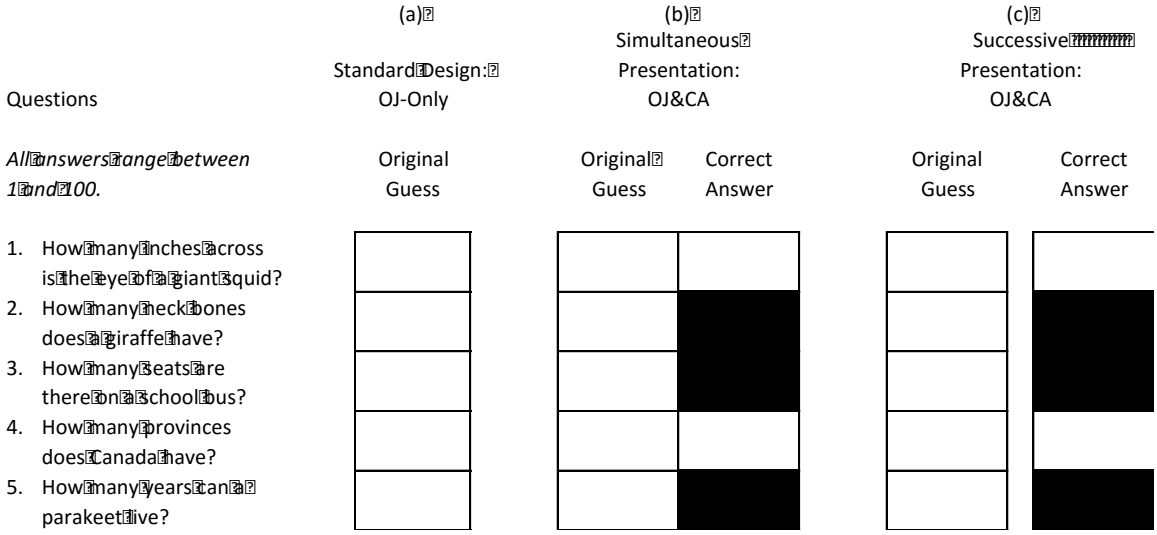

Figure 2. Fragments of the response sheets used in Van Boekel et al. (2016). The filled boxes are for the control questions, for which the correct answers were not provided during Phase 2.

Temporal factors. Temporal factors are those that involve manipulating some aspect related to the timing of the memory-design task. These manipulations may involve varying the length of time between Phases 1 and 2, or increasing the speed at which participants must recall their original judgments.

Memory-design studies require a filler task between Phases 1 and 2 to clear participants' working memory of their original judgments before they are told the correct answers in Phase 2. Manipulations made to the length of this filler task – known as the retention interval – can impact the presence of hindsight bias. For example, in Nestler et al.'s (2010) Experiment 2, participants experienced either a 10 or 60 minute retention interval after completing Phase 1. Participants in the shorter retention interval condition

were significantly better at recalling their original judgments than participants in the long retention interval condition, and thus displayed less hindsight bias. This finding was replicated and extended by Groß and Bayen (2015), who used intervals of either 20 minutes or 46 hours.

Similarly, manipulating the amount of time allowed when recalling original estimates during Phase 3 also affects the amount of hindsight bias that participants display. Calvillo (2013) had participants make their original judgments for 20 almanac-style questions. After receiving answers to half of these questions, participants were instructed to recall their original estimates either under a temporal constraint (less than three seconds per question) or under no temporal constraint. Participants in the no-constraint condition demonstrated less hindsight bias than those in the constraint condition.

Group factors. Group factors are those related to the conditions in which participants engaged in the study. Specifically, group factors examine the influence of engaging in the task individually or as part of a group on the presence of hindsight bias.

One manipulation that has successfully led to the reduction of hindsight bias is making recall judgments as part of a group as opposed to individually. I will refer to this as the *group effect*. In Stahlberg, Eller, Maass, and Frey's (1995) Experiment 2, which used a variant of the memory design, participants worked in groups or on their own to answer questions on a variety of different topics. After receiving the answers to a subset of these questions, participants recalled their original judgments. Participants who worked in a group demonstrated lower levels of hindsight bias than participants who worked alone.

The group effect has been found to vary as a function of the design of the experiment. Choi and Choi (2010) had participants engage in a hypothetical-design experiment, making probability estimates of different science scenarios either in groups or individually. They found that participants in groups demonstrated higher levels of hindsight bias than participants working on their own. This finding is inconsistent with Experiment 1 of Stahlberg et al. (1995), where participants completing a hypothetical-design task in groups (i.e., collectively rated the likelihood of the various studies' outcomes) did not confer any advantage relative to participants completing the task on their own, such that both groups exhibited equal amounts of hindsight bias.

Research from the memory literature provides insight into why participants engaging in the memory-design task are able to avoid hindsight bias. Research has demonstrated that participants working in a group have superior recall performance compared to participants working individually (Hinsz, 1990; Weldon & Bellinger, 1997). Hinsz (1990) proposed that groups have superior memory compared to individuals due to the interaction between three processes: pooling of information, enhanced error correction, and effective decision making. Groups tend to have a larger pool of memories compared to an individual, thereby increasing the likelihood that correct information is recalled in a group setting. When these memories are tested, groups are more effective than individuals at deciding what they collectively can and cannot remember, and therefore can identify and correct errors. As a result, groups have an increased probability that they will identify and exclude false memories, and an increased probability that they will identify and select correct memories.

Hinsz's (1990) three processes can account for the results observed in the group memory-design studies. When participants are working in a group to recall their original judgments, the group will have a larger pool of memories of the original judgments than an individual participant, and will better be able to identify and exclude recall judgments that have been influenced by the knowledge of the correct answers, thereby *reducing* the chance of hindsight bias for participants working in a group. In contrast, when participants working in groups are making probability judgments about an event after the outcome is known (hypothetical design), each individual may strive to impress their team members by leading them to believe that they could foresee the event, *increasing* the chance of hindsight bias for participants working in a group (Choi & Choi, 2010; Stahlberg et al., 1995).

Cognitive factors. Cognitive factors are those related to the participants engaged in the study. These cognitive factors are intrinsic to the participant (e.g. expertise).

One factor that has been shown to reduce or eliminate hindsight bias is domain expertise. Calvillo and Rutchick (2014a, 2014b) found that experts in both poker and politics demonstrated less hindsight bias than less knowledgeable participants when engaging in a memory-design task related to their domain of expertise. This finding, which is called the *expertise effect*, has also been investigated using computer simulations (Hertwig, Fanselow, & Hoffrage, 2003).

It is important to note that the expertise effect has not been consistently documented. In an early meta-analysis, Christensen-Szalanski and Willham (1991) found that expertise was a significant moderator of hindsight bias, with experts in a domain exhibiting less hindsight bias than novices. However, in a more recent meta-analysis, this

relationship was no longer significant (Guilbault et al., 2004). Guilbault and colleagues (2004) hypothesized that this shift may be due to their stricter definition of expertise. An analysis comparing studies with non expert participants to all of the studies identified by Christensen-Szalanski and Willham as "utilizing expert participants" resulted in nonsignificant differences between the two groups, $Q(1) = 0.12$, $p = .73$ (Guilbault et al., 2004), indicating that experts are as likely as non experts to engage in hindsight bias.

Calvillo and Rutchick (2014a, 2014b) suggest that the contradictory findings surrounding the expertise effect are largely driven by the way hindsight bias is studied. They demonstrated that when a memory-design task was used, experts were less susceptible to hindsight bias than they were when taking part in a hypothetical-design task. There is support for this claim from the memory literature. Recall that in memory-design experiments, hindsight bias is driven in part by a person's capacity to remember their original judgments. A well-established line of research has demonstrated that experts not only have superior knowledge when compared to novices, but also possess superior memory (i.e., due to effectively using chunking strategies to improve storage in and retrieval from long term memory) for information related to their domain of expertise (Chase & Ericsson, 1982; Ericsson & Charness, 1994; Ericsson & Kintsch, 1995; Larkin, McDermott, Simon, & Simon, 1980). This memory advantage would support the ability of experts to recall their original judgments during Phase 3, enabling them to avoid hindsight bias. However, with hypothetical-design studies participants' expertise may have the opposite effect, making known events seem even more plausible, and thereby *increasing* the chances of hindsight bias (Calvillo & Rutchick, 2014b).

Summary. The studies reported in this section demonstrate that even though hindsight bias is a robust effect, it can be eliminated by manipulating task factors, temporal factors, group factors, and cognitive factors.

Memory-Based Models of Hindsight Bias

Early theories proposed that hindsight bias occurs because the actual outcome knowledge was "assimilated" with the original judgment, rendering the original judgment inaccessible and permanently altered (Fischhoff, 1977). More recently, Blank et al. (2008) suggested that multiple theories might be needed to explain hindsight bias to fully account for the multi-faceted nature of this phenomenon. There are currently three prominent memory-based models of hindsight bias - the Selective Activation and Reconstructive Anchoring (SARA) model (Pohl, Eisenhauer, & Hardt, 2003), the Reconstruction After Feedback with Take the Best (RAFT) model (Hoffrage, Hertwig, & Gigerenzer, 2000), and the HB13 model (Erdfelder, Brandt, & Bröder, 2007; Erdfelder & Buchner, 1998).[2] They are reviewed here along with a newer theory, the Retrieval-Based Theory that directly motivated the current research (Van Boekel et al., 2016).

The Selective Activation and Reconstructive Anchoring (SARA; Pohl et al., 2003) model builds on the Search of Associative Memory (SAM) model of episodic memory (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981). It proposes that when participants are asked to generate their original judgments during Phase 1 they will use information from their knowledge base, or *image set* - a set of associations between

[2] Also prominent in the literature is the Causal Model Theory (Nestler et al., 2008). It is not considered here because it addresses probability estimates, the dependent variable of hypothetical-design experiments, and not recall judgments, the dependent variable of memory-design experiments. In addition, it focuses on metacognitive processes, whereas the focus here is on memory mechanisms.

images and retrieval cues. This image set is comprised of *images*, individual units of information related to the overarching topic. For example, if participants were asked to answer the question, *How many feet can a kangaroo jump in one leap?*, they will draw from information contained in the corresponding image set to answer the question, such as drawing from their image set for kangaroos. An image is created to represent their original judgment, and this image joins the image set. During Phase 2, when the correct answer is presented, *30 feet*, it is encoded into the image set as a new image. When the question is presented during Phase 3, it is decomposed into retrieval cues. In order to recall their original judgments, participants must rely on their image set. If the image for the original judgment cannot be recalled, then hindsight bias will result.

According to SARA, hindsight bias is the product of two processes (Blank & Nestler, 2007; Pohl et al., 2003). The first, *selective activation*, is a retrieval process. It is a consequence of encoding the image representing the correct answer during Phase 2, which changes the relationship between the images within the image set, including the original judgment, and the retrieval cues. The correct answer image impairs retrieval by acting as an "anchor". When prompted for recall of the original judgments during Phase 3, the correct answer image becomes activated, because the relationship between the original judgment image and the retrieval cues has changed, the probability the anchor will be recalled increases. The second process related to hindsight bias according to SARA is a reconstructive process known as *biased sampling*. When attempting to reconstruct the original judgment, the image representing the correct answer again acts as an "anchor". During Phase 3, images that are similar to the anchor will join the set of

retrieval cues in working memory, where they bias subsequent retrieval attempts, and ultimately the recall judgment.

The Reconstruction After Feedback with Take the Best (RAFT) model explains hindsight bias for the special case of paired comparison questions such as *Who will win the election, Clinton or Bush?* (Hertwig et al., 2003; Hoffrage et al., 2000). It proposes that during Phase 1, when people encounter a paired comparison question, they construct a *probabilistic mental model* (PMM) to make their original judgment. A PMM connects a set of objects in a reference class (e.g., Clinton, Bush, and other candidates for the US presidency) to a set of cues (e.g., economic prosperity, charisma, incumbent status). The "take the best" heuristic is then applied, which uses the single best cue to make the judgments. More precisely, cues are rank-ordered by their predictive validities. The best cue is selected (e.g., re-election) and if it discriminates between the objects, then the best object is retrieved and used to make the original judgment. If the cue does not discriminate, then the next best cue is selected. This process repeats until a viable cue is selected. If none of the cues discriminate, then the original judgment is made by randomly choosing among the objects. Encoding the correct answer during Phase 2 changes the cue values of the corresponding object in memory. In particular, some cues that were previously non-discriminative will become discriminative, some cues that were previously missing will become known, and some of these will erroneously point to the correct answer, not the original judgment. When making recall judgments during Phase 3, if people cannot recall their original judgment, then they must reconstruct it. They will again construct a PMM, rank order the cues by their predictive validities, and apply *the take the best heuristic* until a discriminative cue is reached that supports a judgment.

Critically, this rank-order will be different than in Phase 1 because some cues that were previously non-discriminative or unknown will now be (1) highly ranked, (2) discriminative, and (3) erroneously point to the correct answer. Their increased prominence will bias recall judgments towards the correct answer, producing hindsight bias (Hertwig et al., 2003; Hoffrage et al., 2000).

The third model, the HB13 model, is a multinominal processing tree model that specifies branching sequences of processes that lead to recall judgments during Phase 3 (Erdfelder & Buchner, 1998). Specifically, the HB13 model assesses the contributions of biased and unbiased retrieval and reconstruction processes on hindsight bias. The recollection and reconstruction processes are unbiased for the control items, those which the correct answers are unknown. For experimental items, those which the correct answers are known, these processes become more complicated. For the experimental items, the recollection process may result in accurate retrieval of the original judgment with probability r_E. The reconstruction is executed if the retrieval process fails (with probability $1 - r_E$). Either the original judgment will be reconstructed by the same process that produced it (with probability b), resulting in unbiased recall and an avoidance of hindsight bias, or it will be reconstructed by a process influenced by knowledge of the correct answer (with probability $1 - b$), resulting in hindsight bias. Although HB13 places an emphasis on the reconstruction processes, the estimates of r_E are typically non-zero, implying that the retrieval process (i.e., retrieval) plays a role during Phase 3 (Erdfelder et al., 2007). HB13, or variations of this model, have been particularly successful at accounting for group and individual differences in hindsight bias (Bayen et al., 2006;

Bernstein et al., 2011; Coolin et al., 2015; Erdfelder et al., 2007; Erdfelder & Buchner, 1998; Pohl et al., 2010).

The SARA, RAFT, and HB13 models have been successful in modeling and describing why people engage in hindsight bias, locating the source of hindsight bias in Phase 2, when the correct answer is encoded, proposing that learning the correct answer disrupts the reconstruction processes resulting in hindsight bias. However, in situations where these disruptions are "undone", and hindsight bias is avoided due to manipulations made to Phases 2 and 3 (Erdfelder & Buchner, 1988; Hasher et al., 1981; Van Boekel et al., 2016), the explanatory power of these models diminishes. Specifically, these models focus on the reconstruction processes involved in hindsight bias, and do not provide a mechanistic account of how hindsight bias can be avoided after participants have learned the correct answers.

To address this limitation, Van Boekel et al. (2016) proposed a retrieval-based theory of hindsight bias. It begins with the assumption, common to SARA and RAFT and inherited from more general models of episodic memory (Gillund & Shiffrin, 1984; Hintzman, 1986; Raaijmakers & Shiffrin, 1981), that every conscious experience is encoded as a trace in episodic memory. Once encoded, traces are not modified by subsequent processing, although they can become less accessible through processes like retroactive interference. Traces are retrieved via *similarity-based processing*. Specifically, the higher the similarity between the retrieval cues in working memory and the target trace, the greater the probability that the target trace will be retrieved. However, retrieval can be undermined by interference from competing traces: The higher the similarity between the retrieval cues and all other traces, the greater the probability that

one of the distractor traces will be retrieved instead. More generally, the more *discriminative* the retrieval cues – the more *selectively* they activate the target trace but none of the distractor traces – the greater the probability of selectively retrieving the target trace. When the question text is used as the sole retrieval cue, as is the case in the standard memory-design, it will be equally similar to both traces.[3] Thus, the correct answer trace is as likely to be retrieved as the original judgment trace. This constitutes a form of retroactive interference that biases recall judgments and leads to hindsight bias. Whereas, when participants are asked to recall both their original judgments and the correct answers, this simple manipulation may enable participants to generate a *compound cue* that is highly similar to just the original judgment trace, enabling its selective retrieval and avoiding hindsight bias. Moreover, participants may also generate a compound cue that is highly similar to just the correct answer trace, enabling its selective retrieval.

What distinguishes the retrieval-based theory from the SARA, RAFT, and HB13 models is its emphasis on the retrieval cues available during Phase 3. Specifically, the prediction is that if these cues are *sufficiently discriminative* – if they selectively activate the original judgment trace (target) but not the correct answer trace (distractor) – then hindsight bias will be eliminated. In contrast, SARA, RAFT, and HB13 propose that the recall process is driven by two processes: a simple retrieval process (also referred as a recollection process), and a reconstruction process. Within these models, hindsight bias is largely the result of the reconstruction process that takes over when the retrieval process

[3] In fact, the question retrieval cue might be *more* similar to the correct answer trace than the original judgment trace. This is because the question retrieval cue and original judgment trace were formed closer in time, and thus overlap more on their contextual features.

fails to lead to the original judgment (Den & Erdfelder, 1988). By focusing on the retrieval process, the retrieval-based theory is able to make a priori predictions regarding instructional manipulations that would result in the elimination of hindsight bias, a prediction that the other models, as they currently stand, cannot.

Source Monitoring

In order to test the retrieval-based theory, new methods are required to study hindsight bias, such as the collection of think-aloud protocols. Think-aloud protocols provide the opportunity to identify the retrieval strategies people use when recalling their original judgments. Think-alouds have not been used previously to study hindsight bias, and therefore the ability to use such a methodology in this context is unclear. Evidence for the utility of the think-aloud protocol in situations where recall performance is being tested comes from a related field, *source monitoring.* Before discussing the research where think-alouds have been used to study source monitoring, it is useful to define source monitoring and highlight the similarities between this construct and hindsight bias.

Source monitoring refers to people's capacity to identify the context in which memories are encoded (Johnson, 2006; Johnson, Hashtroudi, & Lindsay, 1993). There are many similarities between source monitoring and hindsight bias. Source monitoring errors arise due to failure in identifing the source of a memory, a process Johnson and Raye (1981) called *reality monitoring*. Reality monitoring errors arise from being unable to discriminate between internally generated memories (what an individual thought versus what they had said) and externally generated memories (which person provided a given piece of information). Similarly, the retrieval-based theory suggests that hindsight bias arises due to participants' failure in discriminating between their memories for the

original judgments (their internally generated memories) and the correct answers (externally provided memories).

One common way of measuring source monitoring errors is the Deese-Roediger-McDermott (DRM) paradigm (Bruce & Winograd, 1998; McKelvie, 2003; Roediger & McDermott, 1995). In this paradigm participants are presented with a list of words related to a concept. For example, a list may contain the words *table, sit, legs, seat, soft, desk, arm, sofa, wood, cushion, rest*, and *stool*, that are all related to the concept *chair* (nonrepresented associate) (Roediger & McDermott, 1995). Although the word "chair" is not included in the list, when participants are asked to recall the words from the list, the nonrepresented associate (*chair*) is falsely recalled with greater frequency (critical intrusion) than other words that were not presented (noncritical intrusion) (McKelvie, 2003). Critical intrusions can be considered an example of a source monitoring error because participants incorrectly identify the nonrepresented associate (an internally generated memory) as belonging to a group of externally provided memories. Again, this type of source monitoring error is similar to the retrieval errors observed in the memory-design task, where participants incorrectly identify externally provided memories (the correct answers) as their internally generated memories (their original judgments). Critical intrusions, like hindsight bias, can be minimized by increasing the discriminativeness of the different memory traces (Brédart, 2000).

Like the effects of hindsight bias, source monitoring errors are more prevelant in young children and older adults (Cycowicz, Friedman, Snodgrass, & Duff, 2001; Sprondel, Kipp, & Mecklinger, 2011; Swick, Senkfor, & Van Petten, 2006). For example, Lindsay, Johnson, and Kwon (1991) found that young children make more source

monitoring errors than adults as the similarity between two sources of information increases. However, as the discriminability of the two sources increases, children perform as well as adults on source monitoring tasks. These results suggest that when children are not supported in discriminating between sources, their capacity to do so is undermined when the similarities among the memory traces are high. These findings are consistent with the results observed by Van Boekel and colleagues (2016). In their Experiment 1 and the RCA surprise condition of Experiment 4, when the adolescent participants were only asked to recall their original judgments, they engaged in hindsight bias. However, when adolescent participants were prompted to recall both their original judgments and the correct answers – a prompt that Van Boekel et al. proposed increased the discriminability of the memory traces – hindsight bias was avoided.

Source monitoring researchers have posited that the greater prevelance of source monitoring errors in young children may reflect the immaturity of brain development, specifically the prefrontal cortex (Cycowicz et al., 2001; Sprondel et al., 2011). Sprondel et al. (2011) found that children (7-8 year olds) performed worse than adolescents (13-14 year olds) and adults (20-29 year olds) on a source monitoring task. Interestingly, no behavioral differences on measures of source monitoring were observed between adolescents and adults. However, there were neural differences between adolescents and adults, specifically in the neural correlates of source monitoring. These results suggest that as the brain develops, its capacity to engage in source monitoring becomes more refined, even though behavioral measures might not be sensitive to this continued refinement (Sprondel et al., 2011).

Several source monitoring studies have employed think-alouds when participants encode the word lists and when they recall these lists. They have found that the think-aloud process did *not* influence participants' overall performance on the DRM when compared to participants in a silent control condition (Goodwin, 2007, 2013; Goodwin, Meissner, & Ericsson, 2001). These findings, in conjunction with the similarities with hindsight bias noted here, motivate the use of the think-aloud procedure in Experiment 2.

The Testing Effect

Recall that retention interval manipulations were identified as a temporal factor that has successfully resulted in the elimination of hindsight bias (Groß & Bayen, 2015; Nestler et al., 2010). These studies manipulated the time between Phases 1 and 2 of the standard memory-design task, where shorter retrieval intervals have been associated with reductions in hindsight bias. Relatedly, the passage of time after Phase 3 can also be investigated as a temporal factor that may influence the presence of hindsight bias.

Pohl and Hell (1996) investigated the passage of time and targeted feedback as two factors that may influence the presence of hindsight bias. They had participants who were knowledgeable about hindsight bias engage in a standard memory-design task. On the first day, participants completed the standard three phases (see Figure 1a). One week later, participants returned to the laboratory and received feedback on their performance. This feedback consisted of a bias-index scale that indicated how well participants recalled their original judgments for each item (i.e. 'X' number of items moved towards the correct answer). A bias-index scale was reported separately for the experimental and control items. After reviewing their feedback, participants were instructed to use their feedback when recalling their original judgments a second time. This same procedure

was repeated again after a one-week delay. Participants' recall of their original judgments for both the experimental and control items did not change, for better or for worse, over time. In other words, their recall of their original judgments did not improve with feedback, resulting in hindsight bias that persisted across time.

Participants in the Pohl and Hell (1996) study engaged in repeated retrieval of their original judgments: immediately after learning the correct answers and at two other time points taking place at one-week intervals. Implicit in this design is the act of engaging in repeated retrieval of the original judgments. This study therefore represents an important link between the study of hindsight bias and the widely studied memory phenomena known as the *testing effect*. Researchers have found that information that has been repeatedly tested is better remembered than information that has been restudied for an equivalent amount of time (Roediger & Butler, 2011; Roediger & Karpicke, 2006b; Roediger & Pyc, 2012). The testing effect has been studied extensively in laboratory settings (Carpenter, 2009; Carpenter & DeLosh, 2006; Carpenter & Pashler, 2007; Cull, 2000, Roediger & Karpicke, 2006a). More recently, it has been studied in authentic classroom settings (Carpenter, Pashler, & Cepeda, 2009; McDaniel, Anderson, Derbish, & Morriesette, 2007; Roediger, Agarwal, McDaniel, & McDermott, 2011).

Although many of these studies have relied on short retention periods of up to one week (for example Butler, 2010), researchers have begun to examine longer retention intervals that may have more practical significance for the field of education (for example Larsen, Butler, & Roediger, 2009). Carpenter et al. (2009) noted the importance of understanding how the testing effect can influence retention over a longer time scale that is more amenable to educational contexts.

Although implicit in the design of the Pohl and Hell (1996) study, the influence of the testing effect on the emergence of hindsight bias was not explicitly examined, in part because a necessary control condition was not included. However, this does not mean the effects of repeated retrieval were not observed. Even though this manipulation did not successfully reduce hindsight bias, participants' magnitude of hindsight bias did not increase over time. Participants engaged in repeated retrieval of their original judgments, which may have strengthened these memories, and thus prevented increases in the amount of hindsight bias observed with the passage of time. Because a control group that did not engage in repeated retrieval was not included, disentangling the influence of feedback and the repeated retrieval was not possible. An interesting goal is understanding if engaging in repeated retrieval mitigates the onset of hindsight bias over the passage of time, as predicted by the testing effect, especially when participants engage in a retrieval task that supports the creation of discriminant retrieval cues. Experiment 3 addresses this goal.

Chapter 3: Research Questions

The results of Van Boekel et al. (2016) and the implications of the retrieval-based theory of hindsight bias more generally suggest a complex interaction between the encoding and retrieval conditions that produce – or fail to produce – hindsight bias. I investigate this interaction and evaluate the retrieval-based theory of hindsight bias in this dissertation.

Research Goal (1)

As discussed above, the dominant memory-based models of hindsight bias do not sufficiently explain or predict the situations where hindsight bias is not observed. Therefore, Van Boekel and colleagues (2016) proposed the retrieval-based theory. According to this theory, when participants create compound retrieval cues that support the discrimination between their original judgments and the correct answers, they will be able to accurately recall both sets of memories. The proposed sufficiently discriminative retrieval cue mechanism can account for why hindsight bias can be avoided in some contexts. However, the discriminative retrieval cue mechanism has not yet been isolated and *directly* tested. Experiments 1 and 2 addresses this important gap in two ways. First, they attempt to replicate and extend the findings of Van Boekel et al.'s (2016) Experiments 1 and 2, which used middle-school participants, to young adults. In one condition participants are asked to only recall their original judgment, the standard practice for memory-design studies of hindsight bias. In the other condition, the retrieval task is modified and participants are asked to recall their original judgments *and* the correct answers simultaneously. Second, Experiment 2 directly tests the discriminative retrieval cue mechanism proposed by the retrieval-based theory by asking participants to

think-aloud when recalling their original judgments, or as they recall both their original judgments and the correct answers.

Research Goal (2)

A second step in formalizing the retrieval-based theory is to identify the boundary conditions where the predictions regarding the presence of hindsight bias change. For example, while compound retrieval cues may help people discriminate between their original judgments and the correct answers over the short term (Van Boekel et al., 2016), the discriminativeness of these retrieval cues may deteriorate with the passage of time, therefore limiting the utility of interventions adopting this approach in classroom situations. Experiment 3 explores the passage of time as a factor that may constrain the utility of the discriminative retrieval cue mechanism. It uses a design similar to Experiments 1 and 2. The critical difference is that after a one-month delay, participants are asked to engage in a second recall task, where they once again recall their original judgments and the correct answers. The question is whether hindsight bias increases, stays the same, or decreases over this long delay.

Research Goal (3)

Investigating whether the passage of time modulates hindsight bias presents an opportunity to evaluate the relationship between this effect and a widely studied memory phenomenon, the testing effect. Again, this is the finding that a person's memory is better for information that has been repeatedly tested relative to information that has been studied for an equivalent amount of time (Roediger & Butler, 2011; Roediger & Karpicke, 2006b; Roediger & Pyc, 2012). Asking participants to repeatedly recall their original judgments and the correct answers over several weeks is very similar to the

procedure of studies of the testing effect (e.g., Butler, 2010; Larsen et al., 2009). Therefore, Experiment 3 investigates whether engaging in repeated recall (immediately after learning the correct answers, and again one month later) leads to better performance (i.e., less hindsight bias) one month later compared to a group that does not engage in the immediate recall task, and only recalls their original answers and the correct answers after the one-month delay.

Chapter 4: Experiment 1

Experiment 1 addressed research goal (1), testing the sufficient discriminative retrieval cue mechanism proposed by the retrieval-based theory of hindsight bias. This experiment sought to replicate and extend the findings of Van Boekel et al.'s (2016) Experiments 1 and 2, which utilized middle-school participants, to young adults. In Van Boekel et al.'s Experiment 1, hindsight bias was observed in middle-school students using a standard memory-design experiment. In their Experiment 2, hindsight bias was eliminated when the retrieval phase was manipulated so that participants simultaneously recalled their original judgments and the correct answers. Under this condition, participants accurately retrieved both their original judgments and the correct answers. These results support the retrieval-based theory, which predicts that if sufficiently discriminative cues are available at retrieval, then people will be able to selectively retrieve their original judgments, and avoid hindsight bias.

In the present experiment, I used the same materials from previous memory design studies (Bernstein et al., 2011; Calvillo, 2013; Hell et al., 1988; Van Boekel et al., 2016). Based on past research, and consistent with the retrieval-based theory, I predicted that when participants are asked only to recall their original judgments, they will engage in hindsight bias. By contrast, when participants are asked to recall both their original judgments and the correct answers, they will create sufficiently discriminative retrieval cues that enable them to accurately recall their original judgments and the correct answers, and thus hindsight bias will be reduced or eliminated.

Method

Participants

Sixty undergraduates from the University of Minnesota were recruited from four recitation sections of two introductory educational psychology classes. Participants received course credit for their participation.

Materials

The present experiment used a measure of hindsight bias developed by prior researchers (Bernstein et al., 2011; Calvillo, 2013; Hell et al., 1988). It consisted of 20 questions testing science content (see Appendix A). The questions were difficult (e.g., *How many hours does a lion sleep in a day?*), and participants were not expected to know the answers.[4] Phase 1 is when participants make their original judgments to the questions, Phase 2 is when participants learn the correct answers to half of these questions, and Phase 3 is when participants recall their original judgments to the items from Phase 1 (see Figure 3a).

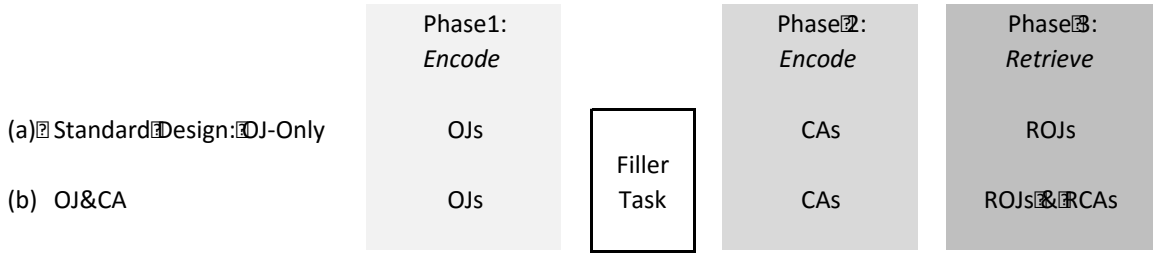

Figure 3. Design and procedure for Experiments 1 and 2. OJ = original judgments; CA = correct answers; ROJ = recall original judgments; RCA = recall correct answers.

[4] The instances where participants correctly guessed the correct answers when making their original judgments were rare for both the OJ-only condition (5.7%) and the OJ&CA condition (4.2%). These were similar to the rates observed in Experiments 2 (OJ-only, 3.1%; and OJ&CA, 3.4%) and 3 (Test, 6.0%; and No-testing, 3.9%).

Participants in the OJ-only condition made their original judgments on a document similar to the example in Appendix A with the exception that it had a blank line instead of the answers. Participants were given a second blank response document to complete during the retrieval phase (see Figure 4a). The materials were manipulated for the OJ&CA condition, where during Phase 3, participants used a response document that simultaneously prompted for their 20 original judgments and for the 10 correct answers provided during Phase 2 (see Figure 4b).

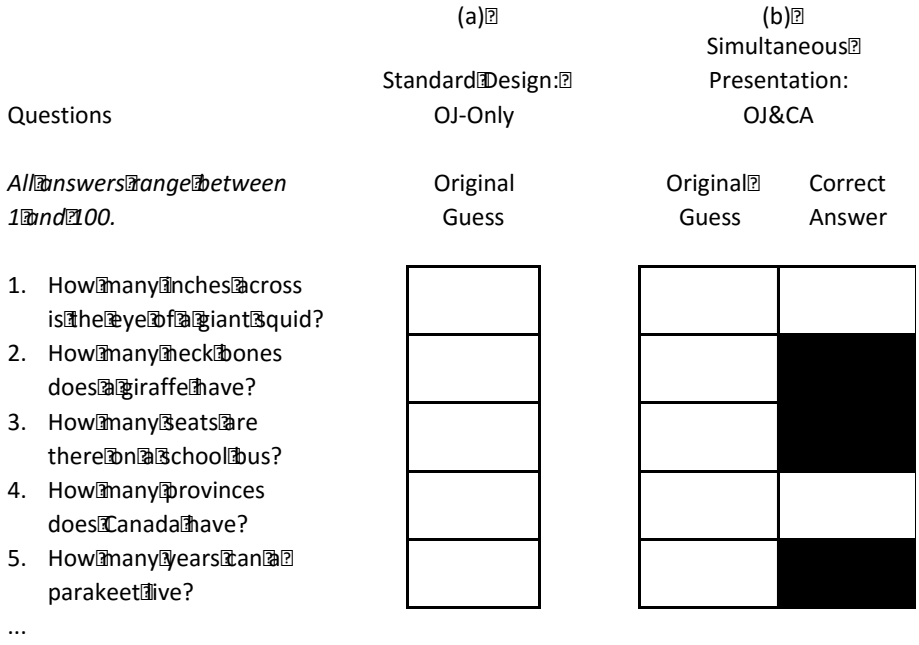

Figure 4. Fragments of the response sheets used in Experiments 1 and 2. The filled boxes are for the control questions, for which the correct answers were not provided during Phase 2.

Design

The experiment used a mixed-design. The within-subjects factor, *Type*, had two levels (experimental, and control). During Phase 2 participants were provided with correct answers to 10 questions (experimental level), but not for the remaining 10 questions (control level). The experimental questions were selected randomly using a

random number generator. The same 10 questions were used as the experimental questions across all three experiments.

The between-subjects factor, *Retrieval Task*, had two levels (OJ-only, and OJ&CA). Two recitation sections were assigned to each of the conditions to form two groups of comparable sizes. During Phase 3 participants either retrieved only their original judgments (OJ-only, $n = 35$), or they retrieved both their original judgments and the correct answers (OJ&CA, $n = 25$).

Two dependent variables were used to measure the accuracy of recall judgments made during Phase 3. The first was the *hindsight bias index* (HBI; Pohl, 2007), which is the standard way to quantify performance in memory-design studies (Bernstein et al., 2011; Calvillo, 2012, 2013). In a standard memory-design task, participants provide their original judgments (OJs) to a set of almanac style questions. Next, participants complete a filler task, after which they learn the correct answers (CAs) to a subset of these questions. Immediately after learning the correct answers, participants are given a "surprise memory task" where they are asked to recall the original judgments (ROJs). If a participant exhibits no hindsight bias their OJs and ROJs should differ from the CAs by approximately the same amount, on average. That is, the following expression, which Pohl (2007) termed the proximity index, should be near zero:

$$|OJ\text{-}CA| - |RJ\text{-}CA| \quad (1)$$

By contrast, if a participant exhibits hindsight bias, then their ROJs should be closer to the CAs than their OJs, and the expression should be positive.

The HBI was computed in three steps. First, the proximity indices for the control items (z_{CON}), the 10 questions for which participants were *not* provided with the correct

answers, were computed. The index was then standardized by dividing it by the standard deviations of all of the participants' responses to that question.

$$\frac{|OJ_i - CA_i|}{s_{OJ_i}} - \frac{|ROJ_i - CA_i|}{s_{ROJ_i}} \quad (2)$$

The median value of the 10 standardized indices (2) was selected to represent $z_{CON.}$ in order to control for the influence of extreme indices. Second, the proximity indices (1) for the experimental items (z_{EXP}), the 10 questions for which participants were provided with the correct answers, were computed in the same way. Again, the proximity index (1) was calculated, and then standardized using the standard deviations of all of the participants' responses to that question (2). The median value was used to represent z_{EXP}. Third, the HBI was calculated as the difference between these two values.

$$HBI = z_{EXP} - z_{CON} \quad (3)$$

The HBI measures the influence of knowing the correct answer on participants' subsequent recall judgments of their original judgments. Again, values close to zero indicate little or no hindsight bias and positive values indicate increasing hindsight bias.

The second dependent variable was the *correct answer index* (CAI). The CAI was created by Van Boekel et al. (2016) to reflect the accuracy with which correct answers were recalled during Phase 3. If a participant exhibits poor recall of the correct answers, then their recall of the correct answers (RCAs) should be similar to their original judgments (OJs). This is driven by the assumption that if participants cannot recall the correct answers, they will use whatever information available to make an estimate of the correct answer. This would likely be accomplished by either recalling their OJ directly, or relying on the same processes used when estimating their OJ. In either case, when

participants rely on these processes to recall the correct answer, the following expression should be near zero:

$$|OJ_i - CA_i| - |RCA_i - CA_i| \quad (4)$$

By contrast, if a participant exhibits accurate recall of the correct answers, then their RCAs should be closer to the CAs than their OJs, and the expression (4) should be positive.

The CAI was computed analogously to the HBI. First, proximity indices for the correct answers (z_{CA}) were calculated. This was done by standardizing the indices (1) by the standard deviations of all of the participants' responses to that question.

$$\frac{|OJ_i - CA_i|}{s_{OJ_i}} - \frac{|RCA_i - CA_i|}{s_{RCA_i}} \quad (5)$$

The median value of these standardized indices (5) was selected to represent the z_{CA}. Second, the CAI was computed as the difference between z_{CA} and z_{CON}.

$$CAI = z_{CA} - z_{CON} \quad (6)$$

Again, a value close to zero represented poor recall of the correct answers, and a positive value represented accurate recall of the correct answers.

Procedure

The procedure followed that of the standard memory-design (Figure 3a), though it was modified for group administration. Van Boekel et al. (2016) have successfully used the same materials and procedures in middle-school classroom settings. The procedure was modified for the present experiment to fit within a recitation section of an undergraduate class.

Phase 1 occurred at the beginning of the participants' recitation section. Participants were informed by their recitation leader that they would be completing an

estimation activity, and were each given a response sheet with 20 almanac-style questions (see Appendix A). They were told that the questions were difficult and that the answer to each one was between 1 and 100. They were then given as much time as they needed to answer the questions. Their answers constituted the original judgments (OJs). This phase took approximately 10 minutes to complete.

The recitation leader then conducted a regular lesson lasting 50 minutes, the purpose of which was to clear the working memory contents of the participants.

During Phase 2, the recitation leader informed the participants that they would be getting the correct answers to some of the questions they had answered at the beginning of the period. They were instructed to listen carefully, and not to talk or record any of the answers. The recitation leader then read aloud the answers to 10 of the 20 questions (e.g., "How many inches across is the eye of a giant squid? 15"). These constituted the correct answers (CAs).

This was immediately followed by Phase 3. Participants in both the OJ-only and the OJ&CA conditions were given a "surprise memory task". The OJ-only condition was given a response sheet with the 20 original questions (Figure 4a) and was asked to recall their original judgments to the questions made during Phase 1. These constituted their recall original judgments (ROJs). The OJ&CA condition was given a response sheet (see Figure 4b) and were asked to simultaneously recall both their original judgments made during Phase 1 and the correct answers provided during Phase 2. The recall of the correct answers constituted their recall correct answers (RCAs).

The order in which the questions were printed during Phase 1 and Phase 3 and the order in which correct answers were read during Phase 2 were fixed across all participants and across all three experiments.

Results

To review, prior to this experiment I had two hypotheses. The first was that participants asked to recall only their original judgments, the OJ-only condition, would demonstrate evidence of hindsight bias (Bernstein et al., 2011; Hell et al., 1988; Van Boekel et al., 2016). The second was that participants asked to recall their original judgments and correct answers simultaneously, the OJ&CA condition, would demonstrate accurate recall of both sets of memories (Van Boekel et al., 2016).

Contrary to my first prediction, participants in the OJ-only group accurately recalled their original judgments from Phase 1, showing no evidence of hindsight bias (see Table 1), one-sample $t(34) = 0.99$, $p = .331$. Possible explanations for this unexpected finding are provided in the Discussion. Consistent with my second hypothesis, participants in the OJ&CA conditions accurately recalled their original judgments from Phase 1, showing no evidence of hindsight bias (see Table 1), one-sample $t(24) = 0.81$, $p = .427$. Also as predicted, participants in the OJ&CA group accurately recalled the correct answers from Phase 2, as indicated by a CAI that differed reliably from zero in the positive direction, one-sample $t(24) = 6.90$, $p < .001$, $d = 2.82$ (see Table 1).

Table 1

Results of Experiment 1 reported M (SD).

	Experiment 1	
	OJ-only	OJ&CA
z_{CON}	-0.02 (0.01)	-0.02 (0.04)
z_{EXP}	-0.02 (0.02)	0.05 (0.08)
z_{CA}	NA	0.88 (0.51)
HBI	0.00 (0.02)	0.07 (0.43)
CAI	NA	0.90** (0.65)

Note. HBI and CAI statistically tested against 0. **$p < .001$. NA = not applicable because correct answers were not provided.

Discussion

The goal of Experiment 1 was to test whether the findings from Van Boekel et al.'s (2016) Experiments 1 and 2 could be replicated using a young adult population. In this, it was partially successful. Unexpectedly, participants in the OJ-only condition did not replicate the results of Van Boekel et al.'s Experiment 1. Rather, they were able to accurately recall their original judgments, and therefore to avoid hindsight bias. However, as expected, participants in the OJ&CA condition were able to accurately recall both their original judgments and the correct answers, replicating the results of Van Boekel et al.'s Experiment 2.

The question is, why were participants in the OJ-only condition able to avoid hindsight bias, especially given the robustness of this effect (Christensen-Szalanski & Willham, 1991; Guilbault et al., 2004)? One possible explanation focuses on procedural limitations. Unlike other studies that have used the same materials, these data were collected in a lecture setting as opposed to a controlled laboratory setting (Bernstein et

al., 2011; Calvillo, 2013). However, it is unlikely that the modifications necessary to administer this task to multiple participants simultaneously were responsible for the surprising results. Both Hell et al. (1988) and Van Boekel et al. (2016) used the same materials and very similar testing conditions to those employed in the present experiment, and yet were able to detect hindsight bias in their samples.[5]

A second, more likely explanation concerns developmental differences in retrieval strategy use during Phase 3. Developmental research has demonstrated that hindsight bias is typically stronger for young children and older adults than for young adults (Bayen et al., 2006, 2007; Bernstein et al., 2011; Coolin et al., 2014; Groß & Bayen, 2015; Pohl et al., 2010). Even though the size of the hindsight bias effect is expected to be smallest in young adult populations, we would not expect an absence of hindsight bias for this population (Christensen-Szalanski & Willham, 1991; Guilbault et al., 2004). Yet, the present study is not the first to observe low hindsight bias indices when using the memory design with a young adult sample. Bernstein et al. (2011) and Calvillo (2013) observed hindsight bias indices for young adults that were less than 0.10.[6] These values are close to the HBIs reported in Experiment 1 for the OJ-only and OJ&CA conditions (*HBI* = 0.00, and *HBI* = 0.07 respectively). Recall that an HBI close to zero represents accurate recall of the original judgments, and therefore no hindsight bias. These results, as well as the results from the OJ-only condition in Experiment 1, stand in stark contrast

[5] Hell et al. (1988) used a 90 item measure of hindsight bias, whereas the Van Boekel et al. (2016) used a 20 item subset of this larger measure.

[6] Actual hindsight bias indices were not reported by Bernstein et al. (2011) for ages above 5 years old. The HBI was reported in a graph (see Figure 4a on page 384) and is approximately 0.01, a value that would mark an absence of hindsight bias.

to Van Boekel et al.'s (2016) Experiment 1 results with adolescents, which found hindsight bias using the same materials and similar testing conditions.

A final explanation is suggested by Calvillo's (2013) second experiment. During Phase 3, the young adult participants that demonstrated minimal hindsight bias were instructed to use *at least* three seconds when retrieving their original judgments. In other words, Calvillo's participants were not constrained by a maximum time limit when making their recall judgments. This temporally unconstrained retrieval period is similar to that used in the present experiment. Thus, it may be that when young adult participants are given sufficient time to recall their original judgments, they *spontaneously* create compound retrieval cues that discriminate between their original judgment traces and the correct answer traces.

Experiment 2 tested this possibility by asking participants to think-aloud as they either recalled both their original responses and the correct answers (OJ&CA group) or recalled only their original responses (OJ-only group). The use of think-aloud protocols provided direct evidence of the retrieval strategies used by young adults, and potentially for the retrieval-based theory of hindsight bias. This enabled identification of cases where participants discriminated between their original judgments and the correct answers, and analysis of whether or not this led to accurate recall of their original judgments.

Chapter 5: Experiment 2

Experiment 2 addressed research goal (1) by looking for direct evidence for the discriminative retrieval cue mechanism proposed by the retrieval-based theory of hindsight bias. To do so, it employed a think-aloud methodology to observe the retrieval strategies participants used when they were recalling their original judgments. This methodology is not uncommon in memory research (Ericsson & Simon, 1993; Fox, Ericsson, & Best, 2010). For example, it has been applied to investigate participants' states of awareness of encoded information (remember vs. know, McCabe, Geraci, Boman, Sensenig, & Rhodes, 2011; Tulving, 1985) and children's capacity to identify memory strategies used to encode information (Jonsson, Wiklund-Hörnqvist, Nyroos & Börjesson, 2014). However, to my knowledge, Experiment 2 is the first to use think-alouds to study hindsight bias. The source monitoring literature reviewed in Chapter 2 (Goodwin, 2007, 2013; Goodwin et al., 2001) provides evidence for the suitability of using a think-aloud protocol to test the discriminative retrieval cue mechanism proposed by the retrieval-based theory.

In Experiment 2, participants were asked to think aloud during Phase 3, as they recalled their original judgments (and, in the OJ&CA condition, the correct answers) in order to identify the recall strategies that resulted in accurate recall. According to the retrieval-based theory, participants can avoid hindsight bias in the OJ&CA condition because the cue to recall both their original judgments and the correct answers supports the formation of compound retrieval cues that better discriminate between these memories. Therefore, it was hypothesized that when participants made verbalizations about discriminating between their original judgments and the correct answers, they

would be more likely to avoid hindsight bias. However, when participants verbalize other strategies that do not support such discrimination, such as engaging in a reconstructive process as proposed by the other memory-based models of hindsight bias, the prediction is that they will be more likely to engage in hindsight bias (Erdfelder et al., 2007; Erdfelder & Buchner, 1998; Hoffrage et al., 2000; Pohl et al., 2003). By contrast, when participants in both the OJ-only and OJ&CA conditions are recalling their original judgments for the control items (items for which the correct answers are unknown), they should be able to successfully use a variety of retrieval strategies, including reconstruction, because the correct answers remain unknown and therefore cannot influence the retrieval process.

Experiment 2 also investigated whether the unexpected finding of Experiment 1 – that participants in the OJ-only condition, who were asked only to recall their original judgments, were able to do so accurately and avoid hindsight bias – would replicate. This was particularly unexpected given that hindsight bias is a robust, highly replicated effect in the literature (Christensen-Szalanski & Willham, 1991; Guilbault et al., 2004; Hawkins & Hastie, 1990; Van Boekel et al., 2016). One hypothesis is that participants in the OJ-only condition, under the tightly controlled laboratory setting of Experiment 2 versus the classroom setting of Experiment 1, would engage in hindsight bias. The alternative hypothesis is that when young adult participants are not restricted by time when making their recall judgments, they can spontaneously create compound retrieval cues that support the discrimination between their original judgments and the correct answers, even when they are not prompted to do so. Experiment 2 uses the think-aloud methodology to test this hypothesis.

Method

Participants

Forty undergraduates from the University of Minnesota participated in the present study and were compensated $12 for their time. Participants were randomly assigned one of two groups, OJ-only (*n*=21*)* and OJ&CA (*n*=19*)*.

Materials

The materials were identical to those used in Experiment 1.

Design

The design was the same as Experiment 1 with the addition of the dependent measure, *Think-Aloud*, which reflects the strategies participants used to recall their original judgments during Phase 3. (The coding scheme is detailed below; see Table 2.)

Procedure

The procedure was similar to that of Experiment 1. Participants were tested individually in a laboratory setting.

Phase 1. Participants were welcomed into the lab, provided informed consent, and immediately began with Phase 1 of the hindsight bias task. They were presented with a response sheet containing the same 20 questions used in Experiment 1. They were told that the questions were difficult and that the answer to each one was between 1 and 100. Participants were asked to read each question out loud and state their answer to the question while the experimenter recorded their responses.[7] They were then given as much

[7] When the memory-design is modified for whole-class administration, it is necessary for the experimenter to read the correct answers out loud to ensure that participants have attended to that information. However, when participants make their OJs and ROJs they read and write their responses on their own. It is possible that this modality switching enhances the discriminability of the memory traces (Hicks & Marsh, 1999) because the

time as they needed to answer the questions. Their answers constituted their original judgments (OJs). This phase took approximately 5 minutes to complete.

Participants then engaged in a series of math cognition tasks lasting about 30 minutes, the purpose of which was to clear the contents of the participants' working memory. Upon completing the math cognition tasks, participants engaged in practice think-aloud tasks. These tasks were presented as a new experiment with the goal of understanding what people say to themselves as they work on different problems, but was actually meant to serve as a think-aloud instructional and practice period. Think-aloud practice was sequenced before Phase 2 because the alternative – before Phase 3 – would have introduced too great a delay between Phase 2 (when the correct answers are presented) and Phase 3 (when the original judgments and correct answers are recalled). Think-aloud instructions and practice tasks were taken from Ericsson and Simon (1993) and other studies using think-alouds to investigate memory processes (Goodwin, 2013; Goodwin et al., 2001; Williams & Hollan, 1981, complete think-aloud instructions can be found in Appendix B).

Participants were first asked to think-aloud as they determined the solution to '24 x 30'. Upon completing this task, participants were again reminded to not filter their thoughts and say everything they were thinking, and then asked to 'Name the state capitals that begin with the letter B' (Williams & Hollan, 1981). If a participant was silent for more than three seconds during their think-aloud they were prompted with the

---

source of the information is different (self vs. experimenter). Therefore, in the present experiment, participants were asked to read all of the questions out loud and state their responses out loud in order to control for these modality shifts.

following question: “What are you thinking right now?” The think-aloud instructions and practice tasks took about 5 minutes.

Phase 2. Phase 1 was followed by Phase 2. Participants were provided a correct answer sheet containing all 20 questions. The correct answers were provided for the 10 experimental questions; for the remaining 10 control questions, the correct answer cell was shaded black. Participants were asked to read the 10 experimental questions and the corresponding answers out loud.

Phase 3. Phase 2 was immediately followed by Phase 3. Participants were given a “surprise memory task.” Participants in the OJ-only group were asked to recall only their original judgments from Phase 1 (see Figure 3a, reproduced here as Figure 5a), and participants in the OJ&CA group were asked to recall both their original judgments and the correct answers (see Figure 3b, reproduced here as Figure 5b). Both groups were asked to read each question out loud, and to think aloud as they engaged in the retrieval process. If participants were silent for more than three seconds during their recall they were prompted with the following question “What are you thinking right now?”.[8] They then recalled their original judgments to the questions – the answers they provided during Phase 1. These constituted their recall original judgments (ROJs). The only difference was that participants in the OJ&CA condition were given a response sheet that asked them to additionally – and simultaneously – recall both the correct answers from Phase 2 (see Figures 4a and 4b, reproduced here as Figure 6). The experimenter recorded all participant responses. All think-alouds were recorded using a voice recorder.

[8] Only one participant required a reminder to think-aloud after the training was completed.

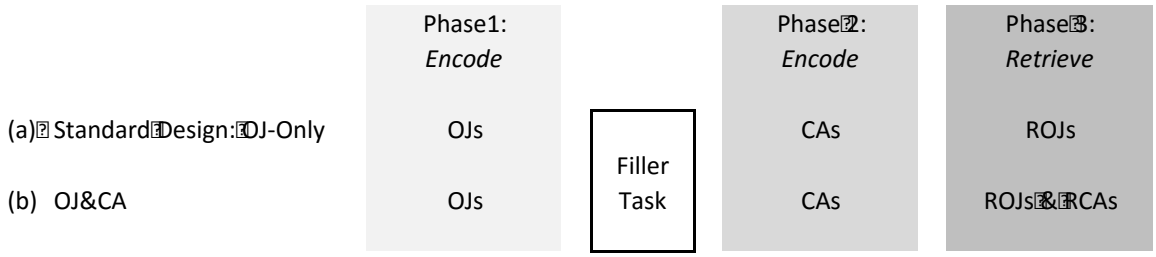

Figure 5. Design and procedure for Experiments 1 and 2. OJ = original judgments; CA = correct answers; ROJ = recall original judgments; RCA = recall correct answers.

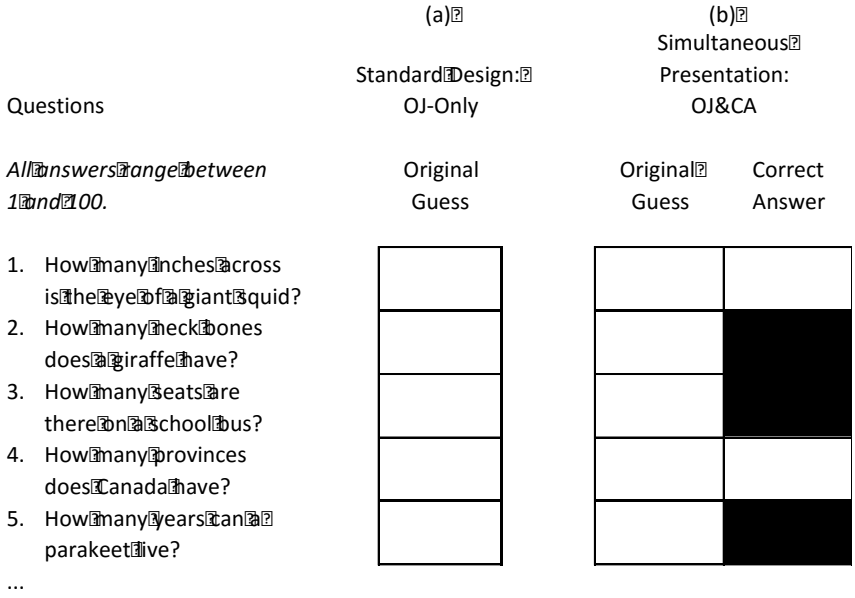

Figure 6. Fragments of the response sheets used in Experiments 1 and 2. The filled boxes are for the control questions, for which the correct answers were not provided during Phase 2.

Think-Aloud Coding

The application of the think-aloud methodology to the study of hindsight bias is new in this experiment, and therefore there was no predetermined coding scheme. Rather, the coding scheme emerged as participants' think-alouds were coded (see Table 2). The coding categories were informed by the mechanisms for explaining the presence or absence of hindsight bias proposed in the SARA, RAFT, and HB13 models and in the

retrieval-based theory. The think-aloud data support a direct test of the retrieval-based theory, through an investigation into the retrieval strategies participants use when recalling their original judgments (and the correct answers for the participants in the OJ&CA condition), and importantly, whether or not the use of a DISCRIMINATION strategy leads to higher rates of successful recall of the original judgments than the other retrieval strategies.

Table 2

Think-Aloud Coding Scheme.

Retrieval Strategy	Definition	Example
RECONSTRUCTION	Comments that described the steps/information participants used when recalling their original judgment	*"I said 47. The reasoning behind that was obviously humans only have so many vertebrate and giraffes are significantly larger, I knew it was under 50 and over 30 on the upper end"*
DISCRIMINATION	Comments that mentioned both the original judgment and the correct answer	*"I think the correct answer was 8, but I was wrong in saying something more, I think like 20"*
ANCHOR	Comments that mentioned using a numerical value contained within the task when making their original judgment or recall	*"The instructions said I couldn't have a value over 100, so I know I went with the upper limit and went with 100"*
PRIOR KNOWLEDGE	Comments that explicitly stated knowing the answer because the participant had learned that information from a reported source (could be correct or incorrect prior knowledge)	*"I read on a poster that cockroaches can live 30 days without a head"*
GUESS	Comment that reported the original judgment or the recall of	*"I said 10 minutes, a complete guess, I don't*

	the original judgment was a complete guess	*know"*
MULTIPLE STRATEGIES	Comments that identified the use of more than one of the above strategies	*"I said 14. I know lions sleep a lot. I didn't think it was 20, I figured they slept at least half of the time which would be 12 out of 24 so I just bumped that up two hours"* (Evidence of both RECONSTRUCTION and DISCRIMINATION)
NO STRATEGY	Comments that did not provide any evidence of strategy use	*"I said something like 64"*

The RECONSTRUCTION category is derived from the SARA, RAFT, and HB13 models (Erdfelder et al., 2007; Erdfelder & Buchner, 1998; Hoffrage et al., 2000; Pohl et al., 2003). Each one emphasizes the reconstructive process as the main cause of hindsight bias. Reconstructions were comments that described the steps/information participants used when retrieving/reconstructing their original judgment.

The retrieval-based theory (Van Boekel et al., 2016) proposed that when participants form compound retrieval cues, they will be able to discriminate between memory traces for their original judgments and the correct answers, and thus avoid hindsight bias. The category DISCRIMINATION was created to record examples of this strategy. Discriminations were comments that mentioned both the original judgment and the correct answer.

Participants were observed using three other recall strategies. On occasion participants reported using numerical values contained within the task itself as their original judgment. For example, a participant may have reported that they remembered an

answer being 100 because that was the highest value that they were allowed to use. The use of this strategy was coded as ANCHOR. Participants also reported that their original judgment or recall of that judgment was a complete guess. The use of this strategy was coded as GUESS. Finally, participants reported that they knew the answer to a question based on prior knowledge. To be given the code PRIOR KNOWLEDGE, participants must have explicitly stated that they knew the answer because they had learned that information from a reported source. This strategy was used infrequently. Participants were coded into this category whether their original judgment/prior knowledge was accurate or not.

Participants were not restricted in the amount of time they had to make their recalls during Phase 3. They were instructed to talk aloud "constantly" while recalling their original judgment; as such, some participants reported using several of the strategies noted above when recalling a single item. These verbalizations were coded as MULTIPLE STRATEGIES.

Many participants did not provide evidence of any recall strategy use, simply stating their original judgment or information unrelated to the task while engaging in retrieval. These responses were coded as NO STRATEGY.

The think-aloud response for each question was considered as an individual thought unit, and was not broken down into smaller units. The reasoning for this segmentation decision was two-fold. First, smaller units did not contain enough information to make a coding decision. Second, the goal of the think-aloud task was to look for evidence for the discriminative retrieval cue mechanism proposed by the retrieval-based theory. Ericsson and Simon (1993) advised that when a think-aloud

protocol is used to address goals related to theory building, the use of larger segmenting strategies is appropriate. One limitation of this decision was that the individual strategies reported within a segment could not be treated as independent. Therefore, when participants verbalized using multiple strategies, it was not possible to code each strategy used as a unique process. All think-alouds were scored by the first author, and then a second trained graduate rater scored half of the think-alouds independently to verify consistency (Kappa = 0.98).

Results

The results are broken up into two sections. The first section presents the results of the hindsight bias task, the second section presents the results of the think-aloud analysis.

HBI and CAI Measures

My original hypothesis was that participants asked to recall only their original judgments, the OJ-only condition, would demonstrate evidence of hindsight bias (Bernstein et al., 2011; Hell et al., 1988; Van Boekel et al., 2016). By contrast, participants recalling their original judgments and correct answers simultaneously, the OJ&CA condition, would demonstrate accurate recall of both sets of information (Van Boekel et al., 2016). However, based on the results from Experiment 1, an alternative hypothesis is that participants would demonstrate accurate recall of their original judgments regardless of their conditions, perhaps because the retrieval period was temporally unconstrained (Calvillo, 2013, Exp. 2), while those in the OJ&CA condition would also demonstrate accurate recall of the correct answers.

Consistent with the alternative hypothesis, Experiment 2 replicated the findings of Experiment 1 (see Table 3). Participants in the OJ-only group accurately recalled their original judgments from Phase 1, showing no evidence of hindsight bias, one-sample $t(20) = -1.30$, $p = .209$. Participants in the OJ&CA condition also accurately recalled their original judgments from Phase 1, showing no evidence of hindsight bias, one-sample $t(18) = -2.40$, $p < .05$, $d = 1.10$. Also as hypothesized, participants in the OJ&CA condition accurately recalled the correct answers from Phase 2, as indicated by a CAI that differed reliably from zero in the positive direction, one-sample $t(18) = 9.94$, $p < .001$, $d = 4.56$.[9]

Table 3

Results of Experiments 1 and 2 reported M (SD).

	Experiment 1		Experiment 2	
	OJ-only	OJ&CA	OJ-only	OJ&CA
z_{CON}	-0.02 (0.01)	-0.02 (0.04)	-0.03 (0.04)	0.02 (0.04)
z_{EXP}	-0.02 (0.02)	0.05 (0.08)	-0.04 (0.05)	-0.02 (0.04)
z_{CA}	NA	0.88 (0.51)	NA	1.02 (0.42)
HBI	0.00 (0.02)	0.07 (0.43)	-0.02 (0.06)	-0.03* (0.05)
CAI	NA	0.90** (0.65)	NA	1.01** (0.44)

Note. HBI and CAI statistically tested against 0. **$p < .001$; *$p < .05$. NA = not applicable because correct answers were not provided.

Before proceeding, I address the negative HBI observed in the OJ&CA condition. This indicates that participants more accurately recalled their original judgments for the questions for which they were provided correct answers during Phase 2 than for questions for which they were not provided correct answers. This finding has been termed *reverse*

[9] Item 4 was removed from the CAI calculation because all participants recalled the correct answer accurately.

hindsight bias. Reverse hindsight bias has been observed experiments using the hypothetical design (Calvillo & Gomes, 2011; Mazursky & Ofir, 1990; Nestler & Egloff, 2009), but it is surprising that this was observed using a memory design. Van Boekel et al. (2016) also observed participants demonstrating better recall for the experimental items than the control items in their third experiment, but was not replicated in their fourth experiment. The same pattern was observed in Experiments 1 and 2 of the present study, a negative HBI was reported here in Experiment 2, but not Experiment 1. Therefore, I conclude that this finding is likely to be spurious, and unlikely to replicate.

A cross-experiment analysis (two-way ANOVA) was conducted to evaluate whether engaging in the think-aloud influenced participants' recall of their original judgments. A between-subjects factor, *Retrieval Context*, indicated whether participants completed Phase 3 while thinking aloud (Experiment 2), or silently (Experiment 1). Neither the main effects for the type of Retrieval Task (OJ-only vs. OJ&CA), $F(1, 96) = 0.52$, $p = .475$, or Retrieval Context, $F(1,96) = 1.47$, $p = .229$, were significant, nor was the Retrieval Task x Phase 3 Retrieval Context interaction, $F(1,96) = 0.86$, $p = .356$. These results suggest that thinking aloud during Phase 3 did not modulate the presence or absence of hindsight bias.

Think-aloud Measures

The think-aloud data support a direct test of the retrieval-based theory. Specifically, I test the proposal that when participants provide evidence of using a DISCRIMINATION retrieval strategy, that they will be better able to discriminate between their original judgments and the correct answers, and thus avoid hindsight bias. In addition, because the results from the HBI and CAI measures replicated those observed in

Experiment 1, I directly test the alternate hypothesis that when young adults participants are not constrained by time, they spontaneously create compound retrieval cues that support discrimination between memory traces of their original judgments and the correct answers.

This section begins by presenting the overall retrieval strategy use trends and success rates for the control items, further broken down into separate analyses for participants in both the OJ-only and OJ&CA conditions. It is important to first consider the control items because these items represent a shared retrieval context across the two conditions, one that would not be influenced by correct answer knowledge for either of the two conditions. Therefore, understanding retrieval strategy use patterns for the control items is critical for establishing that there were no differences across the two conditions in the frequency or accuracy of strategy use, so that any differences with the experimental items can be attributed to the retrieval instructions. This is followed by an analysis of the retrieval strategy trends and success rates for the experimental items, further broken down by both conditions. This analysis, particularly the analysis of the retrieval strategy success rates, enables direct testing of the discriminative retrieval cue mechanism proposed by the retrieval-based theory.

Control items. As previously mentioned, unlike the experimental items, the retrieval context for the control items is similar for participants in both the OJ-only and the OJ&CA conditions. Therefore, the prediction is that across the two conditions participants will be able to successfully use a variety of retrieval strategies, and they will not differ in their strategy use and the success rates of those strategies.

***Retrieval strategy use*.** The present set of analyses examined whether participants in the OJ-only and OJ&CA differed in the *frequency* with which different retrieval strategies were used for the control items, for which they did not learn the correct answers during Phase 2 (see Table 4). Because the correct answers were only provided for the experimental items, it was impossible for participants to use the DISCRIMINATION strategy for the control items.

Table 4

Frequency of Retrieval Strategy Use for the Control Items, reported M (SD).

	Control Items			
Retrieval Strategy	OJ-only	OJ&CA	*t* (38)	*p*
Reconstruction	2.89 (2.94)	2.32 (3.15)	0.03	.975
Discrimination	NA	NA		
Anchor	0.00 (0.00)	0.26 (0.45)	2.67	.011
Prior Knowledge	0.38 (0.67)	0.16 (0.50)	1.18	.244
Guess	0.19 (0.40)	0.05 (0.23)	1.31	.197
Multiple Strategies	0.00 (0.00)	0.00 (0.00)	NT	NT
No Strategy	7.14 (3.48)	7.21 (3.78)	0.06	.953

Note. NA = not applicable because retrieval strategy could not be used; NT = success rate differences were not statistically compared because sample sizes were too small.

As expected, there were minimal differences between the OJ-only and the OJ&CA conditions in the frequency with which different retrieval strategies were used. There were no differences for the following strategies: RECONSTRUCTION, $t(38) = 0.03$, $p = .975$; PRIOR KNOWLEDGE, $t(38) = 1.18$, $p = .244$; GUESS, $t(38) = 1.31$, $p = .197$; NO STRATEGY, $t(38) = 0.06$, $p = .953$.[10] There was a difference for the ANCHOR strategy, $t(38) = 2.67$, $p = .011$, $d = 0.87$, which participants in the OJ&CA condition reported using more than participants in the OJ-only conditions.

[10] Participants from both conditions did not report using multiple strategies for any of the control items. Therefore, no statistics were reported for this strategy.

Retrieval strategy success rates. I next investigated whether the OJ-only and OJ&CA conditions differed in the *success rates* of the different retrieval strategies. These data are presented in Table 5. The most frequently used strategy was RECONSTRUCTION. When participants used a RECONSTRUCTION strategy to retrieve their original judgments their success rates did not differ whether they were in the OJ-only or OJ&CA conditions, two-sample $z = 0.35$, $p = .726$. Similarly, participants' success rates when no evidence of a strategy was provided (NO STRATEGY) did not differ across the two conditions, two-sample $z = 0.29$, $p = .772$. Participants in both conditions rarely used the ANCHOR, PRIOR KNOWLEDGE, and GUESS strategies. When these strategies were used, they were used with varying degrees of success.

Table 5

Success Rates for the Control Items, reported Percentage (N).

	OJ-only	OJ&CA		
Retrieval Strategy	Strategy used successfully (ROJ = OJ)	Strategy used successfully (ROJ = OJ)	*z*	*p*
Reconstruction	60.4% (129)	56.8% (125)	0.350	.726
Discrimination	NA	NA		
Anchor	0.0% (0)	80.0% (4)	NT	NT
Prior Knowledge	100% (8)	66.7% (2)	NT	NT
Guess	25.0% (1)	0.0% (0)	NT	NT
Multiple Strategies	0.0% (0)	0.0% (0)	NT	NT
No Strategy	56.7% (85)	58.4% (80)	0.29	.722

Note. NA = not applicable because retrieval strategy could not be used; NT = success rate differences were not statistically compared because sample sizes were too small.

The think-aloud results for the control items demonstrate that when participants engaged in the same retrieval task – recalling the control items – they used the same retrieval strategies at the same rates (see Table 4), and did so with the same level of

accuracy (see Table 5). Given that differences were not observed with the control items, any differences observed with the experimental items will likely not be due to the general frequency or accuracy of different retrieval strategies, but rather to the instructional manipulation (OJ-only vs. OJ&CA).

Experimental items. The critical predictions concern the experimental items. The first prediction concerns the frequency of strategy use for the experimental items. Specifically, participants should demonstrate more evidence of DISCRIMINATION and MULTIPLE STRATEGIES in the OJ&CA condition than the OJ-only condition. The second – and the most important – prediction tests the retrieval-based theory, and is that when participants use discriminative retrieval cues, their rates of successful recall of their original judgments should be greater than for other retrieval strategies. The final set of analyses tests the reconstructive processes proposed by the SARA, RAFT and HB13 models. These models propose that when participants are aware of the correct answers, this information will bias the reconstruction of their original judgments, leading to hindsight bias. Therefore, it is expected that when participants provide evidence of using a RECONSTRUCTION retrieval strategy, they should not demonstrate accurate recall of their original judgments.

Retrieval strategy use. Recall that participants in the OJ&CA condition were instructed to simultaneously retrieve both their original judgments and the correct answers, meaning their retrieval strategies were limited to the DISCRIMINATION strategy or MULTIPLE STRATEGIES coding. Participants were required to discriminate between their original judgments and the correct answers for all of the experimental items, and therefore to necessarily use the DISCRIMINATION strategy. In addition, they could pair this

strategy with any of the other strategies, resulting in a code representing MULTIPLE STRATEGY use (see the example in Table 2, where the participant provided evidence of using both RECONSTRUCTION and DISCRIMINATION strategies). Because participants in the OJ&CA condition were restricted to DISCRIMINATION and MULTIPLE STRATEGIES, only the frequency of usage for these strategies were compared to the OJ-only condition (see Table 6).

Table 6

Frequency of Retrieval Strategy Use for the Experimental Items, reported M (SD).

	Experimental Items			
Retrieval Strategy	OJ-only	OJ&CA	*t* (38)	*p*
Reconstruction	1.81 (2.87)	NA		
Discrimination	1.29 (1.95)	7.63 (3.06)	7.90	.001
Anchor	0.00 (0.00)	NA		
Prior Knowledge	0.14 (0.48)	NA		
Guess	0.14 (0.48)	NA		
Multiple Strategies	0.38 (0.67)	2.37 (3.06)	2.91	.006
No Strategy	6.24 (3.77)	NA		

Note. NA = not applicable because retrieval strategy could not be used.

As expected, two independent t-tests showed that there were significant differences between the OJ-only and the OJ&CA conditions with respect to their use of DISCRIMINATION, $t(38) = 7.90$, $p < .001$, $d = 2.56$, and MULTIPLE STRATEGIES, $t(38) = 2.91$, $p = .006$, $d = 0.94$. As can be seen in Table 6, and consistent with the instructional manipulation, participants in the OJ&CA condition reported using the DISCRIMINATION and MULTIPLE STRATEGIES significantly more than participants in the OJ-only condition.

Retrieval strategy success rates. The primary goal of Experiment 2 is to use the think-aloud methodology to directly investigate the critical prediction of the retrieval-based theory: participants in the OJ&CA condition, who must recall both their original

judgments and the correct answers, will be more likely to form discriminative retrieval cues that support accurate recall of their original judgments, and thus avoid hindsight bias.

I tested this prediction by examining the success rate of the recall strategies used when participants' recalled their original judgments for the experimental items (see Table 7). This analysis first compared the accuracy rates of the DISCRIMINATION strategy for the OJ-only and the OJ& CA conditions. Participants in the OJ-only condition did not use the DISCRIMINATION strategy frequently, but when they did, they successfully recalled their original judgments with a success rate of 66.7%. The participants in the OJ&CA condition were required to discriminate between their original judgments and the correct answers for all of the experimental items. Participants' success rates when using the DISCRIMINATIVE strategy in the OJ&CA condition (59.6%) did not differ significantly from that of the participants in the OJ-only condition, two-sample $z = 0.692$, $p = .490$. The high success rates observed for the DISCRIMINATION strategy across both conditions provide partial support for the retrieval-based theory. Specifically, these results suggest that when participants discriminate between their original judgments and the correct answers (even when they are not explicitly told to do so, as in the OJ-only condition) they can successfully distinguish between the respective memory traces and avoid hindsight bias. Together, the success rates of the DISCRIMINATION strategy and the frequency with which this strategy was used, even when this strategy was not prompted (OJ-only), support the alternate hypothesis from Experiments 1 and 2, that when young adult participants are not restricted by time during the retrieval phase, they spontaneously and successfully use the discrimination strategy.

Table 7

Success Rates for the Experimental Items, reported Percentage (N).

	OJ-only	OJ&CA		
Retrieval Strategy	Strategy used successfully (ROJ + DJ)	Strategy used successfully (ROJ + DJ)	*z*	*p*
Reconstruction	57.9% (22)	NA		
Discrimination	66.7% (18)	59.6% (87)	0.69	.490
Anchor	0.0% (0)	NA		
Prior Knowledge	66.7% (2)	NA		
Guess	100% (3)	NA		
Multiple Strategies	83.3% (5)	68.2% (30)	NT	NT
No Strategy	58.6% (78)	NA		

Note. NA = not applicable because retrieval strategy could not be used; NT = success rate differences were not statistically compared because sample sizes were too small.

The above results are further supported by a closer examination of the instances when participants successfully used MULTIPLE STRATEGIES. Participants in the OJ&CA condition could, and did, use MULTIPLE STRATEGIES, which again always included the DISCRIMINATION strategy. The success rates for the MULTIPLE STRATEGY use did not differ from the success rates for the use of DISCRIMINATION, two-sample $z = -1.02$, $p = .308$, for the OJ&CA condition. Additionally, it is important to note that of the 6 occurences of MULTIPLE STRATEGY use for participants in the OJ-only condition, the success rate was high (83.3%), and each reported use of multiple strategies involved using both the RECONSTRUCTION and the DISCRIMINATION strategies together.

In order to determine whether the DISCRIMINATION retrieval strategy resulted in superior recall of the original judgments, the success rates from this strategy was compared to the other more widely used strategies. The success rates from the

DISCRIMINATION strategy in the OJ-only condition was not significanlty higher than that of the RECONSTRUCTION strategy, two-sample $z = 0.72$, $p = .578$, and NO STRATEGY, two-sample $z = .78$, $p = .435$, suggesting that although the DISCRIMINATION retrieval strategy is useful in supporting accurate recall of the original judgments, it is not necessarily better than the other more widely used strategies. Participants' reported strategy use for the other categories (ANCHOR, PRIOR KNOWLEDGE, GUESS, and MULTIPLE STRATEGY use) were not frequent enough to compare statistically.

Given the current results, it is difficult to interpret the utility of the DISCRIMINATION retrieval strategy in eliminating hindsight bias because there is no true baseline success rate to which the success of a given retrieval strategy can be compared. One possibility is to use NO STRATEGY as such a baseline (the difficulties of interpretting the retrieval processes when NO STRATEGY is reported are addressed in the Discussion). As reported above, the success rate of the DISCRIMINATION strategy did not differ from that of NO STRATEGY. In other words, the utility of the discrimination retrieval strategy does not seem to confer any retrieval advantage compared to when participants do not provide evidence of using any retrieval strategy. Therefore, conclusions regarding the utility of the discrimination retrieval strategy relative to the other strategies are limited.

The SARA, RAFT, and HB13 models predict that if participants cannot recall their original judgment, they engage in a reconstructive process. When participants are aware of the correct answers, this information biases the reconstructive process (also called biased sampling and PMM (re)construction), which should result in hindsight bias. Contrary to this prediction, when participants engaged in a reconstructive process during recall it did not always lead to hindsight bias. When participants in the OJ-only condition

used the RECONSTRUCTION strategy they successfully recalled their original judgments 57.9% of the time. The success rate did not differ from control to experimental items, two-sample $z = 0.23$, $p = .818$. Also, the success rate of the RECONSTRUCTION strategy did not differ from NO STRATEGY for the experimental items, $z = 0.08$, $p = .936$, providing further support for the surprising finding – from the perspective of the SARA, RAFT, and HB13 models – that engaging in reconstructive retrieval processes can lead to successful retrieval of the original judgments.

Discussion

Experiment 2 addressed research goal (1), which was to test the discriminative retrieval cue mechanism proposed by the retrieval-based theory. To achieve this goal, participants were tested individually, under the strict confines of the laboratory, while engaging in a think-aloud as they recalled their original judgments (and the correct answers). According to the retrieval-based theory, when participants are asked to recall their original judgments and the correct answers during Phase 3, they will form discriminant retrieval cues that differentiate their original judgments from the correct answers, thus eliminating hindsight bias. This proposal was partially supported by the hindsight bias and think-aloud data. Participants in the OJ&CA condition used the DISCRIMINATION retrieval strategy at greater rates than participants in the OJ-only condition, and were able to successfully avoid hindsight bias by demonstrating accurate recall of their original judgments. However, participants in the OJ-only condition also demonstrated accurate recall of their original judgments. Why then were they able to avoid hindsight bias?

One hypothesis is that when participants in the OJ-only condition reported discriminating between their original answer and the correct answer during Phase 3, they would be more likely to correctly recall their original judgment, than in instances where no such discrimination took place. This hypothesis was not supported. When participants in the OJ-only condition reported using a DISCRIMINATION strategy, their original judgment success rate (66.7%) was not significantly different from the success rate for participants in the OJ&CA condition. Additionally, when participants in the OJ-only condition reported using MULTIPLE STRATEGIES, the strategies used together were always DISCRIMINATION and RECONSTRUCTION, and resulted in high rates of successful recall (83.3%), providing further support for the retrieval-based theory. The success rates for DISCRIMINATION strategy did not differ from those of RECONSTRUCTION, or even when no evidence of a retrieval strategy (NO STRATEGY code) was provided.

Taken together, the results from the think-aloud provide partial support for the retrieval-based theory. When participants demonstrated using a retrieval strategy where they discriminated between their original judgments and the correct answers (DISCRIMINATION) they were able to successfully recall their original judgments, thus avoiding hindsight bias. However, these success rates were not significantly higher than the other retrieval strategies used, including instances where no strategies were reported. Further, young adult participants were able to successfully use the DISCRIMINATION strategy whether they were instructed to do so (OJ&CA) or not (OJ-only).

Unlike participants in the OJ&CA condition, who were required to (at the very least) use the DISCRIMINATION retrieval strategy for the experimental items, participants in the OJ-only condition provided evidence of using a recall strategy in only 77 out of the

possible 210 thought units for the experimental items. Participants' infrequent reporting of their retrieval strategy use poses a power issue for analyses aimed at making comparisons across the various strategies. This is a limitation of this experiment; it is likely that the low frequency of the reported retrieval strategy use made it difficult to compare the effectiveness of the various strategies.

To avoid this limitation, think-aloud practice tasks were used to improve participants' ability to think aloud while working on a retrieval task (i.e. "Name the state capitals that begin with the letter 'B'."). However, two practice tasks may not have been sufficient. It is also possible that some of the recall strategies participants used when making their ROJs were automatic processes, which restricted their ability to verbalize their recall process (Ericsson & Simon, 1993).

Based on past research, it was expected that there would be instances when participants' think aloud would not provide evidence of using a recall strategy. Jonsson et al., (2014) asked participants to articulate the memory strategies they were using when attempting to recall a story. To support the participants in describing their strategies, they were given a list of possible strategies. Even with this support, 22% of the time no strategy use was reported. Therefore, it was expected that even with clear instructions and multiple practice tasks, participants would occasionally fail to provide evidence of using a recall strategy when retrieving their original judgment. Note that when participants do not provide evidence of using a recall strategy during their recall, it does not mean that one was not used (Goodwin et al. , 2001). It may have been that participants chose not to verbalize the strategies being used, or that the strategy used is so automatized that it was

not possible for participants to verbalize it (Ericsson & Simon, 1993; Jonsoon et al., 2014). This limitation is further addressed in the General Discussion.

Experiment 2 replicated the findings of Experiment 1 again with young-adult participants, and extended this work by identifying the recall strategies participants use when they avoid engaging in hindsight bias. As in Experiment 1, participants were able to accurately recall their original judgments regardless of the condition. Participants in the OJ&CA condition were also able to accurately recall the correct answers. In Experiment 2 participants were tested individually, under the strict confines of a laboratory, while participants in Experiment 1 were tested in a group setting. Even with these very different testing conditions, participants in the OJ-only condition did not engage in hindsight bias. Therefore, I can rule out the possibility that this unexpected finding was due to the lax testing conditions used in Experiment 1. The possibility that the results of the present experiment were an artifact of the think-aloud procedure could also be ruled out because the results from Experiment 2 did not differ from those observed in Experiment 1.

Given that participants were once again able to avoid hindsight bias using the standard memory-design, the think-aloud protocol was employed to identify what strategies participants were using to support their recall of their original judgments. Experiments 1 and 2 provided partial support for the sufficiently discriminative retrieval cue mechanism proposed by the retrieval-based theory of hindsight bias. They demonstrated through the use of a think-aloud protocol, that when participants used compound retrieval cues they were able to successfully discriminate between their original judgments and the correct answers. Young adult participants were observed

using compound retrieval cues to recall their original judgments even when they were not prompted to do so.

Chapter 6: Experiment 3

Experiment 3 addressed research goals (2) and (3). It investigated the passage of time as a factor that may minimize the ability to selectively discriminate between retrieval cues for participants' original judgments and the correct answers. In addition, it evaluated whether engaging in repeated recall increases, decreases, or has no effect on hindsight bias. To address these goals, Experiment 3 included a *test* condition that was similar to the OJ&CA condition from Experiments 1 and 2 (without the think-aloud). It also included a second retrieval phase one month later (Phase 4, see Figure 7). This delayed test was relevant for the research goal (2) of investigating whether the passage of time minimizes the ability to selectively discriminate between original judgments and correct answers. The *no-testing* condition differed from the *test* condition in that participants did *not* experience Phase 3 on the initial testing day. This condition was relevant for the research goal (3) of investigating whether engaging in repeated recall – specifically, the recall test during Phase 3 on the initial testing day – leads to better performance (i.e., less hindsight bias) during the recall test one month later, during Phase 4.

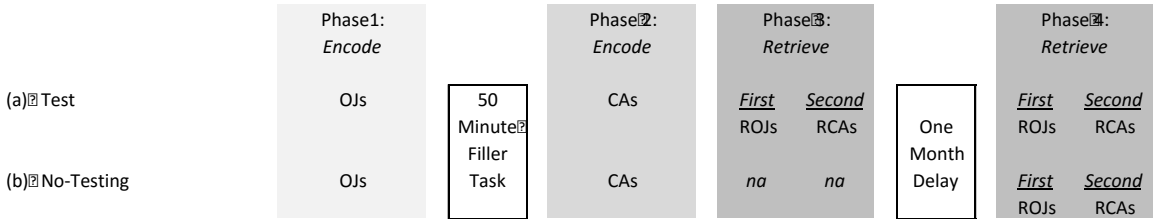

Figure 7. Design and procedure for Experiment 3. OJ = original judgments; CA = correct answers; ROJ = recall original judgments; RCA = recall correct answers; *na* = no retrieval activity.

Based on Van Boekel et al.'s (2016) findings in Experiments 3 and 4 and the results of the current Experiments 1 and 2, I hypothesized that during Phase 3, the first retrieval phase, participants in the test condition would demonstrate accurate recall of their original judgments (no hindsight bias) and the correct answers. Although the retrieval phase in the present experiment was slightly different than Experiment 1 (successive rather than simultaneous recall of original judgments and correct answers), the same explanation for why hindsight bias should be mitigated applies here too. Specifically, because participants were told that they would be recalling both their original judgments and correct answers, they would be able to form discriminative compound cues and selectively retrieve their original judgments and the correct answers (rather than conflate the two).

The new retrieval phase, which occured after a one-month delay (Phase 4), was critical for addressing research goal (2), specifically whether participants' ability to discriminate between the retrieval cues for their original judgments and the correct answers diminishes over time, thus making the correct answer traces as likely to be retrieved as the original judgment. In fact, it is possible that the correct answer traces may be more readily activated because the retention of this information is desirable (Shah & Oppenheimer, 2009). I therefore hypothesized that during Phase 4, participants from both conditions would demonstrate accurate recall of the correct answers. Given the prediction that participants will retain accurate recall of the correct answers, and the likelihood that the discriminativeness of the memory traces would diminish over time (Howe, 1998), it is hypothesized that after a delay of one month, hindsight bias will be observed for participants in both conditions. However, it is hypothesized that the magnitude of

hindsight bias observed will differ across the two conditions. This hypothesis relates to research goal (3): to evaluate whether engaging in repeated retrieval is partially responsible for the continued elimination of hindsight bias.

For participants in the test condition, the testing effect literature (Roediger & Butler, 2011; Roediger & Karpicke, 2006b; Roediger & Pyc, 2012) suggests that the process of engaging in repeated retrieval (during Phase 3) may enhance memories for the recalled original judgments. Therefore, for them, hindsight bias might continue to be reduced during Phase 4, if not eliminated. If participants' ability to discriminate between their retrieval cues is supported by repeated retrieval, then there are two competing hypotheses regarding the magnitude of the hindsight bias that will be observed. The first hypothesis is that hindsight bias would continue to be completely eliminated after a one-month delay. The second hypothesis is that hindsight bias would only be partially reduced after a one-month delay, meaning participants in the test condition will engage in hindsight bias during Phase 4, but to a lesser extent than particpants in the no-testing condition.

Method

Participants

Fifty-four undergraduates from the University of Minnesota were recruited from four recitation sections of two introductory educational psychology classes. Participants received course credit for their participation.

Materials

The materials used in Experiment 3 were identical to those used in Experiment 1 with the exception that during Phases 3 and 4, participants used a two-sided response

sheet (see Figure 8c). One side prompted for just the 20 original judgments made during Phase 1, similar to the procedure for the OJ-only condition in Experiments 1 and 2, which is standard for memory-design experiments. The other side prompted for just the 10 correct answers provided during Phase 2.

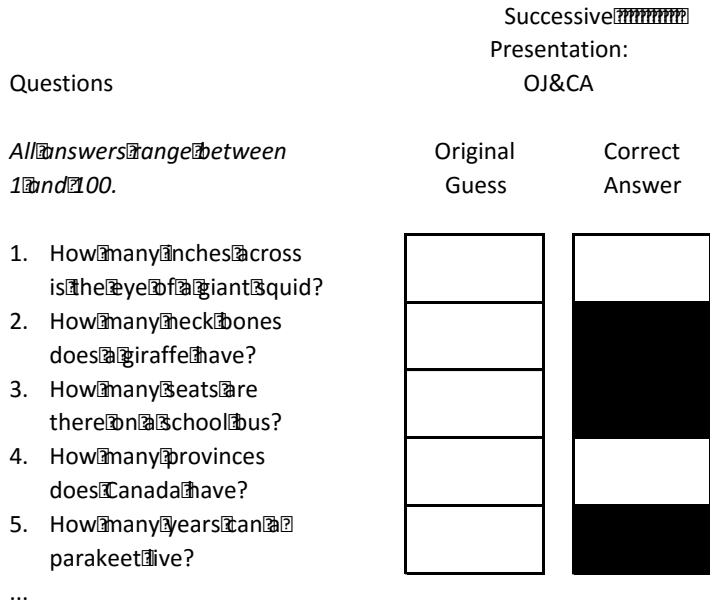

Figure 8. Fragments of the response sheet used in Experiment 3. The filled boxes are for the control questions, for which the correct answers were not provided during Phase 2.

Design

The experiment used a mixed-design. The within-subjects factor, *Type*, had two levels (experimental and control). During Phase 2 participants were provided with the answers to 10 questions (experimental level), but not for the remaining 10 questions (control level). The between-subjects factor, *Test Timing*, was manipulated during Phases 3 and 4 (see Figure 7), and had two conditions (test and no-testing). Two recitation sections were assigned to each of the conditions to form two groups of comparable sizes. The test condition (n = 31) required participants to recall their original judgments and the correct answers during Phase 3, and again after a one-month delay (Phase 4). The no-

testing condition (n = 23) did not include Phase 3; participants were only required to recall their original judgments and the correct answers after the one-month delay.

The two dependent variables used in Experiments 1 and 2, HBI and CAI, were used in the present experiment. These indices were computed for the values obtained during the first retrieval task (Phase 3) for participants in the test condition, and during the retrieval task (Phase 4) one month later for participants in both conditions.

Procedure

The same procedure for Phases 1 and 2 of Experiment 1 was used in the present experiment. During Phase 3, participants in the test condition were asked to recall both their original judgments and the correct answers. They were given a two-sided response sheet. They were told that they would first be recalling their original judgments from Phase 1, followed by the correct answers provided during Phase 2. Participants first completed the side of the response sheet that prompted for recall of their original judgments. Only when all participants finished recalling their original judgments were they told to turn over their response sheets and to complete the side that prompted them to recall the correct answers. Participants were told *not* to turn over their worksheets and return to the first page at any point.

After a one-month delay participants from both conditions completed Phase 4. At the beginning of their class period, they were given a "surprise memory task." The materials and procedure were the same as those used in Phase 3.

Results

Based on the findings of Experiments 3 and 4 of Van Boekel et al. (2016) and the retrieval-based theory, it was hypothesized that during Phase 3, participants in the test

condition would demonstrate accurate recall of their original judgments (no hindsight bias) and the correct answers. This hypothesis was supported. Participants in the test condition accurately recalled their original judgments from Phase 1, showing no evidence of hindsight bias (see Table 7), one-sample $t(30) = 0.10$, $p = .92$. Also as predicted, participants accurately recalled the correct answers from Phase 2, as indicated by a CAI that differed reliably from zero in the positive direction, one-sample $t(30) = 9.05$, $p < .001$, $d = 3.30$.

Table 7

Results of Experiment 3 reported M (SD).

	Phase 3		Phase 4	
	No-Test	Testing	No-Test	Test
z_{CON}		0.03 (0.09)	-0.08 (0.09)	-0.10 (0.25)
z_{EXP}		0.03 (0.10)	0.31 (0.22)	0.12 (0.14)
z_{CA}		0.81 (0.48)	0.23 (0.19)	0.36 (0.33)
HBI		0.00 (0.11)	0.35** (0.26)	0.23** (0.33)
CAI		0.78** (0.48)	0.37** (0.36)	0.46** (0.35)

Note. HBI and CAI statistically tested against 0. **$p < .001$.

A second hypothesis was that after a one-month delay participants in both conditions would exhibit accurate recall of the correct answers. This hypothesis was also supported. Participants in the test (one sample $t(30) = 7.32$, $p < .001$, $d = 2.67$) and no-testing (one sample $t(22) = 4.86$, $p < .001$, $d = 2.07$) conditions had CAIs that were significantly different from zero in the positive direction (see Table 7).

Three hypotheses were made regarding the presence of hindsight bias after a one-month delay. First, it was hypothesized that participants in the no-testing condition would demonstrate hindsight bias. This hypothesis was supported, one sample $t(22) = 6.46$, $p <$

.001, d = 2.75. The final two hypotheses involved the test condition. Participants in the test condition engaged in repeated retrieval practice (Phases 3 and 4), which may have enhanced the memories for their original judgments made during Phase 3 (Roediger & Butler, 2011), resulting in a continued elimination or reduction of hindsight bias. Specifically, it was hypothesized that participants in the test condition may either continue to display accurate recall of their original judgments, or they may engage in hindsight bias, but to a lesser extent than participants in the no-testing condition. Neither of these hypotheses was confirmed. Participants in the test condition engaged in hindsight bias after the one-month delay, one sample $t(30) = 3.88$, $p < .001$, $d = 1.42$. Even though participants in the test condition had a lower average HBI ($M = 0.23$, $SD = 0.33$) than participants in the no-testing condition ($M = 0.35$, $SD = 0.26$), this difference was not significant, independent samples $t(52) = -1.47$, $p = .147$. A similar pattern was found for the correct answers, where participants in the test condition had a higher average CAI ($M = 0.46$, $SD = 0.35$) than participants in the no-testing condition ($M = 0.37$, $SD = 0.36$), but again, these differences were not significant, independent samples $t(52) = 0.95$, $p = .348$.

By calculating the CAI, I was able to investigate the assumption that hindsight bias occurs due to retroactive interference from the correct answers which disrupts the recall of participants' original judgments. This assumption implies that the CAI and HBI should be positively correlated because both are driven by the encoding of correct answers. That is, when the correct answers are strongly encoded, the chance that they will be accurately recalled increases. This enhanced recall accuracy is also likely to retroactively interfere with retrieval of the original judgments and cause hindsight bias. This assumption was not supported during Phase 3 for the test condition, $r(29) = 0.03$, p

= .89. However, this assumption was supported during Phase 4 for both the test, $r(29) = 0.45$, $p < .01$, and the no-testing conditions, $r(29) = 0.48$, $p < .05$.

Discussion

Experiment 3 addresses research goals (2) and (3), investigating the effect of longer retention intervals on the presence of hindsight bias and the possible mitigating role of repeated recall on re-appearance of hindsight bias, respectively. During Phase 3, participants in the test condition were able to accurately recall both their original judgments and the correct answers when prompted to do so successively. This finding is consistent with the retrieval-based approach, which predicts that when participants are instructed at the beginning of Phase 3 that they will be recalling both sets of memories, they form compound retrieval cues that discriminate original judgment traces from correct answer traces, and in this way avoid hindsight bias.

After a one-month delay, participants in the test condition completed a second retrieval phase. As hypothesized, they maintained accurate recall of the correct answers. Also as hypothesized, participants demonstrated hindsight bias. These results suggest that people's capacity to selectively discriminative between retrieval cues for their original judgments and the correct answers deteriorates over time. This finding has important implications for the retrieval-based theory by identifying the passage of time as a factor that constrains the utility of the discriminative retrieval cue mechanism.

To investigate research goal (3) – whether engaging in recall during Phase 3 helped to mitigate the onset of hindsight bias during Phase 4 – the HBI from the test condition was compared to the HBI from the no-testing condition. Although the differences were not significant, participants from the test condition had a lower HBI and

higher CAI than participants in the no-testing condition. These results indicate that engaging in repeated recall (2 times rather than 1) did not enhance participants' memory for their original judgments or their correct answers. Importantly, these findings connect the study of hindsight bias and the testing effect, and demonstrate that the benefits typically associated with repeated retrieval observed in testing effect research do not extend to the memory-design paradigm in hindsight bias.

Two factors possibly explain why the effect of repeated retrieval, though in the right direction, was not significant: the experimental design and sample size. The experimental design used in the present study diverged in one critical way from the traditional method used to study the impact of the testing effect. Experiment 3 required participants to recall two sets of memories, their original judgments and the correct answers, in contrast with the traditional method requiring recall of just one set of memories, their original judgments. According to the retrieval-based theory, successful participants would have to create compound retrieval cues to support the discrimination between these two types of memories. It is possible that over time, the retrieval cues used to selectively retrieve the original judgments and correct answers lose their capacity to maintain the discrimination, leading to the onset of hindsight bias.

In fact, there is evidence to support this claim. Recall that both the CAI and HBI are driven by the encoding of the correct answers. If the correct answers are strongly encoded, then the chances of these memories interfering with the recall of the original judgments increases (Van Boekel et al., 2016). Examining the correlations from Phase 3 and Phase 4 for participants in the test condition, it can be seen that the accurate memories for the correct answers (CAI) during Phase 3 remain strong one month later

during Phase 4, $r(29) = .50$, $p < .05$. However, participants' memories for their original judgments (HBI) do not remain consistent across the same retention interval, $r(29) = .17$, $p < .36$. That is, participants who demonstrated better recall of their original judgments during Phase 3 did not necessarily have better recall of their original judgments after the one-month delay. Therefore, it is possible that repeated retrieval helped bolster participants' memories for the correct answers, thereby undermining and interfering with the discriminant retrieval cues formed during Phase 3 through retroactive interference.

The second explanation is that the inability to detect significant differences between the test and no-testing conditions was due to insufficient power. The observed average HBIs and CAIs are consistent with the hypotheses suggested by the testing effect literature: the HBI was lower for the test condition than the no-testing condition, and the CAI was higher for the test condition than the no-testing condition (see Table 7). However, these differences were not significant. The sample sizes for the two conditions in the present study, test ($n = 31$), and no testing ($n = 23$), were similar to the average sample size ($n = 30.15$) for memory-design studies with almanac-style questions reported in Christensen-Szalanski and Willham's (1991) meta-analysis. However, little is known about the fate of hindsight bias as the time following the first retrieval phase (Phase 3) extends beyond one week (Pohl & Hell, 1996), and therefore it is perhaps not surprising that using the meta-analysis as a guide to determine the sample size for the present study resulted in an underestimation of the power required to detect an effect for retrieval periods spanning one month. In traditional testing effect studies, when longer retention intervals are used, the benefits gained from repeated testing are smaller than the effects that are observed when shorter intervals are used (Carpenter et al., 2009). Given the long

retention interval and the complex retrieval task, the inability to detect differences due to repeated recall across the two conditions may be due to the small sample size.

Chapter 7: General Discussion

The dominant memory-based theories of hindsight bias have identified the importance of the retrieval and reconstruction processes involved in people's recall of their original judgments (Erdfelder & Buchner, 1998; Erdfelder et al., 2007; Hoffrage et al., 2000; Pohl et al., 2003). These theories have privileged the reconstruction process as a critical factor in the presence of hindsight bias. In contrast, the retrieval-based theory focuses on the retrieval process, and is therefore capable of making predictions regarding conditions under which people's original judgments can be recalled accurately (Van Boekel et al., 2016). Specifically, the retrieval-based theory proposes that when participants create sufficiently discriminative retrieval cues, they will be able to selectively access the memory traces for their original judgments and for the correct answers. Experiments 1 and 2 provide findings that partially support the sufficiently discriminative retrieval cue mechanism proposed by the retrieval-based theory. Experiment 3 further adds to our understanding of this mechanism by identifying the passage of time as a factor that constrains the utility of this mechanism in reducing hindsight bias. This chapter first unpacks the results of the three experiments. It then considers the theoretical implications of these findings, and then outlines the limitations of the present set of studies and proposes future areas of research. It concludes by exploring the practical implications of the current findings.

Review of Research Goals and Findings

Research goal (1). The objective of the present study was to investigate factors that reduce or eliminate hindsight bias. In order to better understand why these factors may successfully eliminate hindsight bias, I tested the discriminant retrieval cue

mechanism proposed by the retrieval-based theory. Experiments 1 and 2 partially replicated Van Boekel et al.'s (2016) Experiments 1 and 2 with young adult participants. They found that when participants were asked to recall their original judgments and the correct answers, they could do so accurately, avoiding hindsight bias. However, contrary to expectations, when participants were only asked to recall their original judgments, as is the standard procedure in memory-design studies, participants avoided hindsight bias. This stands in stark contrast to Van Boekel et al.'s Experiment 1, where middle-school participants engaged in hindsight bias using the same materials and procedure. It is also surprising given the ubiquity of hindsight bias in memory-design studies (Christensen-Szalanski & Willham, 1991; Guilbault et al., 2004).

In Experiment 2, a think-aloud method was used during Phase 3 to gain insight into the recall strategies participants were using when they recalled their original judgments. Participants rarely provided evidence of using a retrieval strategy as they recalled their original judgments. However, when participants verbalized discriminating between their original judgment and the correct answer, or jointly using a discrimination and reconstruction strategy, they were able to recall their original judgment with high accuracy rates. These findings support the discriminative retrieval cue mechanism proposed by the retrieval-based theory.

The contrast between the finding of hindsight bias in Experiment 2 of Van Boekel et al. (2016) and the failure to find hindsight bias in the control conditions of the current Experiments 1 and 2 can be phrased developmentally: Why do middle-school participants completing the same memory-design task engage in hindsight bias, while young adult participants do not? The think-aloud data of Experiment 2 provide evidence that young

adults utilize a variety of retrieval strategies successfully when recalling their original judgments, particularly the spontaneous use of discrimination. It may be the retrieval strategies that middle-school students use are less sophisticated than those used by young adults. Evidence supporting this claim comes from the source monitoring literature. Adolescents and young adults typically perform similarly on source monitoring tasks (Sprondel et al., 2011). However, the neural correlates activated when engaging in a source monitoring task differ for adolescents and young adult (Cycowicz et al., 2001; Sprondel et al., 2011). These results suggest that as people's brains develop, their capacity to engage in source monitoring becomes more refined, even though this continued refinement may not be observed in the behavioral data (Sprondel et al., 2011). Returning to hindsight bias, it is possible that the retrieval strategies implicitly used by adolescents are less refined or less sophisticated than those used by young adults. When adolescents use less sophisticated retrieval strategies, they are less likely to discriminate between the memory traces for their original judgments and the correct answers. However, when adolscents are directed to use a discrimination strategy, they are able to do so successfully (Van Boekel et al., 2016). In contrast, it can be seen from the think-aloud results that young adults can spontaneously use a variety of effective retrieval strategies, including discriminating between their original judgments and the correct answers, whether they were instructed to do so or not.

Research goal (2). Given that the creation of discriminative retrieval cues supports participants' ability to selectively retrieve their memory traces for their original judgments and the correct answers, it is logical to ask about the fate of these retrieval cues over time. In Experiment 3, participants in the test condition were asked to

successively recall their original judgments and the correct answers, and demonstrated accurate recall of each, avoiding hindsight bias. Participants were tested again after a one-month delay and engaged in hindsight bias, although they retained their memory for the correct answers. These results suggest that the creation of discriminative retrieval cues are useful over the short term in eliminating hindsight bias, but with the passage of time, they lose their capacity to maintain discrimination, leading to the onset of hindsight bias.

It is possible that the utility of discriminative retrieval cues formed during Phase 3 may have been undermined by retroactive interference from better long-term retention of the correct answers. If so, as the ability to discriminate between the memory traces fades, the probability of interference from related traces increases (Howe, 1998). Therefore, as time passes and participants' ability to discriminate between their original judgments and the correct answers fade, the likelihood that the well-remembered correct answers interfere with the retrieval of the original answers increases, resulting in hindsight bias.

Research goal (3). It is possible that even though participants engaged in hindsight bias after a one-month delay, the act of engaging in repeated retrieval of their original judgments and the correct answers (i.e., during Phase 3) supports retention of these memories to some degree (Roediger & Butler, 2011; Roediger & Karpicke, 2006b; Roediger & Pyc, 2012). This enhanced retention due to repeated retrieval was not supported by the results of Experiment 3. Participants' memory for their original judgments and the correct answers were the same whether they engaged in repeated retrieval (test condition) or only engaged in retrieval a single time (no-testing condition).

Although no statistical differences were observed between the two conditions, it is important to highlight that the observed trends were consistent with the results from the testing effect research. That is, participants in the test condition demonstrated greater mean accuracy in recalling their original judgments (lower HBI) and the correct answers (higher CAI) than participants in the no-testing condition. It is possible that the sample size may have led to insufficient power to detect the differences between the two conditions. Even though the sample sizes were consistent with past studies (Christensen-Szalanski & Willham, 1991; Van Boekel et al., 2016), the inclusion of the passage of time as a variable of interest may have required additional participants to detect a smaller effect.

Together these findings suggest that discriminant retrieval cues may help participants avoid hindsight bias over the short term, but these benefits diminish over time. They also suggest that engaging in repeated recall in the memory design procedure used in this set of experiments may not confer the same advantages observed in the testing effect literature.

Implications for Memory-Based Theories of Hindsight Bias

The series of experiments reported in this dissertation make several important contributions to our understanding of hindsight bias, specifically in terms of the retrieval-based theory. By asking participants to think aloud while recalling their original judgments, Experiment 2 provides partial support for the proposed discriminant retrieval cues mechanism proposed by the retrieval-based theory. Participants demonstrated high recall accuracy rates when they reported discriminating between their original judgment

and the correct answer, and when they reported such a discrimination in conjunction with other recall strategies.

The results from the think-aloud data are also germane for the other memory-based models of hindsight bias – the SARA, RAFT, and HB13 models. These models propose that participants' accurate recall of their original judgments is due to the processes at work during retrieval. Specifically, these models predict that a participant's memory trace for their original judgment will be retrieved when the similarities between this trace and the retrieval cues in working memory are high, and will not be retrieved when there are greater similarities between the correct answer trace and the retrieval cues. When participants cannot retrieve their original judgments, they engage in a reconstructive process, and when participants have knowledge of the correct answers, this reconstructive process will result in hindsight bias. Contrary to this last proposal, participants in the OJ-only condition from Experiment 2 were observed successfully employing a reconstruction strategy when recalling their original judgments for the experimental items. In fact, their accuracy rates when using this strategy were not significantly different from instances when this strategy was used for recalling the control items. This finding is difficult for SARA, RAFT, and HB13 to account for.

According to the HB13 model, the reconstruction process can occasionally lead to accurate reconstruction of the original judgment, even when potentially biased by knowledge of the correct answers (Erdfelder & Buchner, 1998). However, HB13 does not describe the mechanisms supporting this accurate reconstruction in some instances but not others. Similarly, according to the SARA and RAFT models, the correct answers are automatically and unconsciously integrated into the image set or probabilistic mental

model and should therefore factor into the reconstruction process in some way (Blank & Nestler, 2007; Hertwig et al., 2003; Hoffrage et al., 2000; Pohl et al., 2003). Like HB13, these models do not provide a description of the factors that enable successful reconstruction in some situations, and not others. The findings from the think-aloud study suggest that the SARA, RAFT and HB13 models need to be expanded to include the factors at play when participants are able to successfully engage in reconstruction of their original judgments when the correct answers are known.

Limitations and Future Directions

In Experiments 1 and 2, participants successfully recalled their original judgments whether they were given instructions to discriminate between their correct answers and their original judgments or not. That participants were able to avoid hindsight bias when not asked to retrieve both their original judgments and the correct answers was surprising, and stand in contrast to the results of Van Boekel et al.'s (2016) Experiments 1 and 4. Calvillo (2013) observed a similar result with a young adult population. In that study, the participants who were able to avoid hindsight bias were given an unlimited amount of time to make their recall judgments, whereas participants who were constrained by a time limit engaged in hindsight bias. Therefore, I proposed that young adults, and not middle-school participants, are able to avoid hindsight bias because when they are unconstrained by a time limitation during retrieval they can spontaneously create compound retrieval cues that help them discriminate between their original judgments and the correct answers. This hypothesis was supported by the Experiment 2 finding of young adults successfully using a discrimination recall strategy even when they were not prompted to do so. However, Experiment 2 did not include a middle-school sample to investigate their

use of recall strategies. Therefore, a follow-up study should include a middle-school comparison group to explore the differences in recall strategy use between young adults and middle-school participants.

Experiment 2 provides several lessons about the design of future think-aloud studies of hindsight bias. Participants rarely provided evidence of using a retrieval strategy while thinking aloud: 68.9% of the thought units were coded as NO STRATEGY when that was a viable option. This low frequency of strategy reporting limited the power of this design to detect differences between the effectiveness of the various recall strategies.

Future think-aloud studies could potentially improve their power in three ways. First, participants may provide a more detailed think-aloud if they are asked to think aloud retrospectively, as opposed to concurrently as they did in Experiment 2. For example, McCabe et al. (2011) had participants complete a remember-know task, indicating if a word presented during testing was previously studied in an earlier word list (an "old" word) or not (a "new" word). Once the remember-know task was completed, participants were presented with the words they had identified as "old", and asked to retrospectively think-aloud and explain why these items were identified as "old". Similar to Experiment 2's NO STRATEGY code, participants occasionally provided a response that contained no explanation (e.g. "I don't remember what I said."), but unlike Experiment 2, these verbalizations occurred much less frequently than the other categories. It is possible that having participants retrospectively think-aloud about their responses provided during

Phase 3 may result in participants providing more detailed responses when recalling their original judgments.[11]

Second, participants' identification of the retrieval strategies they used may be supported through the use of a document listing possible retrieval strategies. Jonsson et al. (2014) presented participants with five retention strategies that may be used when remembering portions of a text (Repetition, Visualization, Elaboration, Multiple Strategies, No Strategy) in an effort to support their reporting of their retention strategy use. Only 22% of participants reported No Strategy, a rate substaintially lower than the 68.9% observed in the present Experiment 2. It was not possible to use a document outlining alternative retrieval strategies in Experiment 2 because Experiment 2 was the first study to examine hindsight bias using a think-aloud protocol. In addition, inadequate guidance was available from current models of hindsight bias, which provide little detail about their proposed retrieval mechanisms (Blank & Nestler, 2007; Erdfelder & Buchner, 1998; Hertwig et al., 2003; Hoffrage et al., 2000; Pohl et al., 2003). Therefore, prior to running Experiment 2, any list of participants' potential retrieval strategies would likely have been biased or incomplete. In future studies, using a response sheet based on the retrieval strategies observed in Experiment 2, either concurrently (as participants retrieve their original judgments during Phase 3) or retrospectively (after they have retrieved all

[11] There is one important difference between the think-aloud procedure used by Mc Cabe et al. (2011) and those used in Experiment 2. McCabe et al. asked participants to *explain* the reasons for their responses. Asking participants to explain or describe their thoughts alters the retrieval process (Fox et al., 2011). Because McCabe et al. were interested in participants' *explanations* they were required to use a retrospective think-aloud process. This differs from the procedure used in Experiment 2, where participants were explicitly told not to explain their thoughts, but encouraged to say everything that they were thinking.

of their original judgments, and Phase 3 has ended) may help to reduce instances of participants reporting NO STRATEGY.

A related point regards the development of retrieval strategies. Given that adolescents are likely to use less sophisticated retrieval strategies than the young adults in the present studies (Sprondel et al., 2011), future research should investigate whether the retrieval strategies identified in Experiment 2 are also reported by adolescents. Once a retrieval strategy document is established that can be used with both adolescent and adult populations, future research could cross-sectionally examine developmental differences in participants' retrieval strategy use.

Third, it is possible that participants did not provide elaborate think-aloud reports because the recall strategies that they were using were so automatized that it was difficult to articulate them (Ericsson & Simon, 1993; Jonsson et al., 2014). Therefore, in order to detect differences among recall strategies more items may be required. Increasing the number of the almanac-style items (from 20 to 40, for example) may be the best solution to increase the overall power of the study because increasing the amount of information participants are required to remember would increase task difficulty. This in turn, would make it harder for participants to engage in a simple retrieval process when recalling their original judgments. The need to use more elaborate retrieval strategies should increase the number of different retrieval strategies verbalized, reduce the number of NO STRATEGIES, and improve the overall power of the analysis.

Experiment 3 also likely suffers from an issue of power. It was hypothesized that repeated retrieval would support participants' accurate recall of the correct answers and original judgments, mitigating but not completely fending off the onset of hindsight bias

after a one-month delay. Consistent with this prediction, participants who engaged in repeated retrieval did have a larger correct answer index and lower hindsight bias index than the control group, but these differences were not significant. The advantage for repeated retrieval over a long retention interval is expected to be small (Carpenter et al., 2009), therefore the small sample sizes in the two groups may have masked the effect of repeated retrieval.

Implications for Practice

Hindsight bias is a robust phenomenon that is resilient to efforts to reduce the effect (Guilbault, et al., 2004), and there is evidence that its effects can be long-lasting (Experiment 2; Pohl & Hell, 1996). Under the conventional view, hindsight bias has detrimental impacts on learning (Kerr, 1998; Roese & Vohs, 2012; Slovic & Fischhoff, 1977). An alternative view, however, is that hindsight bias may be adaptive, a feature of an efficient memory system (Bernstein et al., 2011; Bradfield, & Wells, 2005; Hawkins & Hastie, 1990; Hertwig et al., 2003; Hoffrage et al., 2000).

The present set of experiments provides evidence that hindsight bias may not be as detrimental as it has traditionally been portrayed. Participants in all three experiments demonstrated accurate recall of their original judgments, as well as the correct answers over the short term. These findings suggest that in scenarios where hindsight bias has traditionally been observed, the original judgments may be accessed by simply supporting the discrimination between the original judgments and the correct answers.

In terms of scientific reasoning, it is important that learners be able to access their original hypotheses so that those hypotheses can be accurately critiqued in light of unexpected results. More generally, it is important for budding scientists to be able to let

go of inaccurate information in favor of credible scientific evidence. Indeed, three of the five facets of scientific reasoning proposed by Varma and colleagues (Varma et al., 2013) are related to evaluating potentially contradictory findings: reasoning from evidence to conclusions, providing explanations, and coordinating theory and evidence. An interesting line of future research should examine the possibility that hindsight may be adaptive, by investigating the conditions where engaging in hindsight bias may interfere with or facilitate learning (Bernstein et al., in press), particularly in the context of scientific reasoning.

Conclusion

In conclusion, the results of the three experiments presented in this dissertation provide support for the retrieval-based theory of hindsight bias. Specifically, they demonstrate that when participants form compound retrieval cues, they are able to discriminate between their original judgments and the correct answers, avoiding hindsight bias. The unexpected finding that participants engaging in the standard memory-design procedure in Experiments 1 and 2 were able to avoid hindsight bias highlights the possibility of developmental differences in participants' natural tendency to create these compound retrieval cues. Experiment 3 suggests that the passage of time may restrict the utility of compound retrieval cues in reducing hindsight bias. Importantly, the present research helps us better understand the factors that constrain hindsight bias. This is important because it informs competing cognitive theories of this effect, and because it has the potential to inform the design of science instruction that minimizes hindsight bias.

References

Anderson, J., Lowe, D., & Reckers, P. (1993). Evaluation of auditor decisions: Hindsight bias effects and the expectation gap. *Journal of Economic Psychology, 14*(4), 711-737.

Anderson, K. (2014). The effects of hindsight bias on experienced and inexperienced auditors' relevance ratings of adverse factors versus mitigating factors. *Journal of Business & Economics Research, 12*(3), 199-208.

Arkes, H. R. (2013). The consequences of the hindsight bias in medical decision making. *Current Directions in Psychological Science, 22*(5), 356-360.

Arnold, M., & Lindsay, D. S. (2007). "I remember/know/guess that I knew it all along!": Subjective experience versus objective measures of the knew-it-all-along effect. *Memory & Cognition, 35*(8), 1854-1868.

Ash, I. K., & Wiley, J. (2008). Hindsight bias in insight and mathematical problem solving: Evidence of different reconstruction mechanisms for metacognitive versus situational judgments. *Memory & Cognition, 36*(4), 822-837.

Bârliba, R., & Dafinoiu, I. (2015). The hindsight bias effect and counterfactual thinking: Clinical predictors. *Journal of Evidence-Based Psychotherapies, 15*(1), 121-133.

Bayen, U., Erdfelder, E., Bearden, N., & Lozito, J. (2006). The interplay of memory and judgment processes in effects of aging on hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1003-1018.

Bayen, U., Pohl, R., Erdfelder, E., & Auer, T. (2007). Hindsight bias across the life span. *Social Cognition, 25*(1), 83-97.

Bernstein, D., Aßfalg, A., Kumar, R., & Ackerman, R. (in press). Looking backward and forward on hindsight bias. In J. Dunlosky & S. Tauber (Eds.), *The oxford handbook of metamemory.* Oxford University Press. Retrieved from http://iew3.technion.ac.il/~ackerman/papers/Bernstein%20et%20al.%20in%20press%20-%20Looking%20Backward%20And%20Forward%20on%20Hindsight%20Bias.pdf

Bernstein, D., Atance, C., Loftus, G., & Meltzoff, A. (2004). We saw it all along: Visual hindsight bias in children and adults. *Psychological Science, 15*(4), 264-267.

Bernstein, D., Erdfelder, E., Meltzoff, A., Peria, W., & Loftus, G. (2011). Hindsight bias from 3 to 95 years of age. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*(2), 378-391.

Bernstein, D., & Harley, E. (2007). Fluency misattribution and visual hindsight bias. *Memory, 15*(5), 548-560.

Bernstein, D., Wilson, A., Pernat, N., & Meilleur, L. (2012). Auditory hindsight bias. *Psychonomic Bulletin & Review*, *19*, 588-593.

Birch, S., & Bernstein, D. (2007). What can children tell us about hindsight bias: A fundamental constraint on perspective-taking? *Social Cognition, 25*(1), 98-113.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp.185-205). Cambridge, MA: MIT Press.

Blank, H., Diedenhofen, B., & Musch, J. (2015). Looking back on the London Olympics: Independent outcome and hindsight effects in decision evaluation. *British Journal of Social Psychology, 54*, 798-807.

Blank, H., Fischer, V., & Erdfelder, E. (2003). Hindsight bias in political elections. *Memory, 11*(4/5), 491-504.

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

Blank, H., Nestler, S., von Collani, G., & Fischer, V. (2008). How many hindsight biases are there? *Cognition, 106,* 1408-1440.

Bradfield, A., & Wells, G. (2005). Not the same old hindsight bias: Outcome information distorts a broad range of retrospective judgments. *Memory & Cognition, 33*(1), 120-130.

Brédart, S. (2000). When false memories do not occur: Not thinking of the lure or remembering that it was not heard? *Memory, 8*(2), 123-128.

Bruce, D., & Winograd, E. (1998). Remembering Deese's 1959 articles: The zeitgeist, the sociology of science, and false memories. *Psychonomic Bulletin & Review, 5*(4), 615-624.

Butler, A. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118-1133.

Calvillo, D. (2012). Working memory and the memory distortion component of hindsight bias. *Memory, 20*(8), 891-898.

Calvillo, D. (2013). Rapid recollection of foresight judgments increases hindsight bias in a memory design. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 959-964.

Calvillo, D. P., & Gomes, D. M. (2011). Surprise influences hindsight-foresight differences in temporal judgments of animated automobile accidents. *Psychonomic Bulletin & Review, 18,* 385-391.

Calvillo, D., & Rutchick, A. (2014a). Domain knowledge and hindsight bias among poker players. *Journal of Behavioral Decision Making, 27*, 259-267.

Calvillo, D., & Rutchick, A. (2014b). Political knowledge reduces hindsight memory distortion in election judgments. *Journal of Cognitive Psychology, 26*(2), 213-220.

Carpenter, S. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569.

Carpenter, S., & DeLosh, E. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268-276.

Carpenter, S., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*(3), 474-478.

Carpenter, S., Pashler, H., & Cepeda, N. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23*, 760-771.

Chase, W., & Ericsson, K. (1982). Skill and working memory. *The Psychology of Learning and Motivation, 16*, 1-58.

Choi, D., & Choi, I. (2010). A comparison of hindsight bias in groups and individuals: The moderating role of plausibility. *Journal of Applied Social Psychology, 40*(2), 325-343.

Choi, I., & Nisbett, R. (2000). Cultural psychology of surprise: Holistic theories and recognition of contradiction. *Journal of Personality and Social Psychology, 79*(6), 890-905.

Christensen-Szalanski, J., & Willham, C. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes, 48*(1), 147-168.

Coolin, A., Bernstein, D., Thornton, A., & Loken Thornton, W. (2014). Age differences in hindsight bias: The role of episodic memory and inhibition. *Experimental Aging Research, 40*(3), 357-374.

Coolin, A., Erdfelder, E., Bernstein, D., Thornton, A., & Loken Thornton, W. (2015). Explaining individual differences in cognitive processes underlying hindsight bias. *Psychonomic Bulletin & Review, 22*(2), 328-348.

Cull, W. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*, 215-235.

Cycowicz, Y., Friedman, D., Snodgrass, J., & Duff, M. (2001). Recognition and source memory for pictures in children and adults. *Neuropsychologia, 39*, 255-267.

Davies, M. (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*, 50-88.

Dehn, D., & Erdfelder, E. What kind of bias is hindsight bias? *Psychological Research, 61,* 135-146.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education, 20*(10), 1155-1191.

Erdfelder, E., Brandt, M., & Bröder, A. (2007). Recollection biases in hindsight judgments. *Social Cognition, 25*(1), 114-131.

Erdfelder, E., & Buchner, A. (1998). Decomposing the hindsight bias: A multinomial processing tree model for separating recollection and reconstruction in hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 387-414.

Ericsson, A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*(8), 725-747.

Ericsson, A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211-245.

Ericsson, A., & Simon, H. (1993). *Protocol analysis: Verbal reports as data (revised edition).* Cambridge, MA: Bradford books/MIT Press.

Evelo, A., & Greene, E. (2013). Judgments about felony-murder in hindsight. *Applied Cognitive Psychology, 27*(3), 277-285.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 288–299.

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3*(2), 349-358.

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

Fox, M., Ericsson, A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137*(2), 316-344.

Ghrear, S., Birch, S., & Bernstein, D. (2016). Outcome knowledge and false belief. *Frontiers in Psychology, 7*(118), 1-6.

Gillund, G., & Shiffrin, R. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1-67.

Gray, R., Beilock, S., & Carr, T. (2007). "As soon as the bat met the ball, I knew it was gone": Outcome prediction, hindsight bias, and the representation and control of action in expert and novice baseball players. *Psychonomic Bulletin & Review, 14*(4), 669-675.

Groß, J., & Bayen, U. (2015). Adult age differences in hindsight bias: The role of recall ability. *Psychology and Aging, 30*(2), 253-258.

Goodwin, K. (2007). Dissociative effects of true and false recall as a function of different encoding strategies. *Memory, 15*(1), 93-103.

Goodwin, K. (2013). Reducing false memories via context reinstatement: The roles of encoding and retrieval contexts. *The American Journal of Psychology, 126*(2), 213-225.

Goodwin, K., Meissner, C., & Ericsson, A. (2001). Toward a model of false recall: Experimental manipulation of encoding context and the collection of verbal reports. *Memory & Cognition, 29*(6), 806-819.

Guilbault, R., Bryant, F., Brockway, J., & Posavac, E. (2004). A meta-analysis of research on hindsight bias. *Basic and Applied Social Psychology, 26*(2&3), 103-117.

Harley, E. (2007). Hindsight bias in legal decision making. *Social Cognition, 25*(1), 48-63.

Harley, E., Carlsen, K., & Loftus, G. (2004). The "saw-it-all-along" effect: Demonstrations of visual hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(5), 960-968.

Hasher, L., Attig, M. S., & Alba, J. W. (1981). I knew it all along: or, did I?. *Journal of Verbal Learning and Verbal Behavior, 20*(1), 86-96.

Hawkins, S., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107*(3), 311-327.

Heine, S., & Lehman, D. (1996). Hindsight bias: A cross-cultural analysis. *The Japanese Journal of Experimental Social Psychology, 3*, 317-323.

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.

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.

Hertwig, R., Gigerenzer, G., & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychological Review, 104*(1), *19*4-202.

Hicks, J., & Marsh, R. (1999). Attempts to reduce incidence of false recall with source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(5), 1195-1209.

Hinsz, V. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social Psychology, 59*(4), 705-718.

Hintzman, D. (1986). "Schema abstraction" in a retrieval-based memory model. *Psychological Review, 93*, 411-428.

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 566-581.

Hom, H., & Kaiser, D. (2016). Role of hindsight bias, ethics, and self-other judgments in students' evaluation of an animal experiment. *Ethics & Behavior, 26*(1), 1-13.

Howe, M. (1998). When distinctiveness fails, false memories prevail. *Journal of Experimental Child Psychology, 71*, 170-177.

Hussain, M., Shah, S., Latif, K., Bashir, U., & Yasir, M. (2013). Hindsight bias and investment decisions making empirical evidence from an emerging financial market. *International Journal of Research Studies in Management, 2*(2), 77-88.

Johnson, M. (2006). Memory and reality. *American Psychologist*, 760-771.

Johnson, M., Hashtroudi, S., & Lindsay, S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3-28.

Johnson, M., & Raye, C. (1981). Reality monitoring. *Psychological Review, 88*(1), 67-85.

Jonsson, B., Wiklund-Hörnqvist, C., Nyrro, M., & Börjesson, A. (2014). Self-reported memory strategies and their relationship to immediate and delayed text recall and working memory capacity. *Education Inquiry, 5*(3), 385-404.

Kerr, N. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Larkin, J., McDermott, J., Simon, D., & Simon, H. (1980). Expert and novice performance in solving physics problems. *Science, 208*, 1335-1342.

Larsen, D., Butler, A., & Roediger, H. (2009). Repeated testing improves long-term retention relative to repeated studying: A randomized controlled trial. *Medical Education, 43,* 1174-1181.

Lindsay, S., Johnson, M., & Kwon, P. (1991). Developmental changes in memory source monitoring. *Journal of Experimental Child Psychology, 52*, 297-318.

Littlefair, S., Brennan, P., Mello-Thoms, C., Dung, P., Trieu, Y., Pietryzk, M., … Reed, W. (2016). Outcomes knowledge may bias radiological decision-making. *Academic Radiology.* Advance online publication. doi:10.1016/j.acra.2016.01.006

Louie, T., Chandrasekar, P., & Wu, M. (2014). Time is of the essence? Investigating how culturally-based perceptions of time affect hindsight bias for task completion. *Drake Management Review, 3*(2), 37-52.

Massaro, D., Castelli, I., Sanvito, L., & Marchetti, A. (2014). The 'I knew it all along' phenomenon: Second-order false belief understanding and the curse of knowledge in primary school children. *European Journal of Psychology of Education, 29*(3), 311-326.

Mazursky, D., & Ofir, C. (1990). "I could never have expected it to happen": The reversal of the hindsight bias. *Organizational Behavior and Human Decision Processes, 46*, 20-33.

McCabe, D. P., Geraci, L., Boman, J. K., Sensenig, A. E., & Rhodes, M. G. (2011). On the validity of remember-know judgments: Evidence from think aloud protocols. *Consciousness and Cognition, 20*, 1625-1633.

McDaniel, M., Anderson, J., Derbish, M., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4/5), 494-513.

McKelvie, S. (2003). False recall with the drmrs ("drummers") procedure: A quantitative summary and review. *Perceptual and Motor Skills, 97*, 1011-1030.

Motavalli, A., Nestel, D. (2016). Complexity in simulation-based education: Exploring the role of hindsight bias. *Advances in Simulation, 1*(3). doi:10.1186/s41077-015-0005-7

Nestler, S., Blank, H., & Egloff, B. (2010). Hindsight ≠ hindsight: Experimentally induced dissociations between hindsight components. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1399-1413.

Nestler, S., Blank, H., & von Collani, G. (2008). Hindsight bias doesn't always come easy: Causal models, cognitive effort, and creeping determinism. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1043-1054.

Nestler, S., & Egloff, B. (2009). Increased or reversed? The effect of surprise on hindsight bias depends on the hindsight component. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1539-1544.

Oeberst, A., von der Beck, I., & Nestler, S. (2014). Reading about explanations enhances perceptions of inevitability and foreseeability: A cross-cultural study with Wikipedia articles. *Cognitive Processing, 15,* 343-349.

Pohl, R. (2007). Ways to assess hindsight bias. *Social Cognition, 25*(1), 14-31.

Pohl, R., Bayen, U. J., & Martin, C. (2010). A multiprocess account of hindsight bias in children. *Developmental Psychology, 46*(5), 1268-1282.

Pohl, R., Bender, M., & Lachmann, G. (2002). Hindsight bias around the world. *Experimental Psychology, 49*(4), 270-282.

Pohl, R., Eisenhauer, M., & Hardt, O. (2003). SARA: A cognitive process model to simulate the anchoring effect and hindsight bias. *Memory, 11*(4/5), 337-356.

Pohl, R., & Hell, W. (1996). No reduction in hindsight bias after complete information and repeated testing. *Organizational Behavior and Human Decision Processes, 67*(1), 49-58.

Reece Jones, P. (1995). Hindsight bias in reflective practice: An empirical investigation. *Journal of Advanced Nursing, 21,* 783-788.

Raaijmakers, J., & Shiffrin, R. (1981). Search of associative memory. *Psychological Review, 88*(2), 93-134.

Roediger, H., Agarwal, P., McDaniel, M., & McDermott, K. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382-395.

Roediger, H. & Butler, A. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20–27.

Roediger, H., & Karpicke, J. (2006a). Test-enhanced learning: Taking memory tests

improves long-term retention. *Psychological Science, 17*(3), 249-255.

Roediger, H., & Karpicke, J. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.

Roediger, H., & McDermott, K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803-814.

Roediger, H., & Pyc, M. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248.

Roese, N., & Maniar, S. (1997). Perceptions of purple: Counterfactual and hindsight judgments at Northwestern Wildcats football games. *Personality and Social Psychology Bulletin, 23,* 1245-1253.

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science, 7*(5), 411-426.

Shah, A. K., & Oppenheimer, D. M. (2009). The past of least resistance: Using easy-to-access information. *Current Directions in Psychological Science, 18*(4), 232-236.

Slamecka, N., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592-604.

Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance, 3*(4), 544-551.

Sprondel, V., Kipp, K., & Mecklinger, A. (2011). Developmental changes in item and source memory: Evidence from an ERP recognition memory study with children, adolescents, and adults. *Child Development, 82*(6), 1938-1953.

Stahlberg, D., Eller, F., Mass, A., & Frey, D. (1995). We knew it all along: Hindsight bias in groups. *Organizational Behavior and Human Decision Processes, 63*(1), 46-58.

Swick, D., Senkfor, A., & Van Petten, C. (2006). Source memory retrieval is affected by aging and prefrontal lesions: Behavioral and ERP evidence. *Brain Research, 1107*, 161-176.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*(1), 1-12.

Van Boekel, M., Varma, K., & Varma, S. (2016). A retrieval-based approach to eliminating hindsight bias. Memory. doi:10.1080/09658211.2016.1176202

Varma, K., Ross, P., Lawrenz, F., Roehrig, G., Huffman, D., McGuire, L., … Jang, Y. (2013). *Unpacking the elements of scientific reasoning.* Poster session presented at the National Association for Research in Science Teaching, Rio Grande, Puerto Rico.

Weldon, M., & Bellinger, K. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 1160-1175.

Williams, M., & Hollan, J. (2010). The process of retrieval from very long-term memory. *Cognitive Science, 5*(2), 87-119.

Wood, G. (1978). The "knew-it-all-along" effect. *Journal of Experimental Psychology: Human Perception and Performance, 4*(2), 345-353.

Yama, H., Manktelow, K., Mercier, H., Van der Henst, J., Do, K., Kawasaki, Y., & Adachi, K. (2010). A cross-cultural study of hindsight bias and conditional probabilistic reasoning. *Thinking & Reasoning, 16*(4), 346-371.

Appendix A

Questions (and correct answers). Note that all correct answers are between 1 and 100.

1. How many inches across is the eye of a giant squid? (15)

2. How many neck bones does a giraffe have? (7)

3. How many seats are there on a school bus? (24)

4. How many minutes does it take light from the sun to reach Earth? (8)

5. How many legs does a lobster have? (10)

6. How many provinces does Canada have? (10)

7. How many days can a cockroach live without a head? (9)

8. How many years can a parakeet live? (15)

9. How many teeth does an alligator have? (76)

10. How many miles per hour can a hippo run? (20)

11. How many countries are in South America? (13)

12. How many muscles does it take to frown? (43)

13. How many teeth does a mosquito have? (47)

14. How many pounds is a sperm whale's brain? (20)

15. How many hours does a lion sleep in a day? (20)

16. How many countries are in Europe? (45)

17. How many moons does the planet Saturn have? (46)

18. How many weeks are female dogs pregnant? (9)

19. How many countries are there in Africa? (53)

20. How many feet can a kangaroo jump in one leap? (30)

Appendix B

Think-aloud instructions and practice items.

Experimenter: "In this experiment I am interested in what you say to yourself as you perform some tasks that I will give to you. In order to do this I will ask you to THINK ALOUD as you work on the problems. What I mean by think aloud is that I want you to say out loud EVERYTHING that you are thinking from the time you start each problem until you give an answer. I would like you to talk loud CONSTNATLY from the time I present each problem until you have given your final answer to the question. Tell me everything that goes through your mind, even if you think it seems irrelevant. I don't want you to try to plan out what you say or try to explain your thoughts, but verbalize your thoughts as they occur. Just act as if you are alone in the room speaking to yourself. It is important that you keep talking. If you are silent for any length of time I will remind you to keep talking aloud. Do you understand what I want you to do?"

Wait for and answer questions.

Experimenter: "Good, now we will begin with some practice problems. First, I want you to multiply two numbers in your head and tell me what you are thinking as you are working out an answer."

Turn on recorder

Experimenter: "What is the result of multiplying 24 x 30?"

Experimenter: "Good. Remember not to filter your thoughts, and say everything that you are thinking. Now let's proceed to the next practice task. Just as before, I want you to think aloud while you complete the following problem. Be sure to think aloud as you think about the question, tell me everything that you are thinking. Any questions?"

Wait for and answer questions.

Experimenter: "Name the state capitols that begin with the letter B?"

Upon completion participants are told that they are finished, but before they go, they should read the answers to some of the questions they answered at the beginning of the session (Phase 2). After reading out the 10 questions and answers, they are given the surprise memory test (Phase 3).