Looking Back: An Examination of Hindsight Bias in Change Detection Ability

Ashley Hall

Marietta College

The thesis has been approved for the
Department of Psychology of Marietta College

By

Dr. Jennifer McCabe, Ph.D.
Thesis Committee Advisor

Dr. Ryan May, Ph.D.
Thesis Committee Member
Looking Back: An Examination of Hindsight Bias in Change Detection Ability

Ashley Hall

Marietta College
Abstract

This study examines hindsight bias in relation to performance in the change detection (i.e., ‘flicker’) paradigm. In order to determine confidence ratings and predictions about performance, participants provided ratings at three time points. Following a verbal description of the flicker paradigm, the pre-test was administered. Next, three flicker trials were administered, followed by the mid-session test, three more flicker trials, and finally the post-test. Critically, the post-test asked participants to remember and report their initial confidence ratings and predictions about their performance. The hypotheses were that the participants would have an initial overestimate of ability to detect change that would decrease after participation in the change detection paradigm, and that post-test responses would be significantly different than the initial ratings, showing hindsight bias. Results showed that there was an initial overestimate of confidence in change detection ability, and that hindsight bias was evident in the measure of predicted performance level, but not in confidence ratings. These findings are important in many aspects of life that require us to understand the role of hindsight bias in recalling past events, such as in legal settings involving remembered reports of witnessing a crime.
Looking Back: An Examination of the Hindsight Bias in Change Detection Ability

When someone knows something to be true, it is incredibly difficult to ignore that information and reason about their thoughts or those of others before this information was known (Bernstein, Atance, Loftus, & Meltzoff, 2004). This knowledge leads to a biased representation of events or facts once the outcome is known (Blank, Musch, & Pohl, 2007). This is referred to as the hindsight bias, more commonly known as the “knew it all along” effect (Bernstein et al., 2004). Blank et al. (2007) suggested reasons why hindsight bias is important. These reasons include that hindsight bias is very common among many diverse domains such as history, politics, medicine, economics, legal settings, and everyday decision-making. Hindsight bias is also incredibly hard to avoid and reduce because many people do not realize they display the bias. Our perceptions of many events can also be affected by hindsight bias in that outcomes may seem more predictable than they probably were, inhibiting our ability to learn from experience. By believing that we “knew it all along,” we are unlikely to reflect on our analyses and our decision-making processes (Blank et al., 2007).

Hindsight bias has been extensively researched in multiple domains (e.g., Bernstein et al., 2004; Blank et al., 2007; Hoffrage, Hertwig, & Gigerenzer, 2000; Werth, Strack, & Forster, 2002), but not in relation to change detection. Change detection simply refers to the ability to detect a change in one’s visual world. Counter to most people’s intuition, large changes to the visual world often go unnoticed. This is important to study given all the real-world situations where change detection is important, including coping with traffic and noticing when someone walks into the room (Rensink, 2002). As studied separately, the phenomena of change detection and hindsight bias have both shown to be important aspects of human cognition. This is no less true when considered together, such as in legal situations. People must be able to detect change
and be able to accurately state what was seen and when and, importantly, to be confident in the accuracy of those judgments. Whether or not a change in a crime scene is detected may determine the outcome of a case. Also, when testifying, witnesses must be accurate in their judgments without being clouded by hindsight. At the very least, if hindsight bias is playing a role in the accuracy of judgments, it would be essential that the witness, as well as others in the legal system, be aware of the extent of this type of inaccuracy in human cognition.

It was the goal of the current study to investigate the extent to which people’s judgments and memory for judgments of difficulty in relation to a change detection paradigm are altered by actual experience with the paradigm.

**Hindsight Bias**

As previously discussed, hindsight bias is defined as a person’s tendency to change a previous judgment after learning about the outcome of a situation or the correct answer to a question (Worth et al., 2002). Hindsight bias has been examined using the *hypothetical design* and the *memory design* (Fischoff, 1975; Hoffrage et al., 2000; Worth et al., 2002). The *hypothetical design* compares two groups, one which has knowledge of the outcome but is instructed to ignore it and one without this knowledge. These two groups are compared on how outcome knowledge affects their judgments and whether or not they are aware of the effects. The more common way to examine hindsight bias, however, is the two-phase *memory design* (Worth et al.). The first of the two phases consists of answering relatively difficult questions, with this answer later referred to as the original answer. For example, “On average, how long does a light bulb burn?” (Worth et al., p. 324). After some delay, the second phase takes place. In this phase, the subjects are provided the correct answers to the questions previously asked and are then asked to recall their original answers to the questions, this being termed the *recalled estimate.*
The results of the memory design generally show that the recalled estimates tend to lean more towards the correct answer than the original answers. Otherwise stated, the subjects overestimate what they originally knew (Worth et al.).

Fischoff (1975) conducted a series of experiments on hindsight bias to explore two hypotheses about why hindsight bias occurs, the first of which states that knowing the outcome of a situation makes that outcome see more probable. The second hypothesis states that people who receive the outcome knowledge are unaware that that information may affect their perceptions. Fischoff coined the term *creeping determinism* to explain the first hypothesis, that the outcome is perceived to be inevitable. The results of Fischoff’s experiments support the creeping determinism hypothesis, and also show that participants were unaware that the outcome knowledge affected their perceptions. Rather, they were under the impression that the inevitability of the outcome was foreseeable prior to the actual outcome.

Additional explanations for the hindsight bias have been offered. Agans and Schaffer (1994) conducted a study to explore the influence of outcome knowledge and the availability heuristic on hindsight bias. The *availability heuristic* involves using information you are familiar with (e.g., that is prevalent in the media), instead of factual information, to make an estimate, leading to an overestimate of the frequency of an event. For example, the probability of death by disease is underestimated, whereas probability of death via homicide is overestimated. The media feeds on dramatic events, and thus, are more likely to report homicides than disease, although disease is takes more lives than homicide by more than one hundred times (Agans & Schaffer). Participants were asked to read a scenario that described a misfortune (e.g., lung cancer, an automobile accident, or homicide). Some of the participants were presented with outcome knowledge (*hindsight*) while others were not (*foresight*). For example, those with
outcome knowledge were told that a person was murdered while walking alone, whereas those with no outcome knowledge were simply told that a person walked alone. Participants were asked to predict whether the person in the scenario, other people, or both the person in the scenario and others would suffer from the misfortune described. Those with hindsight knowledge underestimated the likelihood of disease while overestimating the likelihood of homicide. Participants in the foresight group predicted disease to be more likely than both accidents and homicide, and accidents more likely than homicide as well. This suggests that the availability heuristic plays a role in hindsight bias. Those in the foresight condition were more accurate in their predictions than were those in the hindsight condition who had outcome knowledge.

Participants in the hindsight condition may have been more likely to use and rely on current social knowledge; that is, they did not think that they needed more information as they were already given the scenario and the outcome of the situation (Agans & Schaffer).

Explanations for hindsight bias may also vary depending on the design (hypothetical or memory) used to examine it (Blank & Nestler, 2007). Blank and Nestler compared three cognitive process models of hindsight bias: (1) selective activation and reconstructive anchoring (SARA), (2) reconstruction after feedback with take the best (RAFT), and (3) causal model theory (CMT). SARA states that hindsight bias is a result of either selective activation or biased sampling. Selective activation means that certain images are activated as a result of previous knowledge and these images contribute to the judgments being made. Biased sampling refers to feedback or outcome knowledge that acts as a cue or an anchor when trying to remember the original estimate, making the last judgment closer to the feedback information than in the original response. RAFT explains hindsight bias as adaptive learning. Feedback or outcome knowledge updates the existing information, making the retrieval of the original information
difficult in that there is new information to consider. CMT is the authors’ combination of the work of several authors (Hawkins & Hastie, 1990; Schkade & Kilbourne, 1991; Wasserman, Lempert, & Hastie, 1991; Nario & Branscombe, 1995; Roese & Olson, 1996; Pezzo, 2003; Blank & Nestler, 2007). CMT explains hindsight bias as it relates to event outcomes, such that people try to make sense of outcomes by creating relationships between the outcome and what they originally thought, sometimes creating the perception of the inevitability of that outcome. Both SARA and RAFT suggest that hindsight bias is not within conscious control, whereas CMT suggests conscious efforts are made to make sense of the outcome (Blank & Nestler).

Other researchers (Erdfelder & Buchner, 1998; Hawkins & Hastie, 1990; Pohl, 2004; Stahlberg & Maass, 1998; and Bayen, Erdfelder, Bearden, & Lozito, 2006) have examined additional factors that may influence hindsight bias, specifically the recollection and reconstruction biases. When recall of the original response is impaired by the presentation of the correct answer, the recollection bias occurs. There are many mechanisms that instigate the recollection bias, one of which is that the correct answer is then used as an anchor to replicate the original response. In contrast, the reconstruction bias occurs when the original response cannot be recalled. Reconstruction bias may be a byproduct of knowledge updating and/or may include trying to reconstruct the processes used to establish the original response. Bayen et al. also discuss a multinomial modeling approach to hindsight bias, which states that during the second phase of the memory design, in which participants are asked to recall their original response, individuals will attempt to recall their actual response. However, if this response cannot be recalled, then reconstruction of the original response begins (Bayen et al.). Generally, hindsight bias seems to result from a combination of cognitive factors. Both memory and judgment play a role, such as failing to recollect the original response and then using more recent information to
bias the recalled estimate, respectively (Agans & Schaffer, 1994; Blank & Nestler, 2007; Hawkins & Hastie, 1990; Schkade & Kilbourne, 1991; Wasserman et al., 1991; Nario & Branscombe, 1995; Roese & Olson, 1996; Pezzo, 2003; Erdfelder & Buchner, 1998; Pohl, 2004; Stahlberg & Maaas, 1998; and Bayen et al., 2006).

Surprisingly, similar levels of hindsight bias have been demonstrated across multiple age ranges. To determine the age at which hindsight bias develops, Bernstein et al. (2004) utilized a visual hindsight bias paradigm. The participants, including 3-, 4-, and 5-year-olds, and undergraduate students, were asked to identify degraded images that would gradually clarify. They were also asked to estimate when someone else of their age might identify the object as well. This was to demonstrate that hindsight bias, as the individuals already knew what the image was. The results of their study showed that all participants were vulnerable to hindsight bias. Also, when the identity of the object was known, all participants overestimated the ability of a peer to identify the object. These results suggest that hindsight bias is present in children as young as three years old, and continues to present itself into adulthood (Bernstein et al.).

Bayen et al. (2006) conducted research on hindsight bias in regards to aging. They used the memory design to assess hindsight bias in younger and older adults. The results show that there is no consistent difference between younger and older adults in hindsight bias, that the bias can be larger, equal to, or smaller when comparing younger and older adults, depending on the experimental set-up (Bayen et al.).

In sum, factors that may contribute to the hindsight bias include knowledge updating, the availability heuristic, recollection bias, and reconstruction bias. Hindsight bias, thus, seems to be a result of both memory and judgment factors. Also, individuals tend to be unaware that outcome knowledge may influence their recalled estimates. Hindsight bias seems to be present in
individuals at all ages, with no improvement or decline in performance with age. Hindsight bias is also incredibly difficult to avoid and reduce, making it an important phenomenon for continued research.

**Change Detection**

Hindsight bias has not formerly been examined in a change detection paradigm, which is the focus of this study. As previously discussed, change detection refers to the ability to detect a change in one’s visual world. Surprisingly, rather large changes in the visual world often go unnoticed.

One aspect of cognition that has been linked to change detection is visual attention. Many research studies suggest that attention to the changing region is necessary in order to detect a change (Beck, Levin, & Angelone, 2007; Hollingworth, Schrock, & Henderson, 2001; Levin & Simons, 1997; Rensink, O'Regan, & Clark, 1997; Simons, 2000; Simons & Rensink, 2005). Rensink et al. (1997) conducted the classic study to examine the cognitive components involved in change detection. They developed the *flicker paradigm* in order to examine whether change blindness, the name given to the phenomenon of high difficulty associated with detecting changes in scenes, may be attributable to an attentional mechanism. The flicker paradigm presents two real-world images: one original image (A) and one modified version of the original image (A₁). These real-world images typically contain many details, only one of which changes between A and A₁. The original image alternates with the modified image repeatedly, with a brief blank screen inserted between each image. The change in the modified version of the image is generally one that is meant to be highly visible with the blank screen removed. That is, when the images are alternated back and forth with no blank screen in between, local motion cues
allow for the immediate detection of the change. However, with the blanks inserted between the images change detection is remarkably difficult (Rensink et al.).

Rensink et al. (1997) were also interested if the degree of interest of the changing object had an influence on change detection. Degree of interest was classified as central or marginal, determined by five naïve observers asked to describe the scenes. Objects of central interest were those that were mentioned by at least three of the observers while those of marginal interest were not mentioned at all. As previously mentioned, detecting changes in the flicker paradigm is incredibly difficult, taking an average of 7.3 alternations to detect changes of central interest and more than 17 alternations to detect changes of marginal interest (Rensink et al., 1997).

The flicker paradigm eliminates two factors discussed by Rensink et al. (1997) that may increase the difficulty of detecting changes, namely insufficient time to build a representation of the scene and disruptions to eye movements. If those were actually the cause of difficulties in detecting a change, then detection would be simple with the flicker paradigm because subjects have an unlimited amount of viewing time for the alternating images, and because they are free to move their eyes around the visual image as they deem appropriate. However, if attention is the key factor, as Rensink et al. (1997) predicted it would be, the difficulty would remain. The results of this study suggest that visual perception of change in the absence of motion cues (such as is the case with the flicker paradigm) only transpires when attention is focused on the object changing. When focused attention is not present, visual memory is overwritten by subsequent images; thus, comparisons between images cannot be made (Rensink et al., 1997).

In addition to Rensink et al.’s (1997) interpretation of impaired change detection ability, it is also important to consider alternative explanations for change blindness. Simons (2000) presented possible explanations for the inability to detect changes, including overwriting, first
impressions, nothing being stored, nothing being compared, and feature combination, all of which are likely to occur as explanations for change blindness. *Overwriting*, as suggested by Rensink et al. (1997), is that the original image is replaced in visual memory by the modified image. Because the first image is replaced there cannot be any comparison, and thus, no change detection. Similarly, *first impressions* states that the original image is retained and the modified image is not encoded, also hindering change detection ability. *Nothing being stored* simply states that neither the original or modified image is stored in the visual memory. If nothing is stored change detection is not possible. *Nothing being compared* differs in that both images are stored but are not compared with each other. This is more often the case in incidental change detection tasks, in which observers initially do not know that a change will occur, rather than an intentional change detection task, such as the flicker paradigm, in which observers expect a change to occur and actively search for it. Finally, *feature combination* is when the original and modified images are combined to form one image. Some features are retained from the original image and also from the modified image and are combined to form a “coherent” representation of the scenes.

Simons and Levin (1998) extended Rensink et al.’s (1997) research by hypothesizing that attention may be necessary, but not sufficient, for change detection. Simons and Levin (1998) suggested that a limitation of the Rensink et al. (1997) study, as well as other previous studies, could be that the flicker paradigm could not fully assess the ability to detect change in real-world situations. Taking this into account, Simons and Levin (1998) conducted a study using unsuspecting pedestrians as participants in a real-world change detection paradigm. The participants were stopped by what appeared to be a lost pedestrian asking for directions but was actually an experimenter. During the interaction, two people carrying a door rudely passed through the conversation, one of which switched with the seemingly lost pedestrian. Even with
what should be such an obvious change, only seven of fifteen pedestrians reported noticing the change. Interestingly, all of these seven were roughly the same age as the experimenter. Those who did not notice the change were older than the experimenter. Simons and Levin suggest that a possible explanation for this difference is that the younger participants were more likely to notice changes because they viewed the experimenter as similar to themselves. In order to test this hypothesis, the look of the experimenters was changed to make them appear different from the participants. Results showed that of the twelve participants who were all of roughly the same age as the experimenters, only four noticed the change. Based on this, one factor involved in change detection in the real world may be the importance of the change to the observer; in this specific case, the similarity to the observer.

Another aspect of change detection investigated by researchers recently is the development of change detection ability over the life span. Shore, Burack, Miller, Joseph, and Enns (2006) utilized the flicker paradigm to explore change detection ability across four age groups: 5-7 year olds, 7-9 year olds, 9-12 year olds, and 18-30 years old. Their results suggest that the ability to detect change improves with age; this is in contrast to developmental research on the hindsight bias, which shows consistent rates of hindsight bias in multiple age groups (Bayen et al., 2006).

A newer focus in the research literature on change detection has been participants’ predictions and beliefs about their change detection abilities. Beck, Levin, and Angelone (2007) conducted a series of experiments to explore the relationship between participants’ beliefs about roles of intention and scene complexity in change detection ability. Before taking part in the task, participants predicted relatively equal performance for incidental tasks, in which a change is not expected, and intentional tasks, in which observers expect and actively search for a change.
However, the performance in intentional tasks was significantly higher, by more than forty percent, than that of incidental tasks. Additionally, as the complexity of the scene increased, performance decreased. The difference between change detection performance for incidental and intentional tasks also increased as scene complexity increased. This suggests that intention to find a change plays a much greater role in more complex scenes. Beck et al.’s (2007) results suggest that there is only a partial understanding of the relationship between change detection, intention, and scene complexity. In relation to the current study, the flicker paradigm that was used is an intentional task, as the participants expect and search for the change. Performance for intentional tasks is significantly higher than incidental tasks; therefore, performance should be better when the observer expects a change rather than if they were not told to detect a change.

Directly relevant to the current research, another study of participants’ predictions and beliefs about change detection abilities found that participants who were presented only with verbal descriptions of change detection scenarios grossly overestimated their ability to detect the changes (Levins & Simons, 2000). In fact, when these confidence ratings were compared with the authors’ previous work (Levin & Simons, 1997), they found that whereas only 11% of participants actually noticed the changes in previous research, an average of 83% stated they would detect the same changes. These results provide a basis for one hypothesis of the current study, that initial confidence ratings will be relatively high in comparison to actual performance.

As established earlier, change detection, particularly in the flicker paradigm, can be very difficult. In addition, people often overestimate their ability to detect changes. Change detection has been studied extensively but not in relation to hindsight bias, which is the focus of the current study.

*Overview of the Present Study*
The purpose of this study was to examine initial (i.e., pre-test) and mid-session confidence ratings and predicted performance levels in regard to the ability to detect changes in the flicker paradigm, and also to measure the extent of hindsight bias in relation to initial confidence ratings, using a post-test. Participants predicted their ability to detect changes, participated in three flicker trials, rated their ability to detect changes in subsequent trials, participated in three more flicker trials, and finally were asked to recall and restate their initial predictions.

Based on past research, two main hypotheses were proposed. First, because prior research has shown an overestimation of the ability to detect change prior to experiencing change detection (i.e., ‘flicker’) trials (Levins & Simons, 2000), it was predicted that participants will initially show this overestimation. When asked to rate their confidence and predicted performance levels again after experiencing the flicker paradigm, ratings should decrease in relation to the initial estimates because participants have experienced the difficulty of the task first-hand. Finally, at the end of the session when participants are asked to remember and restate their initial confidence ratings, it was hypothesized that there would be a significant decrease in confidence compared to the initial ratings, which would indicate the presence of hindsight bias. Similarly, a significant increase in the predicted number of alternations variable from pre- to post-test would demonstrate hindsight bias.

Method

Participants

Participants consisted of 47 undergraduate students at Marietta College, composed of students enrolled in psychology courses. Students in these courses received course credit in exchange for their participation.
Materials

To assess the initial confidence ratings in the ability to detect a change, the flicker paradigm was explained and demonstrated to each participant. The general design of the flicker paradigm (see Appendix A) was shown to participants to simplify the explanation. The script of the explanation of the design of the flicker paradigm is found in Appendix B.

Each participant participated in a total of 6 trials of the flicker paradigm task, which have been retrieved from the University of British Columbia website (2003). The images were viewed on a television screen. For each flicker trial, the original image was presented for 240 ms, the blank screen was presented for 80 ms, and the modified image was presented for 240 ms (Rensink et al., 1997). All 6 of the scenes presented the same type of change, such that an object disappeared in the modified image and reappeared in the original image. The scenes are labeled Airplane, Chopper & Truck, Farm, Harborside, Street Corner, and Tourists (University of British Columbia). Airplane was a photo of military service men loading an airplane on a sunny day with a clear blue sky. The object that disappeared was the side engine of the plane. Chopper & Truck was a photo of a military helicopter above a military truck on a sunny day, with the shadow of the helicopter as the object that disappeared. Farm was a photo of a few red barns with a tree with red blossoms hanging in the foreground, with a branch from the tree as the object that disappeared. Harborside was a photo of multiple fishing boats docked in the harbor. A child was in the foreground reaching towards a boat. The object that disappeared was a tackle box on a boat in the front of the photo. Street Corner was a photo of a busy street corner with people walking in the background and cars in the foreground. The reflection on one of the car’s windshield was the disappearing object. Finally, Tourists was a photo taken, from behind, of two people examining a sculpture outdoors on a sunny day with trees in the background. One of the
trees near the sculpture was the object that disappears. The type of change was kept consistent across the trials in order to simplify the study and not allow for differences in change detection success due to type of change.

A pre-test, a mid-session test, and a post-test were utilized. The first two tests asked the participants to predict how many image alternations would be necessary to detect the change in upcoming trials, whereas the final test asked participants to remember and report their original ratings. All of the tests also included a rating scale that measured confidence in detecting the change in the flicker paradigm, which was measured on a scale of 0 to 100 (100 being that they would detect a change 100 percent of the time) (see Appendices C and D).

Procedure

Participants were tested individually, in sessions lasting approximately 30 minutes. Upon collection of the signed consent forms, the experimenter explained the flicker paradigm to the participants and showed the accompanying static images on the television screen (see Appendix A). Then, the participants were asked to complete the pre-test (Appendix C). Once the pre-test was completed, they participated in three flicker trials. The number of alternations needed to detect the change, which is the typical dependent variable in this paradigm (Rensink et al., 1997), was recorded. Each trial continued for about 60 seconds, or 92 alternations, as this is the typical time allotted for each flicker trial (Rensink et al., 1997). If the changes were not detected within 60 seconds, the participant moved on to the next trial. If the change was detected sooner than the duration of 92 alternations, the trial continued for the complete duration in order to keep the time between tests constant. Upon completion of these three trials, the second rating scale was administered (see mid-session test in Appendix D). Both the pre-test and the mid-session test were completed through use of a computer program. Next, the last three flicker trials were
completed, using the same procedure as stated above. The six flicker trials were counterbalanced among participants, such that participants did not view the scenes in the same order.

After all the flicker trials were completed, participants were asked to complete a filler task, a word search, for five minutes, in order to allow for more time between the pre- and post-test. Upon completion of the filler task, participants were told that the computer malfunctioned and that their pre-test ratings had been lost. Thus, participants were asked to remember and report their initial confidence in their change-detection ability and the predicted number of alternations to detect a change, exactly as they had done prior to participating in the flicker trials, this being the recalled estimate (see Appendix C). At the end of the session, participants were debriefed and dismissed.

Results

Repeated-measures ANOVAs were used to compare both the predicted number of alternations needed to detect a change and the confidence ratings across three time points. These measured the initial pre-task estimates (i.e., pre-test), whether or not they changed after experience with three trials of the flicker paradigm (i.e., mid-session test), and finally if the participants showed hindsight bias when reporting the recalled estimate (i.e., post-test).

Prior to conducting the inferential tests, an outlier analysis was conducted for both dependent variables using boxplots to determine if any data points might be considered extreme outliers. None were found; thus, the following analyses included all of the original data points.

The alpha level was set at .05 for all analyses. Partial $\eta^2$ is reported as an effect size measure, where $\eta^2 < .09$ is considered small, $.09 < \eta^2 < .25$ is considered medium, and $\eta^2 > .25$ is considered large (Cohen, 1988).
The first ANOVA compared the predicted number of alternations needed to detect a change in the pre-test ($M = 14.74$, $SD = 17.32$), the mid-session test ($M = 45.96$, $SD = 26.23$), and the post-test ($M = 22.74$, $SD = 20.95$) (see Table 1). This resulted in a significant main effect of time of test, $F(2, 47) = 40.52$, $p < .001$, $\eta^2 = .47$. Post hoc tests revealed a significant difference between the pre-test and both the mid-session ($p < .001$) and post-test ($p = .003$) ratings. A significant difference was also found between the mid-session and the post-test ($p < .001$).

The second ANOVA compared the confidence rating means, which could range from 0 to 100, in the pre-test ($M = 63.53$, $SD = 18.92$), the mid-session test ($M = 55.72$, $SD = 22.86$), and the post-test ($M = 63.21$, $SD = 18.17$) (see Table 1). This analysis also showed effects of time of test, $F(2, 47) = 6.50$, $p = .002$, $\eta^2 = .13$. Post hoc tests revealed a significant difference between the mid-session test and both the pre-test ($p = .032$) and the post-test ($p = .029$). However, a significant difference was not found in the critical comparison between the pre- and post-test.

Exploratory correlational analyses were also conducted to determine whether there was a relationship between confidence ratings and predicted number of alternations at the three time points and actual performance on the six flicker trials. Because every participant did not detect a change in every trial, three variables were used to measure actual flicker performance: (1) the mean number of alternations needed to detect the change, omitting data from the trials in which a change was not detected; (2) the mean number of alternations needed to detect the change, with the use of the maximum number of alternations (i.e., 92) as the score for trials in which no change was detected; and (3) the proportion of trials in which a change was detected (out of 6).

There was a significant positive correlation between flicker task performance and mid-session ratings for predicted number of alternations needed to detect a change, using both
Method 1, $r(45) = .590, p < .001$, and Method 2, $r(45) = .383, p = .008$, to quantify flicker task performance, but there was no significant correlation using Method 3, $r(45) = -.150, p = .313$.

Similarly, the relationship between the flicker task performance and mid-session confidence ratings was significant, using both Method 2, $r(45) = -.418, p = .003$, and Method 3, $r(45) = .501, p < .001$, but not Method 1, $r(45) = -.231, p = .119$. All other relationships between predicted (i.e., at the pre-test and post-test time points) and actual performance measures were non-significant ($p s > .05$). See Table 2 for correlations and Table 3 for descriptive statistics.

**Discussion**

The purpose of this study was to examine confidence ratings and predicted performance levels in the ability to detect changes in the flicker paradigm stimuli, and also to measure the extent of hindsight bias in relation to initial ratings. The first of the two hypotheses predicted that participants would initially show an overestimation of change detection ability (e.g., Levin & Simons, 2000). When asked to rate their confidence and predicted performance levels again after experiencing the flicker paradigm, ratings were expected to decrease in relation to the initial estimates because participants had experienced the difficulty of the task first-hand. This hypothesis was supported in that participants’ ratings were significantly different from the pre-test to mid-session test time points, showing an increase for predicted number of alternations and a decrease for confidence ratings.

At the end of the session, when participants are asked to remember and restate their initial confidence ratings, it was hypothesized that there would be a significant decrease in confidence compared to the initial ratings, which would indicate the presence of hindsight bias. Contrary to this hypothesis, confidence ratings for the pre- and post-test were almost identical. Also, the predicted number of alternations needed to detect a change was expected to be higher at the post-
test compared to the pre-test ratings, if hindsight bias was present. This hypothesis was supported in that the post-test scores were significantly higher than those in the pre-test.

The fact that hindsight bias was shown in the predicted number of alternations variable, but not in confidence ratings, was unexpected. In general, differences between the two scales of measurement may help explain why hindsight bias was only found in one of the measures. For one, confidence was rated on a more arbitrary scale than that of the number of alternations. For example, one participant’s rating of 52 could be similar to another participant’s rating of 37 on the confidence scale, whereas points on the number of alternations scale should have similar meaning for all participants. The fact that there was no hindsight bias found in the confidence ratings suggests that these scores were overall more memorable. This may be due to the fact that the 0 to 100 confidence scale has more easily imaginable points (e.g., a rating of 50 could be imagined as a hash mark in the center of the 0 to 100 number line), and may be more familiar to participants, compared to the number of alternations scale. Thus, there may exist a difference between confidence and number of alternations that explain hindsight in the latter, but the not former. Further research is needed to explore this potential difference.

To help address this inconsistency between scales, future research studies could ask participants to report predicted confidence for multiple time points during the flicker task (i.e., how confident they are that they would detect a change within the first 20, 40, and 60 seconds of an upcoming flicker trial). Collecting multiple data points in this way may allow for a detection of hindsight bias in the change detection paradigm, if indeed it exists. Another way to modify the design to make it more parallel to the classic two-phase memory design (e.g., Fischhoff, 1975) would be to provide information about participants’ actual performance levels after completing the flicker trials. It would be interesting to see if providing accurate numbers regarding
performance might cause even greater hindsight bias in the number of alternations variable, and might allow for the detection of hindsight bias in the confidence ratings.

Of secondary interest were the intercorrelations between actual performance levels and predicted ratings of performance. Results showed that there were no relationships between confidence ratings and predicted number of alternations, as measured in the pre- and post-test, and actual performance on the flicker trials. The only significant relationships were between actual performance and predicted ratings at the mid-session test. For these mid-session ratings, there was a significant relationship between predicted number of alternations and actual performance, using two out of three methods for quantifying flicker performance, and a significant relationship between confidence ratings and actual performance, again using two out of three operationalization methods. This suggests that once participants had experience with the flicker task, their predictions about confidence and future performance were consistent with their levels of actual performance on future flicker trials. However, as expected, their predictions prior to having experience with the task were not consistent with future performance (Levin & Simons, 2000).

In sum, consistent with past research showing hindsight bias in a variety of paradigms (e.g., Bernstein et al., 2004; Blank et al., 2007; Hoffrage et al., 2000; Worth et al., 2002), hindsight bias was evident in this study. This is the first demonstration of hindsight bias using the change detection paradigm, and results indicated that the bias is shown only in the predicted number of alternations variable. Hindsight bias is common among many situations and is incredibly difficult to avoid or reduce, likely because it is due to a combination of cognitive mechanisms, such as memory and judgment (e.g., Blank & Nestler, 2007; Erdfelder & Buchner, 1998; Hawkins & Hastie, 1990; Pohl, 2004; Stahlberg & Maaas, 1998; Bayen et al., 2006). Most
Looking Back

often, this bias is not even recognized, also contributing to the difficulty of avoiding and reduce hindsight bias. It is incredibly important to understand the phenomenon of hindsight bias as it may affect our perceptions. With this bias, outcomes seem more probable than they most likely were. This inhibits our ability to learn from experience by making us unlikely to reflect on our analyses and decision making processes (Blank et al., 2007). As stated earlier, it is imperative to understand hindsight bias in relation to change detection, for example in legal settings. This understanding may lend itself to taking a closer look at cases which involve witness testimony. Recognizing this limitation in human memory could make a huge difference in the future of those on trial. The results of this study demonstrate that hindsight bias is present in change detection tasks.

Both hindsight bias and change detection have been studied extensively independently, it would be beneficial to continue examining them together. Future research using similar methodology, perhaps a replication of the present study, might want to give participants immediate feedback on the flicker trials. For example, upon the end of the trial, tell them how many alternations were needed for them to detect the change. This could potentially result in a change in the confidence ratings. From past research, we have learned that change detection, contrary to popular belief, is actually very difficult (Levin & Simons, 2000; Rensink et al., 1997). Researchers are also becoming more aware of the limitations in human memory in regards to hindsight bias. In order for these two phenomena to be understood as they relate to each other, further research is important. Examining real-world situations, as opposed to laboratory settings, obviously increases external validity. For example, hindsight bias could be examined in the context of a crime scene in which change detection is needed, thus increasing generalizability to legal settings. In sum, although hindsight bias was exhibited in one of the
predicted change detection ability measures, a final judgment about the consistency of this relationship awaits future research.
References


Appendix A

General design of the flicker paradigm (University of British Columbia, 2003).
Appendix B

Script of explanation of the flicker paradigm (Rensink et al., 1997).

The flicker paradigm presents two images; one original image (A) and one modified version of the original image (A'). The original image alternates with the modified image repeatedly, with a blank screen inserted between each image. Each flicker trial you observe will contain a change. Your task is to detect the change as quickly as you can. I will then ask you to describe the change to be sure that you have correctly detected it.
Appendix C

Pre-test and Post-test

Participant Number (for experimenter)

1. Age
2. Sex
3. Have you ever participated in a change detection task?
4. Were you familiar with the flicker paradigm prior to this study?
5. How many alternations of images do you think it would take for you to detect a change in the flicker paradigm? (1 alternation = one presentation of the original image and one presentation of the modified image; choose a number between 1 and 92)
6. How confident are you that you would detect a change? (Choose a number between 0 and 100; 0 being that you will not detect a change in any case and 100 being that you will detect a change in 100% of cases)
Appendix D

Mid-session test

Participant Number (for experimenter)

1. On average, how many alternations of images do you think it will take for you to detect a change in subsequent trials? (1 alternation = one presentation of the original image and one presentation of the modified image; choose a number between 1 and 92)

2. How confident are you that you will detect a change in subsequent trials? (Choose a number between 0 and 100; 0 being that you will not detect a change in any case and 100 being that you will detect a change in 100% of cases)
Table 1

*Means and Standard Deviations for the Predicted Number of Alternations and Confidence Ratings in the Pre-Test, the Mid-Session Test, and the Post-Test*

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Mid-session test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternations</td>
<td>14.74 (17.32)</td>
<td>45.96 (26.23)</td>
<td>22.74 (20.95)</td>
</tr>
<tr>
<td>Confidence</td>
<td>63.53 (18.92)</td>
<td>55.72 (22.86)</td>
<td>63.21 (18.17)</td>
</tr>
</tbody>
</table>

*Note.* Confidence ratings were measured on a scale from 0 (indicated no confidence in detecting a change) to 100 (indicating full confidence in detecting a change).
Table 2

*Correlations for Predicted Number of Alternations and Confidence Ratings in Pre-Test, Mid-Session test, and Post-Test, in Relation to Three Actual Performance Measurements: Number of Alternations Needed to Detect a Change When Omitting Undetected Trials, Number of Alternations Needed to Detect a Change When Using Maximum Number of Alternations for Undetected Trials, and Overall Proportion of Trials in Which the Change was Detected*

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Mid-session test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alternations</td>
<td>Confidence</td>
<td>Alternations</td>
</tr>
<tr>
<td>Omitted</td>
<td>.009</td>
<td>-.019</td>
<td>.590*</td>
</tr>
<tr>
<td>Proportion</td>
<td>.255</td>
<td>-.042</td>
<td>-.150</td>
</tr>
<tr>
<td>Maximum</td>
<td>-.197</td>
<td>.047</td>
<td>.383*</td>
</tr>
</tbody>
</table>

*Note.* *p* < .05. **p** < .01.
Table 3

Means and Standard Deviations for the Flicker Paradigm Performance Measures: Number of Alternations Needed to Detect a Change When Omitting Undetected Trials, Number of Alternations Needed to Detect a Change When Using Maximum Number of Alternations for Undetected Trials, and the Number of Trials Detected Out of 6 in Which the Change was Detected

<table>
<thead>
<tr>
<th>Flicker Paradigm Performance Measure</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Alternations (Omitted Trials)</td>
<td>36.29 (15.37)</td>
</tr>
<tr>
<td>Number of Alternations (Max Number 92)</td>
<td>55.41 (16.29)</td>
</tr>
<tr>
<td>Number of Trials Detected Out of 6</td>
<td>3.85 (1.32)</td>
</tr>
</tbody>
</table>

Note. There were 6 total flicker trials administered.