UNIVERSITY OF CALIFORNIA,

IRVINE

A New Method of Studying Confidence Malleability:
Self-Sourced Misinformation as Post-Identification Feedback

DISSERTATION

submitted in partial satisfaction of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in Psychology & Social Behavior

by

Rachel Leigh Greenspan

Dissertation Committee:

Professor Elizabeth Loftus, Chair

Professor Linda Levine

Professor Nicholas Scurich

TABLE OF CONTENTS

	Page
List of Tables	iii
List of Figures	iv
Acknowledgements	v
Curriculum Vitae	vi
Abstract of the Dissertation	x
Introduction	1
Dissertation Outline	4
Literature Review	6
The Relationship Between Eyewitness Confidence and Accuracy	6
Verbal and Numeric Confidence	10
Post-Identification Feedback	14
The Misinformation Effect	26
Choice Blindness and Self-Sourced Misinformation	32
Summary	36
Pilot Study A	39
Method	39
Results	43
Discussion	45
Pilot Study B	46
Method	46
Results	47
Discussion	49
Study 1	51
Method	51
Results	62
Discussion	70
Study 2	75
Method	75
Results	79
Discussion	88
General Discussion	93
Limitations and Future Directions	94
Conclusion	96
References	97
Appendices	110

List of Tables

		Page
Table A.1	Paired Confidence Judgments- Pilot Study A	117
Table A.2	Numeric Translation of Verbal Confidence Scale- Pilot Study A	118
Table B.1	Paired Confidence Judgments- Pilot Study B	119
Table B.2	Numeric Translation of Verbal Confidence Scale- Pilot Study B	120
Table 1.1	Example Confidence Elaborations- Study 1	121
Table 1.2	Concurrent Detector Example Responses- Study 1	122
Table 1.3	Posttest Questions- Study 1 Restricted Sample	123
Table 1.4	Posttest Questions- Study 1 Full Sample	124
Table 2.1	Concurrent Detector Example Responses- Study 2	125
Table 2.2	Posttest Questions- Study 2 Restricted Sample	126

List of Figures

		Page
Figure 1.1	Initial Identification Confidence- Study 1	127
Figure 1.2	Confidence Change Over Time- Study 1	128
Figure 1.3	Identification Outcome and Confidence Change- Study 1	129
Figure 1.4	Detection Status and Confidence Change- Study 1	130
Figure 1.5	Posttest Questionnaire Subscales- Study 1	131
Figure 1.6	Posttest Composite and Detection Status- Study 1	132
Figure 2.1	Initial Identification Confidence- Study 2	133
Figure 2.2	Confidence Change Over Time- Study 2	134
Figure 2.3	Detection Status and Confidence Change- Study 2	135
Figure 2.4	Posttest Questionnaire- Study 2	136
Figure 2.5	Posttest Questionnaire Subscales- Study 2	137
Figure 2.6	Posttest Composite by Initial Confidence- Study 2	138

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE- 1321846. This work was also supported by funding from the University of California, Irvine School of Social Ecology Dissertation Data Collection Fellowship, the University of California, Irvine Department of Psychology and Social Behavior Alison-Clarke Stewart Award, the University of California, Irvine Department of Psychology and Social Behavior, and the University of California, Irvine Undergraduate Research Opportunities Program.

Curriculum Vitae

Rachel Leigh Greenspan

Education

Ph.D. Psychology and Social Behavior *2018*
University of California, Irvine
Major: Social Psychology
Minors: Quantitative Methods and Psychology & Law
Advanced to Candidacy: May 2017
Dissertation: *A new method of studying confidence malleability: Self-sourced misinformation as post-identification feedback*

M.A. Social Ecology *2015*
University of California, Irvine
Thesis: *Choice blindness as misinformation: Memory distortion in an eyewitness identification task*

B.S. Psychology, *summa cum laude* *2013*
University of Florida

B.A. Criminology, *cum laude* *2013*
University of Florida

Grants and Awards

National Science Foundation Graduate Research Fellowship ($96,000) *2014-2018*

Alison Clarke-Stewart Dissertation Award ($1,000) *2018*

UC Irvine School of Social Ecology Dissertation Data Collection Stipend ($1,000) *2017*

Pacific Chapter of American Association for Public Opinion Research Student Paper Award ($250) *2017*

UC Irvine Newkirk Center for Science and Society Grant ($11,707) *2017*

UC Consortium on Social Science and the Law Summer Fellowship ($2,333) *2017*

UC Irvine Center for Psychology and Law Distinguished Fellows Program *2015 & 2016*
Mentor: Dr. Park Dietz

University of Florida Four-Year Scholar Award *2013*

Publications

Grady, G. H., **Greenspan, R.L.**, Liu, M. (2018). What's the best size for matrix-style questions in online surveys? *Social Science Computer Review.*

Cochran, K. J., **Greenspan, R. L.,** Bogart, D. F., & Loftus, E. F. (2018). (Choice) blind justice: Legal implications of the choice blindness phenomenon. *University of California, Irvine Law Review.* 8, 85.

Loftus, E. F., & **Greenspan, R. L.** (2017). If I'm certain, is it true? Accuracy and confidence in eyewitness memory. *Psychological Science in the Public Interest, 18*(1), 1–2.

Cochran, K. J., **Greenspan, R. L.,** Bogart, D. F., & Loftus, E. F. (2016). Memory blindness: Altered memory reports lead to distortion in eyewitness memory. *Memory & Cognition.* 44(5), 717-726.

Greenspan, R. L., Scurich, N. (2016). The interdependence of perceived confession volunatriness and case evidence. *Law and Human Behavior*, *40*(6), 650-659.

Conference Presentations and Posters

*indicates undergraduate student mentee

Grady, G. H., **Greenspan, R. L.**, Liu, M. (2017, December). *What's the best size for matrix-style questions in online surveys?* Paper at the Pacific Chapter of American Association for Public Opinion Research Annual Conference, San Fransisco, CA.

*Xiao, J., *Dufall, C., **Greenspan, R. L.,** Loftus, E. F. (2017, May) *Does debriefing work? False memories persist five days post-debriefing.* Poster at the University of California, Irvine Undergraduate Research Symposium, Irvine, CA.

Greenspan, R. L. (2017, April). *Eyewitness identification: Best practices and future directions for law enforcement.* Presentation at the UCI Associated Graduate Student Symposium, Irvine, CA.

Greenspan, R. L., Loftus, E. F. (2017, March). *Assessing the use of eyewitness identification reforms in law enforcement agencies.* Paper at the American Psychology-Law Society Conference, Seattle, WA.

Greenspan, R. L., Cochran, K. J., Loftus, E. F. (2016, January). *Choice blindness as misinformation: Memory distortion in an eyewitness identification task.* Poster at the Annual Society for Personality and Social Psychology Conference, San Diego, CA.

Greenspan, R. L., Cochran, K. J., Loftus, E. F. (2015, October). *Choice blindness as misinformation: Memory distortion in an eyewitness identification task.* Poster at the University of California Conference on Social Science and Law, Irvine, CA.

Greenspan, R. L. (2015, June). *Choice blindness as misinformation: Memory distortion in an eyewitness identification task*. Presentation at the Psychology and Social Behavior Colloquium Series, Irvine, CA.

Greenspan, R. L. (2015, May). Panel member at the Distinguished Fellows Student Experience Luncheon, Irvine, CA.

*Bhakta, K., **Greenspan, R. L.**, Loftus, E. F. (2015, May). *False memories in the context of body image.* Presentation at the University of California, Irvine Undergraduate Research Symposium, Irvine, CA.

Greenspan, R. L., Cochran, K. J., Loftus, E. F. (2015, March). *Now you see him, now you don't: Choice blindness and eyewitness identification.* Presentation at the University of California, Irvine Associated Graduate Student Symposium, Irvine, CA.

Greenspan, R. L., Scurich, N. (2015, March). *The non-independence of perceived voluntariness of confessions*. Poster session presented at the American Psychology-Law Society Annual Meeting 2015, San Diego, CA.

Cochran, K. J., **Greenspan R. L.**, Bogart, D., Loftus, E. F. (2015, March). *Choice blindness in an eyewitness misinformation paradigm*. Presentation at the International Convention of Psychological Science, Amsterdam, The Netherlands.

Teaching Experience

University of California, Irvine *2013-2015 (7 quarters)*
Teaching Assistant

The Social Animal, Eyewitness Testimony, Research Design, Lifespan Development, Health Psychology, Psychology Fundamentals

Guest Lectures: The Social Animal, Psychology and the Law

University of Florida *2011-2012 (4 semesters)*
Teaching Assistant

Psychology of Law, Psychology and the Law

Guest Lecture: Psychology of Law

Mentorship

UC Irvine
Undergraduate Research Opportunity Program Advisor *2014-Present*
Advised Student Projects:
Bhakta, K. (2015): False memories in the context of body image ($500)
Duffall, C., Xiao, J. (2017): Does debriefing work? False memories persist five days post debriefing ($950)
Rodriguez, K., Arun, V. (2018): The role of implicit and explicit biases in the misinformation effect (*under review*)
Duffall, C., Xiao, J. (2018): A new method of studying confidence malleability (*under review*)

Competitive Edge Peer Mentor *2017-Present*
Competitive Edge NSF Reviewer and Editor *2017*
Graduate InterConnect Program Peer Mentor *2014 & 2016*

Department of Psychology and Social Behavior
Graduate Student Mentoring Program Coordinator *2017-2018*
Graduate Student Peer Mentor *2016-Present*

Service Activities

American Psychology-Law Society Conference Reviewer	*2017*
UC Irvine Coordinated Governance Group Housing Committee	*2016-Present*
Society for Personality and Social Psychology Conference Reviewer	*2016 & 2017*

Department of Psychology and Social Behavior

Graduate Student Recruitment Coordinator	*2016 & 2018*
Professional Development Seminar Coordinator	*2016-2017*
Comprehensive Exam Redevelopment Task Force	*2016-2017*

Abstract of the Dissertation

A New Method of Studying Confidence Malleability:

Self-Sourced Misinformation as Post-Identification Feedback

By

Rachel Leigh Greenspan

University of California, Irvine, 2018

Distinguished Professor Elizabeth F. Loftus, Chair

Eyewitness confidence is often used by judges and jurors as a cue to accuracy. Despite this, confidence is not always related to accuracy as confidence is malleable over time and subject to suggestion. One way in which confidence can be influenced by outside factors is through post-identification feedback. The post-identification feedback effect is the finding in which participants report remembering higher confidence in their identification and a better memory for the crime when they are told they correctly identified the suspect relative to those who do not receive feedback. In the current dissertation, research on post-identification feedback is merged with studies on the misinformation effect which shows that exposing people to misleading information after viewing an event can alter their later memories for that event. The main goal of the dissertation is to investigate whether giving participants a misleading reminder about their identification confidence can affect their later recall.

Initially, two pilot studies were conducted to explore the kind of scale to best use to measure and manipulate confidence. In the main dissertation studies, participants completed a two-session experiment. In session one, they watched a mock crime video, identified the suspect from a lineup, and gave their confidence in this identification. In session two, they were reminded of their confidence. However, for some participants, this reminder was manipulated to be 20 points higher or lower than what the participant originally reported. In Study 2, the effect of this kind of misinformation feedback was contrasted with typical feedback informing the participant whether they correctly identified the suspect.

In both studies, resulted revealed that nearly all participants failed to detect the manipulation between their original confidence statement and the manipulated one provided to them. This manipulation had ripple effects such that participants led to believe their confidence was higher than originally reported later remembered having more confidence in their identification and having a better viewing experience at the time of the crime with parallel effects for the manipulation in the opposite direction. Study 2 revealed that although misinformation and typical feedback are similar, there are some differences. Specifically, typical disconfirming feedback is less powerful than typical confirming feedback, but this is not true for the two kinds of misinformation feedback. Implications for police procedure and the importance of only relying on initial confidence as a cue to accuracy are discussed.

Introduction

"The vagaries of eyewitness identification are well-known; the annals of criminal law are rife with instances of mistaken identification" (*United States v. Wade*, 1967). In the United States in 2012, law enforcement officers made over 12 million arrests with over 500,000 for violent crimes (National Research Council, 2014). Although data about the number of these cases that involve eyewitness identification are not available, it is likely to be quite large. A 1989 survey of prosecutors estimated the number of cases involving eyewitness identifications at 70,000 (Goldstein, Chance, & Schneller, 1989). This estimate may dramatically underestimate the number of eyewitness cases today as it was based off the average of 2.5 million arrests occurring each year in 1989.

Mistaken eyewitness identifications lead to several kinds of errors in the justice system. If a witness does not correctly identify the true perpetrator of the crime, then the perpetrator may fail to be arrested or convicted and go on to commit further crimes. Moreover, due to the persuasive nature of eyewitness testimony, when a witness identifies an innocent suspect this misidentification can lead to a wrongful conviction (Borchard, 1961; Loftus, 1979). There are several mechanisms by which this occurs. The overreliance on eyewitness confidence represents one such mechanism (Bradfield & McQuiston, 2004). Laypeople and jurors strongly depend on confidence as a cue to accuracy (Cutler, Penrod, & Dexter, 1990). Jurors believe the testimony of highly confident witnesses and view this evidence as strongly probative of guilt. However, even inaccurate witnesses can be highly confident and so depending on confidence as a cue to accuracy may lead to the conviction of suspects that are actually innocent.

One reason for the complex relationship between confidence and accuracy is that confidence is not one fixed value that remains stable over time (Quinlivan, Wells, & Neuschatz, 2010). Rather, confidence is malleable and a variety of factors from the time of the initial identification until the witness testifies at trial can influence people's confidence in their identification decision. While initial confidence gathered from a pristine lineup may be highly related to accuracy, confidence reports given by witnesses at later points, such as when testifying at trial, are not related to accuracy (Wixted & Wells, 2017). One common method by which witnesses' confidence reports may inflate over time is when they are exposed to post-identification feedback (PIF).

The PIF effect describes the process by which confidence inflation occurs after a witness receives confirming feedback from a lineup administrator. This typically ensues when an officer informs a witness after the identification that they correctly identified the suspect. Research has shown that confirming feedback of this type causes witnesses to retrospectively remember having greater confidence in their initial identification relative to witnesses who receive no feedback (Wells & Bradfield, 1998). Moreover, feedback causes ripple effects on memory. Compared to those who do not receive feedback, witnesses exposed to confirming PIF report not only that they were more confident at the time of their initial identification, but also that they got a better view of the suspect, that they paid more attention at the time of the crime, and other similar judgments. These memory judgments are important as they are criteria used by the courts to determine the reliability of an identification (*Niel v. Biggers,* 1972). Thus, feedback not only distorts the confidence-accuracy relationship, it also distorts other judgments that tries of fact use to determine witness reliability. This effect has been replicated throughout the research literature and meta-analyses indicate it occurs with a large effect size (Steblay, Wells, & Douglass, 2014).

The most common recommendation suggested to combat the effects of PIF is to use double-blind lineups. Double blind lineups are those in which the lineup administrator does not know the identity of the suspect. Using this form of lineup ensures that the administrator cannot provide feedback to the witness. While the use of double blind lineups eliminates the possibility of feedback from the lineup administrator at the time of the initial identification and has many other benefits in reducing intended or unintended suggestive behavior by administrators, this recommendation does not fully protect from the negative effects of feedback. Witness confidence can be affected by a variety of other factors after the initial identification procedure such as during later interviews, trial preparation, interaction with the media, or when simply recounting the event to friends and family (National Research Council, 2014). Despite this, most research on the effect of feedback on eyewitness confidence has focused on the same time point (after the initial identification). Moreover, the type of feedback used in these studies is the same: a statement by the lineup administrator that the witness identified the correct suspect.

In this dissertation, I extend the research on PIF to study confidence malleability at a new time point, a follow-up interview one week after an initial double-blind lineup. As law enforcement agencies continue to adopt double blind lineups, the likelihood of typical feedback about the accuracy of the witness' identification will decrease. Thus, it is increasingly important to investigate whether feedback at follow-up interviews has a similar deleterious effect on memory as initial feedback. To examine this, I investigated feedback in a novel form: self-sourced misinformation. This type of feedback was chosen as it may be more likely to occur during a follow-up interview whereas typical feedback is more likely to occur at the initial lineup.

The misinformation effect is the finding in which individuals exposed to misleading post-event information after viewing an event often incorporate this information into their memories (Loftus, 2005). For example, a witness to a car crash may read a newspaper article after the accident that identifies the color of the car as blue, when it was actually red. After reading this news article, the witness may now remember the color of the car as red. This witness would incorporate the inaccurate post-event information into their memories and falsely recall the true details of the accident. Post-event information can come in many forms: reading a news article, being exposed to suggestive questioning, or talking to other witnesses.

Misinformation feedback represents a unique way of studying PIF. Misinformation feedback through suggestive questioning is a more indirect kind of feedback than direct statements by the lineup administrator about whether the witness picked the suspect. Misinformation feedback is often not detected by participants and in fact is most impactful on memory when it is not noticed (Tousignant, Hall, & Loftus, 1986). Given the increased attention towards using double-blind procedures and avoiding suggestive questioning, subtle, misinformation feedback may more closely mirror the kinds of feedback witnesses in real cases receive. This type of feedback has not yet been studied and it is important to investigate as evidence suggests suggestive questioning spontaneously occurs during follow-up interviews (Maclean, Brimacombe, Allison, Dahl, & Kadlec, 2011).

Dissertation Outline

In the next sections, I review the existing literature on eyewitness confidence. I start by summarizing the research history on the relationship between eyewitness confidence and accuracy to describe the conditions under which these two variables are and are not related. In addition, I briefly discuss the contradiction between the confidence scales used in research studies and confidence statements gathered in real cases. Following this, I review the research on the PIF effect to explain how post-event suggestion can impact a witness' confidence and memory for a crime. Finally, I summarize the literature on the misinformation effect, choice blindness, and the downstream consequences of developing a false memory.

Two pilot studies are first reported regarding testing the feasibility of a new type of confidence scale that optimizes ecologically validity. The purpose of these pilot studies was to assess whether participants would understand this new scale and whether it was the type of item that could be manipulated in a misinformation study.

In the two main studies, I merge the paradigms for the misinformation and PIF effect. In Study 1, I explore how self-sourced misinformation impacts a witness' retrospective confidence in their identification. Moreover, I test whether the misinformation has downstream consequences for a witness' memory for other testimony relevant aspects of their witnessed experience. Study 2 directly contrasts misinformation feedback and typical feedback (i.e., feedback that implies the witness made a correct identification) to assess whether these two similar types of feedback result in similar effects.

Literature Review

The Relationship Between Eyewitness Confidence and Accuracy

Much of the research on the relationship between eyewitness confidence and accuracy occurred because of the U.S. Supreme Court's decision in the case of *Neil v. Biggers* (1972). This case revolved around the factors that should be used to determine whether a specific identification violates the 14th amendment's due process protection. Prior to this case, the primary factor used to determine admissibility of an eyewitness' identification concerned suggestibility (*Stovall v. Deno*, 1967). Under this standard, if the procedure was unnecessarily suggestive, then the court would suppress it. The *Biggers* case moved the criterion from suggestiveness to accuracy. Under this standard, if the procedures used created a substantial likelihood that the witness would misidentify the suspect, then the court would suppress it. The court explicated five criteria to use when assessing accuracy: view of the suspect, amount of attention paid at the time of the crime, accuracy of the witness' description of the perpetrator, time between the crime and identification, and the witness' confidence at the time of the identification. These five factors focus on the potential accuracy of the witness' identification, rather than on the presence of suggestive identification practices (Wells & Murray, 1983).

Four of these factors (view, attention, accurate description, and time elapsed) occur at the time of the crime and can only be estimated by the witness. The fifth factor (confidence) occurs during the interaction between the witness and the lineup administrator at the time of the identification. Thus, aspects of the lineup under control of law enforcement (e.g. use of witness instructions) have the potential to influence seemingly only this fifth factor (Wells, 1978). This made eyewitness confidence a particularly promising topic for study.

In early research on witness confidence, the typical study involved participant-witnesses viewing a crime video, making a lineup identification, and reporting their confidence (e.g. Leippe, Wells, & Ostrom, 1978). In most of these early studies, accuracy was dichotomized, and confidence assessed on a Likert-scale. And so, to assess the confidence accuracy relationship, researchers used the point-biserial correlation coefficient (Wixted & Wells, 2017).

These initial studies showed discouraging results for the positive relationship between confidence and accuracy, typically find either no or low positive correlations between these two variables. Correspondingly, researchers at this time strongly urged against the use of confidence in the courtroom: "I cannot reinforce strongly enough… the judicial system should cease and desist from a reliance on eyewitness confidence as an index of eyewitness accuracy" (Deffenbacher, 1980, p. 258). The first meta-analysis on this topic confirmed the tenuous relationship and showed that roughly half of the included studies found a significant, positive relationship while the other half found a null or reversed relationship (Deffenbacher, 1980). The optimality hypothesis was conceived to explain these results. This hypothesis proposed that under optimal conditions for encoding, storage, and test the confidence accuracy relationship should be strong and that the relationship weakens when conditions deteriorate. While the optimality hypothesis did not receive wide attention in the literature, it changed the emphasis of research from whether confidence and accuracy are related to focusing on conditions under which confidence may be diagnostic of accuracy.

A main moderator that affects the confidence accuracy relationship is the outcome of the lineup. Early research collapsed across all types of lineup decisions: choosing a member of the lineup or rejecting the lineup (i.e. the witness saying the suspect is not present, also called making a non-identification). These two groups are referred to as choosers (correct or filler identification) and non-choosers (non-identification), respectively. Meta-analytic results showed a distinct difference in the correlations between the confidence-accuracy relation of choosers, *r* = .41, and non-choosers, *r* = .12 (Sporer, Penrod, Read, & Cutler, 1995). Results also showed that the average level of confidence for correct choosers was higher than for incorrect choosers. This led the authors to conclude that "when limited to witnesses who make positive identifications under laboratory conditions, confidence appears to be a somewhat stronger predictor of accuracy" (Sporer et al., 1995, p. 322), a dramatic shift from the conclusions of the research literature just 15 years prior.

Conclusions about confidence and accuracy continued to shift after the publication of an influential paper by Juslin, Olsson, and Winman (1996) that suggested that the point-biserial correlation was not the appropriate method to analyze data for this topic. The point-biserial correlation answers the question of whether witnesses who make a correct identification are, on average, more confident than witnesses who make an incorrect identification. However, the legal system asks a different question: if an eyewitness claims they are X% confident, how likely is it that their identification is accurate (Wixted & Wells, 2017). Calibration, rather than the point-biserial correlation, answers this question. Juslin et al. (1996) showed that calibration can be high for eyewitnesses even when the point-biserial correlation is low due to the restricted range and unimodal distribution of typical confidence data.

In a recent new synthesis of the literature, Wixted and Wells (2017) suggest a new set of conditions important to the confidence accuracy relationship. Their review concludes that when eyewitnesses make an identification from a "pristine" lineup with high confidence, then the witness' confidence is a very good signal of their accuracy. They outline five conditions for a pristine lineup: that the lineup has only one suspect, that the suspect does not stand out, that officers instruct the witness that the suspect may or may not be in the lineup, that the lineup is double blind, and that officers obtain the confidence statement immediately after the initial identification. If an eyewitness makes a positive identification under these conditions, then high confidence indicates high accuracy. However, if these conditions are not met, then the confidence-accuracy relationship of even highly confident witnesses can be compromised. This does not necessarily mean that the confidence-accuracy relationship under non-pristine conditions is always impaired, just that it can be (Mickes, Clark, & Gronlund, 2017). Future research is necessary to indicate confidence calibration when only some combinations of conditions for pristineness exist.

Whereas the relationship between high confidence and accuracy depends on the "pristineness" of the identification context, the relationship between low confidence and accuracy does not. Wixted and Wells (2017) assert that low confident identifications are always low value. Regardless of the conditions, identifications made with low confidence always suggest the possibility of error. The lack of pristine testing conditions affects the confidence accuracy relationship by interfering with the cues that eyewitness use to create their confidence statement. In a pristine lineup, witnesses can only rely on strength of their own memory when judging their confidence. When non-pristine conditions are used, such as non-blind lineups, these outside factors provide additional cues to witnesses that they can rely on when determining their confidence. This impairs the relationship between confidence and accuracy.

One of the most important criteria for lineup pristineness is that the officer documents the confidence statement immediately after the identification (Loftus & Greenspan, 2017). While converging evidence now points to a strong relationship between confidence and accuracy under pristine conditions, this relationship only exists when officers gather the witness' confidence statement immediately after the identification (Wixted & Wells, 2017). Years of research has demonstrated the malleability of witness confidence and so only the initial confidence statement given immediately after the identification from a pristine lineup relates to accuracy (Steblay et al., 2014). The National Research Council (2014) specifically recommends that "law enforcement document the witness' level of confidence verbatim at the time when she or he first identifies a suspect, as confidence levels expressed at later times are subject to recall bias, enhancements stemming from opinions voiced by law enforcement, counsel and the press, and to a host of other factors that render confidence statements less reliable" (p. 74).

Verbal and Numeric Confidence

This recommendation from the NRC about verbatim confidence statements means that when witnesses describe their confidence, officers document it in the witness' own words.
For instance, after picking someone from a lineup, a witness might say that they are "mostly sure" in their choice. If asked to translate this to a number, the witness might state they are 80% certain. Documenting confidence verbally, rather than numerically, has several advantages. Police and prosecutors may prefer this method as it conveys less of a possibility for error than numeric judgments. When conveying the "80%" judgment to a jury, the possibility for error is highlighted while the possibly of error in the verbal judgment ("mostly sure") is much harder to quantify.

Verbal confidence reports may also be preferred as, in general, people feel more comfortable reporting probabilistic judgments in their own words rather than with numbers (Renooij & Witteman, 1999). However, a paradox exists between those conveying probability judgements and those receiving them. In a study of patients and physicians, physicians preferred conveying probabilistic information in words (e.g., “probable” likelihood of illness) but patients preferred receiving this information in numeric form (Brun & Teigen, 1988).

Despite the recommendation of reports from the NRC to document confidence in the witness’ own words, nearly all published research on witness confidence uses numeric scales. Researchers likely prefer numeric scales for a variety of reasons. Verbal expressions of confidence cannot easily be rank ordered like numeric expressions and may be more susceptible to the effects of context (Renooij & Witteman, 1999). This problem is compounded by the fact there is an unlimited range of verbal responses witnesses can choose to use to explain their confidence (Hamm, 1991). These factors combine and underly the point that numeric responses are by far easier and more straightforward to analyze and explain than verbal responses.

While these factors help to explain why researchers often chose numeric scales to gather confidence from participants, they also highlight a potential problem in that empirical results, and thus the related procedural recommendations, based on numeric responses may not generalize to the confidence used in real cases. A handful of studies have focused on the question of whether verbal and numeric confidence judgments produce similar results. One important aspect of these kinds of studies is the method by which participants give their verbal confidence. While numeric scales can vary from 6-, 11- or 101-point measures, the range of possible options for verbal responses is infinitely larger. To attempt to standardize these answers, rather than using free response, most researchers studying verbal confidence use a prompted verbal scale. This type of scale is usually Likert-type with all points along the spectrum labeled ranging from no confidence to complete confidence. The design of these scales is critical to their use. If the scale has too many response options, is not ordered properly, or uses language not typically expressed by participants, then the advantage of greater ecological validity using this measure might be outweighed by other methodological concerns.

Several methods have been used to develop prompted verbal scales. In one of the more ecologically valid studies, participants rated confidence statements given by witnesses from real cases gathered through archival analysis. Through these ratings, verbal responses were coded into low, medium, and high confidence groups. For example, “I am not sure” was coded as a low confidence response, “moderately sure” was coded as medium confidence, and “very sure” was coded as high confidence. Once a potential word ordering has been developed, testing often involves giving participants the confidence list and asking them to translate each response option into a number. These numeric translations are then rank ordered to confirm that participants have a common understanding of the scale (Wesson & Pulford, 2009; Windschitl & Wells, 1996).

Most of the research about creating prompted verbal scales has not been in the area of witness confidence. Rather these studies tend to focus on judgments about communicating probability of likelihood (Wesson & Pulford, 2009). Only a few studies in this area focus on witness confidence. In one study testing the difference between verbal and numeric scales, participants watched a series of mock crime videos and made identifications from either target-present or target-absent lineups (Weber, Brewer, & Margitich, 2008). After these lineups, participants gave their confidence either numerically or on a prompted verbal scale. This prompted verbal scale used was based on that from Windschitl and Wells (1996) and included 11 options ranging from *impossible* to *certain*. Calibration curves showed no difference in the confidence-accuracy relation between the verbal and numeric scale. Dodson and Dobolyi (2016) tested the confidence-accuracy relation using prompted verbal scales that both varied in the number of response options and varied in how many scale points were labelled. Calibration curves revealed that scale format did not affect the confidence-accuracy relationship.

In addition to the issues with scale development, despite their ecological validity, verbal scales are challenging to use as they may not be as well understood by jurors as numeric responses. In one study, a subgroup of participants acted as witnesses and gave a verbal free report of their confidence and then translated this into a number (Dodson & Dobolyi, 2015). A second group of participants then read these verbal responses and guessed the numeric response the witness meant by their verbal report. These numeric translations were then matched with what the witness intended number. Results showed that participants consistently underestimated the intended numeric response by the witness. Underestimation increased as the witness' intended confidence level increased. Moreover, variability in estimations increased when participants gave a justification for their answer. So verbal statements that included an explanation in addition to a statement of confidence (e.g. "I am very certain. I remember his hair.") produced greater variability in translation by participants than if the justification had not been included. This was especially true for witnesses who made a non-identification.

Overall, despite the fact that real witnesses are giving their confidence in their own words, experimental research continues to nearly solely focus on numeric confidence. The research reviewed here indicates reasons why—numeric confidence is easier to analyze and may be more easily understood by others. However, it is important to confirm that the findings developed and tested under numeric confidence conditions replicate to verbal confidence. Prompted verbal scales represent one way to meet this gap. They more closely approximate the kinds of response participants give in the real world but retain the advantage of numeric scales in standardized ordering.

Much of the past research on verbal versus numeric confidence judgments focuses on whether scale form impacts the confidence-accuracy relationship (Weber et al., 2008). Another robust finding in the eyewitness literature that could be studied using a prompted verbal scale is PIF. A central aspect of the PIF is confidence inflation. Confidence modality might play a role in that feedback might differentially impact verbal and numeric responses. Furthermore, in their review paper, Wixted and Wells (2017) argue that "perhaps the biggest threat to our ability to rely on confidence in eyewitness identification occurs when witnesses receive post-identification feedback that suggests they made an accurate identification" (p. 18). Given the significant policy implications of PIF research and its importance in establishing a strong confidence-accuracy relationship, it is particularly important that PIF research expand to study confidence in the way it is gathered in the field. Next, the literature regarding the PIF is reviewed.

Post-Identification Feedback

The seminal study on PIF tested the effects of informing a witness about the accuracy of their identification before they reported their confidence in their lineup choice (Wells & Bradfield, 1998). In this study, participants watched security camera footage and made an identification of the suspect from a target absent lineup. Immediately after, participants received either confirming feedback (e.g. good you identified the suspect), disconfirming feedback (e.g. sorry the actual suspect was someone else), or no feedback. Following this, participants responded to a series of eleven dependent measures about their witnessed experience. These questions assessed three broad categories: "qualities of the witnessed event," "qualities of the identification task," and "summative qualities of the viewing experience" (Wells & Bradfield, 1998, p. 362). Qualities of the witnessed event included items such as whether the witness had a good view of the crime and how much attention they paid during the event. Qualities of the identification task included judgments such as the reported ease in which the witness made their identification and, critically, their confidence in their identification. The final category, summative qualities of the viewing event, was a broader category that included questions such as how willing the witness would be to testify about their identification in court. All these items assess witnesses' retrospective memory (e.g. how much attention did the witness pay at the time of the crime, what was the witness' confidence at the time of the identification).

Results showed that, compared to the no feedback condition, witnesses who received confirming feedback reported more confidence in their identification (Wells & Bradfield, 1998). Confirming feedback not only affected witnesses' retrospective confidence, but also inflated their reported view of the suspect, their ease of identification, and several other of the dependent measures. Overall, confirming feedback inflated witness' judgments for ten out of the eleven studied variables. The effect sizes for this manipulation were quite strong for confidence (d = 1.56) as well as for the composite of the ten significant measures (d = 0.75). Disconfirming feedback had a weaker effect on witnesses' retrospective memory than confirming feedback. Compared to the no feedback condition, only four of the eleven dependent variables (not including confidence) differed significantly from the control condition and with a much smaller effect size (d = 0.27).

This study first assessed the effects of confirming and disconfirming feedback on witnesses' retrospective confidence. Confirming feedback strongly affected witnesses' memory for their confidence as well as for other aspects of their witnessed experience. Some of these factors, such as attention and view, were specifically identified by the Supreme Court as key to assessing an eyewitness' accuracy and reliability (*Niel v. Biggers*, 1972). This study showed that feedback about the accuracy of the witness' identification can not only affect their memory for their confidence but can also have downstream affects for other key criteria for evaluating their testimony (Wells & Bradfield, 1998). The term *post identification feedback effect* has been used to describe this class of findings.

PIF differentially affects accurate and inaccurate witnesses. Social psychological theory provides a possible explanation as to why this occurs. According to Bem's (1972) self-perception theory, people determine their attitudes from their behavior when they have weak internal cues. Thus, if a person's memory for an event is strong, they should rely on their own memory for the event to determine their confidence (Bradfield, Wells, & Olson, 2002). If a person's memory for an event is weak or decayed, then their own memory only provides weak evidence to support their confidence judgment. In this case, PIF may strongly influence the witness as this feedback provides an external cue for how confident the witness should be. Experimental research supports this proposition by showing that confirming feedback distorts the confidence-accuracy relationship by inflating the confidence of inaccurate, but not accurate witnesses (Bradfield et al., 2002).

Most studies that investigate PIF use target absent lineups with biased instructions (i.e. the researcher gives the witness a lineup without the true suspect and does not inform the witness the true suspect may not be in the lineup). While this form of lineups is largely regarded as suggestive and goes against best practice recommendations, using target absent lineups with biased instructions maximizes sample size for experimental studies (Smalarz & Wells, 2014). Using target-present lineups with unbiased instructions splits the sample between correct identifications, filler identifications, and non-identifications. A large initial sample size would be necessary to test for interaction effects between these groups, specifically because of the likely uneven distribution of identification outcomes. However, in the field, target present lineups with unbiased instructions are likely to be used and so it is important to confirm whether the PIF effect holds true under these conditions. Semmler, Brewer, and Wells (2004) tested the effect of PIF using unbiased witness instructions for both target absent and target present lineups. Results revealed that the typical PIF effect (i.e. confidence inflation and ripple effects on other aspects of the witnessed event) occurred for both kinds of lineups and for all types of lineup decisions: correct identification, mistaken identification, and lineup rejections. PIF effects also occur for other kinds of identification tasks such as show-ups (Key, Wetmore, Cash, Neuschatz, & Gronlund, 2017) and for earwitness identifications (Quinlivan et al., 2009).

PIF is a robust effect occurring under a wide range of testing conditions. In one study, confirmatory feedback inflated witness confidence even when assessment of confidence happened several days later as well as when witnesses received delayed feedback (Wells, Olson, & Charman, 2003). Critically, research shows that confidence inflation caused by PIF occurs with real witnesses to crimes (Wright & Skagerberg, 2007).

The Selective Cue Integration Framework. A predominant theoretical account to explain the effects of PIF is the Selective Cue Integration Framework (Charman, Carlucci, Vallano, & Gregory, 2010). The Selective Cue Integration Framework (SCIF) proposes three stages by which confidence inflation occurs. Many of the proposals of the SCIF originate from classic findings of the attitude change literature (e.g. Nickerson, 1998; Petty, Haugtvedt, & Smith, 1995). The first stage, the assessment stage, has two main claims. The first proposes that witnesses do not spontaneously assess their confidence, but rather actively construct their confidence judgment only when asked to report it. The second claim mirrors that of early researchers studying PIF. This claim states that when witnesses have strong internal cues they will rely less on external cues.

In the proceeding search stage, witnesses with weak internal cues search for external cues or information to support their decision. This information search process does not occur in an unbiased manner, rather people will seek out and accept confirming evidence and criticize and discount disconfirming evidence (Lord, Ross, & Lepper, 1979). Witnesses with weak internal cues search for external sources to support their identification decision discounting those that disconfirm their identification but considering those consistent with their identification. This stage provides an explanation for the consistent finding in the PIF literature that confirming feedback has a stronger effect than disconfirming feedback (Steblay et al., 2014). According to the SCIF, witnesses heavily scrutinize disconfirming feedback and discount it as it clashes with their identification (Charman et al., 2010). Thus, this feedback is discounted and not integrated when witnesses construct their confidence judgment.

If an external cue confirms the witness' identification decision, then the final evaluation stage occurs. Here, witnesses appraise the external cue from the search stage to check for the presence of factors that may challenge the cue's credibility. For instance, if the cue comes from a contaminated source, it will be discounted (Skagerberg & Wright, 2009). But if there are no factors that undermine the credibility of the external cue, witnesses will integrate this cue when constructing their confidence statement.

Types of feedback. In the seminal study on PIF and much of the research that followed, the feedback witnesses received was in the form of a statement by the lineup administrator about the accuracy of the witness' identification. Confirmatory feedback told the witness "Good. You identified the actual suspect." whereas disconfirming feedback instructed the witness "Actually, the suspect was number X" (Wells & Bradfield, 1998, p. 363). Much of the literature investigating the effect of PIF has used similar, if not identical instructions (e.g. Charman et al., 2010; Quinlivan, Neuschatz, Douglass, Wells, & Wetmore, 2011). Because of this, PIF and feedback about the accuracy of an identification may be seen as identical constructs. However, feedback can and does come from a variety of sources and during varied timepoints.

Lineup feedback can come from sources other than the lineup administrator. After witnesses make an initial identification, they are exposed to a broad array of influences during which feedback can occur. If the witness identifies the suspect, it is likely they will be brought back into the police station to review their statement as the case progresses. They will be deposed and cross-examined by the prosecution and/or defense lawyers. Through all of this, numerous opportunities exist for feedback to influence the witness. Despite this, only a handful of studies have investigated the effects of feedback from sources other than the lineup administrator, such as from a co-witness (e.g., Luus & Wells, 1994). In one study, participant-witnesses were exposed to a simple briefing prior to cross-examination (Wells, Ferguson, & Lindsay, 1981). This briefing contained no leading information and simply warned the participant that the lawyer would try to discredit their testimony and would likely ask about details of the event. Even this non-suggestive manipulation caused witnesses to increase their retrospective confidence and this effect was mainly driven by inaccurate witnesses. Real briefings by prosecutors and defense attorneys may be more suggestive than this. Most defense attorneys and many prosecutors are aware of the malleability of confidence, but also believe that jurors are unaware of this fact (Wise, Pawlenko, Safer, & Meyer, 2009). This creates a situation in which the use of leading questions about confidence may be particularly likely to occur.

Mitigating the PIF effect. Given that PIF from a variety of sources can have destructive effects on witness memory, researchers have investigated whether and how the effects from PIF can be reduced or eliminated. The SCIF provides some testable hypotheses about how PIF may be moderated. The first stage of SCIF suggests that if witnesses have strong internal cues then they will base their confidence decision off this strong memory trace and will thus not be influenced by feedback (Charman et al., 2010). Therefore, if it is possible to solidify this internal memory trace, then feedback may have a reduced effect.

Using a prior thought manipulation provides one way in which to manipulate strength of a memory trace. This manipulation instructs participants to think about their confidence, view, and other factors after making their identification but prior to receiving feedback. The intent is to strengthen the witness' own recall and cue them into assessing their own confidence prior to receiving feedback. One type of prior thought manipulation is the confidence prophylactic which asks participants to specifically think about their confidence prior to feedback (Steblay et al., 2014). Results show that witnesses asked to think about factors such as confidence prior to feedback show less confidence inflation than those not given this instruction (Wells & Bradfield, 1999). However, if prior thought manipulations occur after feedback, this eliminates the beneficial effect of this process. This provides further evidence for the SCIF proposition that witnesses do not construct their confidence statement until asked to do so. However, the prophylactic effect of prior thought only inoculates witnesses from confidence inflation in the short term. When confidence measures were delayed one week, the benefit of prior thought was eliminated and participants showed the typical PIF effect (Quinlivan et al., 2009). This rebound effect possibly ensued because the witness' memory of the prior thought manipulation deteriorated faster than their memory for the feedback.

Another possible stage of the SCIF in which the PIF effect may be reduced is during the evaluation stage (Charman et al., 2010). If witnesses believe that the external cue (i.e. feedback) comes from a contaminated source, they should discount this cue. In laboratory test of this proposal, researchers induced suspicion by leading witnesses to believe that the administrator who gave the feedback was not accurate or truthful. When witnesses became suspicious of the source of the feedback, they did not show the PIF effect (Quinlivan et al., 2010).

While prior thought and post-feedback suspicion manipulations show some evidence of working in the lab, they are unlikely to be practical solutions in the field as the manipulations would be atypical behavior for officers in real cases. Some solutions researchers have proposed are more applied and less theoretically based. The most common recommendation to reduce the PIF effect is that officers conduct double-blind lineups in which the lineup administrator does not know the identity of the suspect (Steblay et al., 2014). This eliminates the possibility of feedback from the lineup administrator as the administrator inherently cannot provide accuracy feedback to the witness. Double-blind lineups are particularly effective when the witness is informed that the administrator does not know who the suspect is. In this way, ambiguous feedback such as "you've been a good witness" can be discounted by witnesses as indicating whether they picked the correct suspect (Dysart, Lawson, & Rainey, 2012).

Jurors' perception of confidence inflation. Using double-blind lineups enables officers to obtain a clean confidence statement from a witness that occurs without suggestion or feedback. If suggestion or feedback occurs later, such as during pre-trial briefings, then at least there exists recorded evidence of the witness' original, pre-feedback confidence. The logic follows that if confidence inflation occurs prior to trial, then the witness' original statement can be introduced as evidence to inform the trier of fact about how the witness' confidence changed over time (Jones, Williams, & Brewer, 2008). Ideally, jurors would rely on the initial, unbiased confidence statement and not later reports that may be the result of suggestion rather than the witness' original memory. Empirical evidence on whether jurors do this is mixed. Initial evidence suggested that jurors take into account this inconsistency (between initial and later confidence) and generate evaluations more favoring the defense (Bradfield & McQuiston, 2004).

However, at trial, when jurors hear multiple confidence statements, there are a variety of ways they can reconcile this inconsistency. They may believe the witness' confidence increased over time for invalid reasons, such as that the witness wants the perpetrator to be convicted and so purposefully increased their confidence. On the other hand, they may believe the witness' confidence increased for valid reasons, such as that when the witness thought about the situation over time they genuinely became more confident by realizing new information about the crime (Jones et al., 2008). This latter explanation is called a confidence epiphany. These explanations play a role in how jurors evaluate testimonial inconsistency. Jurors who attributed the inconsistency of a witness to a confidence epiphany rated that witness more favorably compared than those who attributed the inconsistency to other reasons. Nevertheless, only a few studies have investigated the confidence epiphany explanation with varied results as to its effectiveness (Jones et al., 2008; Paiva, Berman, Cutler, Platania, & Weipert, 2011).

Not only does the explanation jurors receive about the cause of confidence inflation matter, but the method by which they receive this information also plays a role in how they evaluate confidence inconsistency. When jurors learned about confidence inflation by watching a video of the witness' original confidence statement, this was more influential in their evaluations of the witness than when this information was presented via a written transcript (Douglass & Jones, 2013). This effect occurred because, in the video materials, jurors had access to information about the witness' non-verbal behaviors that was not available in the written transcript. These non-verbal behaviors, rather than the inconsistency between the two statements, mediated the relationship between condition and evaluation of the witness.

One notable factor in these studies is the focus on confidence. These studies investigate how jurors view confidence inflation in the form of multiple confidence statements over time. The main proposition being that if jurors rely solely on the first, unbiased confidence statement then double-blind lineups effectively eliminate the negative effects of feedback. Yet feedback affects more than witness' retrospective confidence. It also affects their reported recollections of view, attention, and other forensically relevant variables (Smalarz, 2015; Steblay et al., 2014). As these recollections may be less often documented than confidence, it is unlikely they could be introduced as evidence and no research has been conducted on how these statements would be perceived by jurors. Thus, entering a witness' initial confidence statement as evidence at trial, even if effective, would only help prevent a portion of the negative effects of feedback.

Overall, the findings about how to reduce or eliminate the PIF effect are discouraging. Some solutions such as prior thought manipulations and post-feedback suspicion focus on reducing the PIF effect at the individual level. While there is some evidence of the effectiveness of these measures, they are more theoretically based and unlikely to be used in the field. Other recommendations focus on reducing the negative impact of post-identification feedback at the officer or jury level by using double-blind lineups and introducing the witness' first confidence statement as evidence at trial. There is inconclusive evidence about the effectiveness of this remedy. Some results suggest that jurors focus on the initial confidence statement by downgrading their assessment of a witness who shows confidence inflation. But this effect does not always carryover to influence verdict decisions and the results vary by what method the initial confidence statement is show to jurors (Bradfield & McQuiston, 2004; Douglass & Jones, 2013). Moreover, if witnesses report a genuine reason for their inflated confidence, this may eliminate the beneficial effect of showing the initial confidence statement (Jones et al., 2008).

However, just because jurors do not completely account for the effect of confidence inflation when exposed to inconsistent witness statements does not mean that using double-blind lineups should be discouraged. Double-blind lineups have a wide array of benefits in ensuring an unbiased lineup procedure and is consistently recommended as a best practice that should be used by all law enforcement agencies (National Research Council, 2014; Wells et al., 1998). Indeed, double blind lineups do effectively prevent against feedback at the time of the lineup procedure as, intrinsically, blind lineup administrators cannot provide feedback as to the witness' accuracy. Yet, this recommendation alone does not protect against confidence change over time.

Timing of feedback. One gap in the PIF literature is the focus primarily on feedback at the time of the lineup. This focus may have spurred from the procedures officers used during the initial time that the initial PIF studies were conducted. During this time, instructions warning the witness that the perpetrator may not be present in the lineup and the use of double blind lineups occurred infrequently. Since the late 1990s, there has been a dramatic shift in the adoption of reforms to improve eyewitness evidence. More agencies now use double-blind lineups and standardized witness instructions (Police Executive Research Forum, 2013). These reforms both help improve the overall quality of eyewitness evidence and specifically reduce the probability of feedback from the lineup administrator. As law enforcement agencies continue to adopt these reforms, the probability of initial PIF will continue to diminish.

This fact increases the importance of studying PIF at later time points. Witnesses inherently will receive feedback about the accuracy of their identification at some point either directly or indirectly, such as during secondary interviews (Wells et al., 2003). In fact, simply learning that the person they identified has been charged with the crime can be viewed as a type of feedback for a witness (Steblay et al., 2014). While researchers typically acknowledge that feedback can really occur anytime between an identification and trial, little to no research has tested this question empirically. This understudied question is of particular importance as it would provide some of the first evidence about how confidence inflation can occur even when double-blind procedures are used to elicit the initial confidence statement. This topic will become more important as the use of double blind procedures continues to increase. Witness confidence inflation will likely still occur and so it is important to document and understand this process so that new procedures can be developed and tested to alleviate the impact of a broader range of feedback.

Although PIF can occur from a variety of sources (e.g. lineup administrator, co-witness, prosecutor), this does not mean each of these manipulations tests a different effect. Rather, the convergent findings about feedback from a variety of sources suggests that the cognitive processes that cause confidence inflation “transcend the specific paradigm in which the phenomenon is examined and the specific manipulations used within those paradigms” (Charman et al., 2010, p. 214). So, all kind of feedback may operate through similar cognitive mechanisms. One untested mechanism by which confidence inflation may occur is suggestive questioning. At follow-up interviews with police, witnesses may be asked to again recount their memory for the crime and their identification (Maclean et al., 2011). Even if the initial identification is double-blind, these later interviewers likely occur with principal investigator of the case. If the case goes to trial, this process continues with pre-trial briefings by prosecutors.

Another source of suggestive feedback may come from follow-up interviews by police after the initial identification. Even in the initial identification is pristine and double-blind, the witness is likely to have further interaction with police after the identification, especially if they identify the suspect. During this interview, the officer may summarize what the witness has said and ask for more information. In one of the only studies on this topic, researchers randomly assigned participants to the role of witness or investigator (Maclean et al., 2011). Witnesses watched a mock crime video alone and were given typical confirming feedback about their identification or not. They then were interviewed by the participant investigator. The investigators were given general guidelines about the kinds of questions to ask the witnesses but were encouraged to develop their own questions and style. Later, a separate group of participants watched the videotaped interactions between the witness and investigator and rated the manner in which the investigator asked the witness about their identification confidence. Results revealed over half of investigators asked about confidence in a leading manner. This included questions such as “so you are pretty confident in the choice you made?”

This study shows that, after an initial identification, investigators may freely ask about confidence in a suggestive fashion. This may be particularly likely to occur when confidence is gathered verbally rather than numerically. Investigators may unintentionally summarize the witness’ confidence statement using their own words in a manner that implies the witness was more confident that what they originally reported. While this proposition has not yet been tested in the eyewitness literature, other studies regarding memory malleability have investigated how presenting participants with a modified version of their own memory report can cause changes in their later memory. Studies on the misinformation effect and the choice blindness phenomenon demonstrate how suggestive questioning can lead to changes in witness’ memory for previously seen events.

The Misinformation Effect

The misinformation effect describes the set of findings in which participants exposed to misleading post-event information can incorporate this information into their memories and recall it as part of their memory for a true event (Loftus, 2005). In the classic study of this effect, participants viewed a slideshow of an automobile accident. One of these slides depicted a car stopped at a stop sign. Participants later completed a memory test during they read which several misleading questions. One of these implied that the car in the slideshow had stopped at a yield sign. Participants exposed to this misleading question were significantly more likely to report remembering seeing a yield sign in the slideshow compared to participants who did not receive this leading question (Loftus, Miller, & Burns, 1978). Participants incorporated the details from the misleading question into their memories and it caused distortion in their memory for the original event.

In many ways, PIF can be conceptualized as a form of post-event suggestion (Sagana, Sauerland, & Merckelbach, 2014; Wells et al., 2003). Like post-event information in a misinformation study, PIF occurs after the original event and suggests inaccurate information beyond the witness' memory of the event. In both misinformation and PIF studies, this post-event information causes changes in the witness' memory compared to participants who do not receive post-event suggestion. In typical misinformation studies, the impact of the post-event information is tested directly. For instance, post-event information implies a yield sign and participants' memory for the type of sign is tested (Loftus et al., 1978). In PIF studies, the effect of the post-event information is tested in a more indirect manner. The feedback contains suggestions about the accuracy of the witness' identification but, rather than testing witness' memory for their identification, PIF studies are primarily interested in testing witness' memory for their confidence and viewing experience. This tests more of the downstream consequences of post-event information.

Reducing the misinformation effect. Over the past 40 years, the misinformation effect has been studied in a wide array of circumstances and this research has shown a variety of ways to increase or decrease the chances of misinformation susceptibility. For instance, misinformation is particularly effective when time has elapsed between the original event and the post-event questioning as well as when the final memory test occurs close in time to the post-event information. With an increased retention interval between the event and the misinformation, memory for the original event has more of a chance to decay. Also, because the test is given shortly after the misinformation, the memory for this information is particularly salient (Loftus et al., 1978). This is similar to research on PIF that shows that people are more affected by feedback when they have a weak internal memory of the event (Charman et al., 2010).

Misinformation also has the greatest influence when it goes unnoticed. According to the Discrepancy Detection Principle, misinformation is most likely to be incorporated into a witness' memory when they do not detect the discrepancy between the original event and the post-event information (Tousignant et al., 1986). In this way, misinformation feedback differs from PIF. PIF is direct and explicitly intended to be recognized and attended to by participants. Misinformation feedback is subtler and most effective when it is not noticed by participants.

Like PIF, efforts to reduce the misinformation effect have focused both on manipulations that occur prior and subsequent to the feedback or suggestion. Efforts to contest the effect of misinformation after the fact have proven largely unsuccessful in completely combating its influence. Once participants incorporate misinformation into their memories, it is difficult to remove. Warnings after the fact about the presence of misinformation do not cause a witness' memory to revert to its pre-suggestion state. Warnings prior to misinformation have shown some success most particularly if they are specific to the effects of misinformation and not just a general warning that some false information may be present (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). This may operate on a similar process suggested by the SCIF. When participants receive a warning about the presence of misinformation, they can proactively discount this information and prevent it from affecting their memory.

In a typical misinformation study, the key analysis of interest compares the rate at which participants remember the post-event misinformation in the control as compared to the experimental condition. If participants in the misinformation condition report remembering the false information at a significantly higher rate than participants in the control condition, then this demonstrates the misinformation effect. Much of the research in this area has focused on topics such as under which conditions misinformation is most likely to occur, who is most susceptible to misinformation, the effects of different paradigms on creating false memories (e.g. Patihis et al., 2013; Tousignant et al., 1986; Wylie et al., 2014). Comparatively little research has concentrated on what happens after the false memory has been implanted.

Consequences of misinformation. Interestingly, one of the first studies on the effects of leading questions on memory did investigate this question. In this study, after viewing a slideshow of a car accident, participants read leading questions suggesting that that the car was moving at a high speed when the accident occurred (Loftus & Palmer, 1974). Finding the typical misinformation effect, participants exposed to the leading questions reported remembering the car traveling faster than participants exposed to non-leading questions. At a follow-up test, participants reported whether they remembered seeing broken glass at the scene of the accident which was not shown in the slideshow. Participants exposed to the misleading information were more likely to report remembering broken glass at the scene. This study provided the first empirical test of the consequences of implanting false information. Like PIF, not only did the leading questions cause participants to report that the car was moving faster (information directly suggested by the leading question), but it also caused participants overall memory for the scene to change in ways not directly implanted by the misinformation (remembering broken glass).

Despite this study occurring early in the literature on the development of false memories, few other studies initially investigated the consequences of false memories. Nonetheless, several significant questions can be answered by exploring the ripple effects of misinformation. Determining whether false memories can have later consequences on a person's thoughts, beliefs, or behaviors may provide an avenue for distinguishing true and false memories if false memories do not have these long-term effects (Laney & Loftus, 2017). This topic of study can also provide evidence as to whether the misinformation effect occurs because of demand characteristics. If participants in misinformation studies detect the presence of post-event information, then it is possible that they report remembering this information not because it is incorporated into their memories, but because they are trying to meet the hypotheses of the researcher. Demonstrating that misinformation has effects beyond just the memory test for the target item indicates that demand characteristics do not primarily drive these effects. Similarly, studying this topic also addresses concerns that misinformation studies investigate false beliefs or increased confidence in a false event rather than a true false memory (Smeets, Merckelbach, Horselenberg, & Jelicic, 2005).

Driven by some of these questions, researchers began to again study this topic in the mid-2000s. Many of these studies focused on the downstream consequences of implanting misinformation about eating behaviors. In these studies, participants came into the lab and completed a food inventory questionnaire (Bernfstein, Laney, Morris, & Loftus, 2005). They later received false feedback suggesting that, as a child, they had gotten sick eating hard boiled eggs. Those who received this feedback reported more confidence in experiencing this event as a child compared to control participants. For those participants who believed this feedback, the implanted false belief affected their reported intention for future eating behaviors in that they not only reported less intention to eat hard boiled eggs in the future but also less intention to eat related foods such as egg sandwiches. Later researchers also successfully convinced participants that they enjoyed eating a healthy food (asparagus) as a child (Laney, Morris, Bernstein, Wakefield, & Loftus, 2008). This false belief effected participants reported willingness to order asparagus at a restaurant and caused them to report being willing to pay more for asparagus at the grocery store.

False beliefs about eating behavior not only affect behavioral intentions, but also actual behavior. Participants led to believe they got sick eating egg salad sandwiches as a child consumed less sandwiches from a buffet one week later (Geraerts et al., 2008). Moreover, this effect persisted over an extended period. At a four-month follow-up session, participants who believed the initial feedback consumed less egg sandwiches than control participants. This behavioral effect has been replicated with other kinds of food (e.g. peach yogurt; Scoboria & Bernstein, 2011). A key factor in the effectiveness of this procedure is that the suggestion is individualized for each participant (Scoboria et al., 2012).

While the study of the consequences of false memories has primarily occurred in the domain of eating behaviors, a handful of other studies have followed in other domains. Similar to the finding that participants were willing to pay more for asparagus after believing the suggestion that they loved asparagus as a child, participants who fell for a suggestion that they had a negative experience with Pluto at Disneyland reported that they would pay less for a Pluto stuffed animal compared with the control group (Berkowitz, Laney, Morris, Garry, & Loftus, 2008; Laney et al., 2008).

In a unique field study of the behavioral consequences of false memories, researchers used a combination of suggestive memory techniques in order to improve children's memory and behavior during routine dental exams (Pickrell et al., 2007). These techniques included giving the child concrete examples of their positive behavior during the visit as well as verbalization, in which the researcher playacted how the child could tell their parent how well they did at the dentist. The child then did this after the exam. Children in the experimental condition reported feeling less scared and less pain compared to controls. Moreover, when observers watched videotapes of the dental visits, children in the experimental condition were rated as behaving better than controls. Thus, the memory manipulation techniques not only affected children's memory for the procedure, but also affected their reported pain and their behavior in ways that independent observers could detect.

As the dental study emphasizes, a variety of suggestive techniques can be used to implant post-event information into memory. These techniques include modified co-witness report, news articles, and guided imagination (e.g. Garry, Manning, Loftus, & Sherman, 1996; Paterson & Kemp, 2006). In all of these forms, the post-event information occurs from a source originating outside of the self (Stille, Norin, & Sikstro, 2017). When a participant's own memory report is manipulated and then given back as the source of the post-event information, this is called self-sourced misinformation and is studied through the choice blindness paradigm.

Choice Blindness and Self-Sourced Misinformation.

Choice-blindness is the phenomenon in which participants are often unaware of a mismatch between their intended choice and the outcome presented to them (Johansson, Hall, Sikström, & Olsson, 2005). Unlike misinformation, the choice blindness paradigm originates from decision theory research. Because of this, much of the early research on choice blindness manipulates a person's preference, decisions, or beliefs on attitude measures rather than manipulating their own memory reports.

In the initial study of choice blindness, participants viewed two female faces and chose which they found more attractive (Johansson et al., 2005). The researcher then handed the participants this picture and asked them to explain their choice. Unbeknownst to participants, the researcher used a sleight-of-hand manipulation and the picture the participant received was actually the non-chosen option. Only 26% of participants detected this manipulation. Participants who detect the manipulation are referred to as detectors and participants who fail to detect the manipulation are referred to as non-detectors. Moreover, participants often confabulated reasons for selecting the picture they received even though this was originally the picture they rejected. This failure to notice the discrepancy between intention and outcome constitutes the choice blindness effect.

In a follow-up study, after participants completed ratings of the facial pairs and justified their decisions, they completed a second round of ratings in which they again viewed the original pairs of faces and chose which they found more attractive (Johansson, Hall, Tärning, Sikström, & Chater, 2014). On non-manipulated trials, participants showed high consistency, choosing the same face as more attractive on both trials 93% of the time. For manipulated trials, consistency dropped to 57%. This shows that, like misinformation and PIF, choice blindness manipulations have downstream effects for participants' preferences and intentions. For manipulated trials, the difference in consistency was primarily driven by the inconsistency of non-detectors (44%) compared to detectors (83%). Thus, consistent with the Discrepancy Detection Principle, choice blindness manipulations have stronger ripple effects amongst those who fail to notice the manipulation (Tousignant et al., 1986).

Detection is the key dependent variable in choice blindness studies and is typically measured in two ways. The first measure of detection, called concurrent detection, assesses participants' immediate response to the manipulation. After being shown the manipulated (i.e. non-chosen) face, if a participant were to state that this picture was not the one they had chosen, this participant would classified as currently detecting the manipulation (Johansson et al., 2005). This measure has the benefit that it most clearly indicates a group of participants who detected the change. There is likely to be a low rate of error in incorrectly coding non-detectors as detectors. However, some participants may notice the manipulation immediately after it occurs but be unwilling to report if for a number of reasons and thus this measure likely underestimates the true number of detectors.

The second measure of detection is retrospective detection. Retrospective detection is measured at the end of a study. Participants complete a funneled debriefing procedure in which they are questioned as to whether they noticed anything odd or unusual with the procedure or materials. Participant are eventually fully debriefed and asked to report whether they noticed the manipulation. Participants who report suspicion during this debriefing are classified as retrospective detectors (Johansson, Hall, Gulz, Haake, & Watanabe, 2007). This measure likely overestimates the proportion of true detectors. Participants may report suspicions for reasons unrelated to the manipulation or may simply guess that they were in the manipulated condition rather than truly noticing the manipulation. Because of this, most researchers focus on concurrent detection and use retrospective detection only as an upper-bound estimate of the true number of detectors (Taya, Gupta, Farber, & Mullette-Gillman, 2014).

Failing to detect a change in one's preference for a preferred face is a low consequence mistake and this may partially account for the low rate of detection in these early studies (Johansson et al., 2005). However, failure to detect choice blindness manipulations occurs even with more personally relevant and consequential beliefs such as moral and political attitudes or even one's reported engagement in illegal behaviors (Hall et al., 2013; Sauerland et al., 2013). In one study, participants completed a measure of norm-violating behavior. This scale included not only troublesome behaviors such as cheating on an exam, but also law violating behaviors such as stealing a bike or shoplifting. Participants received manipulated versions of their responses to this scale and elaborated on the reasons for these responses. Roughly 15% of these manipulated items went unnoticed by participants (Sauerland et al., 2013). This higher detection rate does indicate that the consequentiality of a behavior does impact the likelihood of detection. But it is noteworthy that while this concurrent detection rate is much higher than in previous studies, a significant minority of participants still failed to notice this change in a serious and consequential domain.

Although choice blindness originated from decision theory, there are many similarities between choice blindness manipulations and post-event suggestions used in the misinformation literature (Stille et al., 2017). In both paradigms, participants receive suggestive post-event information about either a previous decision or a prior memory. These suggestions are then incorporated into a person's memory and can have downstream consequences for future intentions and behaviors. Similar factors such as detection of the manipulation are proposed to mediate both effects. The main difference is that misinformation uses post-even information from an outside source whereas choice blindness manipulations come ostensibly from the self (Cochran, Greenspan, Bogart, & Loftus, 2016). When choice blindness manipulations are used on a participant’s previous memory report, rather than their reported preferences or attitudes, choice blindness can be viewed as a new kind of misinformation.

In one field study, confederates posing as tourists approached pedestrians in a large city to ask for directions (Sagana, Sauerland, & Merckelbach, 2013). Shortly thereafter, a separate researcher approached these pedestrians and asked them to identify the tourists from a lineup. After a short retention interval, the researcher showed the participant a photograph of the person they selected and asked the participant to explain their choice. However, for some participants the photograph they received was not the person they selected, but rather a random other lineup member (manipulated group). Participants concurrently detected the manipulation 31% of the time with a further 28% retrospectively detecting. Other studies have investigated the consequences of self-sourced misinformation. After witnesses received manipulated feedback about their own lineup identification, they were significantly more likely to change their lineup decision at a second lineup compared to those receiving no feedback (Cochran et al., 2016). Similar to the results of Johansson et al. (2014) this change was primarily driven by non-detectors.

Summary

In summary, the paradigms for studying PIF, misinformation, and choice blindness share many similarities. All three focus on how suggestive questioning or misleading post-event information can impact memory and beliefs in a variety of ways. In this dissertation, I will combine aspects of these three phenomena to study the malleability of eyewitness confidence. Specifically, these studies aim to meet several gaps in the existing literature. Although the PIF effect is robust and meta-analyses have confirmed the strong effect of feedback, these studies are limited in their methodology in that most use the same type of feedback after an initial non-blind lineup (Steblay et al., 2014). The dissertation studies aim to address whether misinformation feedback can lead to the same results as typical feedback after an initial double-blind lineup. Doing so identifies whether double-blind lineups alone provide a sufficient safeguard against confidence inflation.

In addition to this applied question, the dissertation studies contribute to the ongoing literature about the downstream effects of post-event feedback. Compared to typical confirmatory feedback, misinformation feedback is more limited in scope. Feedback that tells the witness they have correctly identified the suspect carries a wide range of implications. Even outside observers would likely assume that witnesses who make a correct identification had a better view of the suspect and had paid more attention at the time of the crime. On the other hand, misinformation feedback is more limited in scope. It only addresses witness confidence and does not carry any implications for the typical battery of posttest measures used in the PIF literature. Thus, if misinformation feedback can similarly affect witness' memory for their viewing experience, this indicates the strong power of misinformation to impact memories and beliefs about events not directly implied by the manipulation.

For the two main studies in the dissertation, participants completed a two-session study. In session one, they watched a mock crime video, made an identification from a target-present lineup, and gave their confidence in their identification. In session two, participants were exposed to suggestive feedback. For participants in the misinformation conditions, this feedback was in the form of self-sourced misinformation. Participants were reminded of their confidence report from session one. However, this report was manipulated such that it implied the participant was either more or less confident in their initial identification than what they actually reported. At the end of the study, all participants reported upon their retrospective memory for their witnessed experience including their retrospective confidence.

Prior to conducting the two main studies, two pilot studies were developed to test whether the typical numeric scale used in PIF studies could be replaced by a prompted verbal scale. These two pilot studies differ from the main studies in several ways. The two main studies are more applied in natural and aim to investigate processes that might occurring during real cases. On the other hand, the two pilot studies answer more basic questions about research design choices. The pilot studies provided essential data as to whether a non-numeric confidence scale could be used as an item to provide self-sourced misinformation about. Validating a new type of scale was especially important as confidence was both measured and manipulated in the main studies. In addition, the studies make an important contribution to research about witness confidence as they add to the emerging literature about how participants evaluate prompted verbal scales. In the proceeding section, the method and results of the two pilot studies are reported. The two main studies in the dissertation investigating misinformation as a form of PIF are reported following these.

Pilot Study A

Method

Overview and Purpose

Although the study of PIF is applied in nature and has resulted in several policy recommendations for law enforcement and the courts, the literature review demonstrates that these studies primary document confidence numerically even though verbal reports are standard in real cases. An exploratory goal of the dissertation was to investigate whether a prompted verbal scale could be used instead of a numeric scale the main studies in order to more closely approximate real-world conditions. Two pilot studies were conducted to explore whether a prompted verbal scale was appropriate for use in the main studies of the dissertation.

The proposed methodology of these main studies would involve giving participants false feedback about their confidence in their lineup choice. This feedback would be presented in the form of a reminder about the witness' earlier confidence statement. This reminder would be manipulated, and the response shown to participants would be two points farther on the scale than their original response. For instance, after making an identification, a participant might select the fourth response option on the scale (labeled *fairly certain*) as their confidence statement. Then, they would later be reminded about their confidence. However, this reminder would not repeat the participant's initial confidence. Instead, the reminder would be for the response option two points higher on the scale (labeled *quite certain*).

Two main research questions were addressed in the pilot studies. Firstly, would participants correctly recognize the difference between response items two points across on the scale. That is, would participants correctly identify that *quite certain* indicated higher confidence than *fairly certain*. If the majority of participants cannot correctly make this assessment, then the proposed feedback would not be an effective way of providing participants with suggestive misinformation. Referring back to the previous example, the purpose of providing the participant with false feedback that they were *quite certain* (rather than *fairly certain*) was to provide suggestive feedback that misinformed the person they were more confident in their identification than they actually said. If the participant believes that *fairly certain* and *quite certain* are synonymous and express the same amount of confidence, then this feedback would not meet its intended purpose.

The second main research question focused on the makeup of the scale. This question aimed to identify whether the scale completely allowed participants to express a full range of confidence judgments and whether the scale was ordered in a manner commonly understood by most people.

Participants

Participants (*N* = 77) were recruited using Amazon Mechanical Turk (MTurk). MTurk is an online marketplace in which individuals (requesters) can post tasks for others (workers) to complete for monetary compensation (Buhrmester, Kwang, & Gosling, 2011). These tasks vary widely, and workers can browse all available tasks and chose which to complete. After completion, payment is deposited in the workers account. Although MTurk participants are not completely representative of the demographics of the United States, workers are, on average, more diverse than other online samples and from college undergraduate samples. Participants motivation for completing tasks varies including earning money, to enjoy interesting task, and to kill time. Compensation rates vary and do not affect data quality (Litman, Robinson, & Rosenzweig, 2015).

The current study used the TurkPrime platform for data collection. TurkPrime is a platform that links with MTurk, and, differently than the main MTurk site, is designed specifically for social science experiments. TurkPrime's software offers several advantages over the main MTurk site in that it allows for exclusion of participants that have already taken a previous study, it allows for longitudinal data collection, and it allows for linking with Qualtrics to set up an autonomous payment structure (Litman, Robinson, & Abberbock, 2017). The current study was posted to MTurk using TurkPrime. Participants were excluded if they had ever participated in a project by the lead researcher in the past.

Participants in the current study were on average 34 years old (*SD* = 11.6) and mostly male (58.4%). The majority identified as White/Caucasian (70.1%) with a minority identifying as Black/African-American (11.7%), Hispanic/Latino (6.5%), Asian-Asian/American (10.4%). Participants were highly educated with 67.6% earning at least a college degree. The study took approximately 10 minutes to complete and participants received $0.50 as compensation.

Materials

The prompted verbal scale used in this study was derived from Windschitl and Wells (1996). This scale is the most commonly used in the eyewitness literature (e.g., Dodson & Dobolyi, 2016; Weber et al., 2008). The scale was modified in several ways from its original form. First, the scale was changed from bipolar to unipolar. During free reports, witnesses tend to explain their confidence in terms of degrees of certainty not degrees of uncertainty (Behrman & Richards, 2005). Additionally, assessing degrees of uncertainty is likely to be more cognitively taxing for participants than assessing degrees of certainty as it is an uncommon meta-cognitive judgment. Because of these factors, the scale used in the current study assessed degrees of confidence on a unipolar scale. In addition to these modifications, the end points of the original scale used by Windschitl and Wells (1996) were also modified to better fit with the overall scale items (see Appendix A).

In discussing the scale in the next sections, when a response option is referred to as “higher” on the scale, it indicates the response shows more confidence. When a response option is referred to as “lower” on the scale it means the response option indicates less confidence.

Procedure

After completing the informed consent, participants were informed that the purpose of the study was to understand how people view eyewitness evidence. Participants then read a brief description of a typical identification procedure. Instructions informed them that a previous study had been conducted in which people made an identification from a lineup and gave their confidence in this identification. Participants were told they would read two of the confidence statements given by these previous participants. Instructions explained that participants should read the two statements and then decide which indicated a person more confident in their identification.

At this point, participants answered 10 questions displayed separately on their own page. For each question, participants read two confidence statements supposedly given by previous participants and selected which showed greater certainty. These statements were shown in the form of “I am ____ certain” with the blank being replaced by a response on the prompted verbal scale. Each question also had a third option to select indicating that the two statements showed equal levels of certainty. For eight of the ten questions, the two statements displayed were two scale points away from each other. The remaining two questions paired the end points of the scale with the neighboring response option to test the new endpoints developed for use in this study. These ten questions were asked in a random order for all participants.

After completing these paired comparison questions, participants then read instructions for a second task. In this task, they translated the scale response options into a numeric percentage. Each question presented participants with one of the ten confidence statements (e.g., “I am ____ certain”) used in the previous task and then asked them to assess how confident that person was on a sliding scale from 0-100%. Each of the 10 certainty statements was displayed on a separate page and in a random order. Participants completed demographic questions at the end of the study.

Hypotheses

This pilot study was intended to be exploratory in nature. Based on its use in past research studying the confidence-accuracy relationship using a prompted verbal scale, it was predicted that participants numeric translations would generally concur with scale ordering (Weber et al., 2008). Although no specific hypotheses were made regarding the extent to which participants would correctly respond to the paired comparison judgments, a general benchmark of 80% was set a priori. This benchmark approximated the percentage of participants that would need to correctly discriminate between the pairs for the prompted verbal scale to be acceptable for use in the main dissertation studies.

Results

Paired Comparison Judgments

For each question in the first task, responses were coded as either correct, incorrect, or equal. Correct responses were those in which the participant selected the response that was higher on the scale (i.e., displayed more confidence) and incorrect responses were those in which the participant selected the response lower on the scale. Equal judgments are those in which the participant responded that the two statements indicated equal confidence.

Overall, participants correctly identified the difference between the two response options most of the time (61.7%). However, a sizable portion of participants believed that these two responses indicated witnesses displaying the same amount of confidence (17.7%). Even more concerning, 16% of the time participants judged that a response lower on the scale indicated the witness was more confident than a response two points higher on the scale.

On average, participants seemed better able to identify a two-point scale difference on the lower half of the scale than the upper half of the scale (see Table A.1). Participants were most likely to discriminate correctly on the pair between *not at all certain* and *somewhat certain* and least likely to discriminate correctly between *very certain* and *almost totally certain.*

Numerical Translation

After responding to the paired comparison judgments, participants also answered questions regarding what percentage they would give each of these confidence statements on a 0-100% scale (see Table A.2). While the paired confidence judgments assess whether participants discriminate between different items on the scale, it does not provide information about how the ordering of the scale matches on to participants' beliefs about how verbal confidence statements should be ordered or how well the scale full reaches the endpoints of numeric judgments.

The numeric translation results indicated great variability in participants' understanding of the confidence statements. The average standard deviation was quite large, particularly on the lower end of the scale. Results further suggested that the scale is not an ideal measurement tool for low confidence respondents. By the second scale point, participants, on average, rated that response as being nearly 50% certain. The scale also struggled to fully cover extremely confident respondents. For most of the scale, ratings averaged between 50%-80% certain with the top response translating as under 90%. Particularly in the middle of the scale, ratings were quite similar with the middle eight responses all falling with 32 percentage points of each other. Focusing on median, rather than mean response, the scale performs better. If participants median responses were rank ordered, the scale would exist in largely the same order as proposed. However, problems with the endpoints persist with the first response receiving a median ranking of 14 and the second response receiving a mean ranking of 51.

Discussion

Overall, the results of this study indicate that participants do not share a common understanding of the proposed prompted verbal scale. Although the numeric translation judgements trend in the correction direction, in that points higher on the scale tend to be rated as more confident than points lower on the scale, the numeric translation of many of the scale points fall quite closely together. Moreover, response options at the endpoints do not fully encompass the ends of a numeric scale suggesting that when using this scale participants with very low confidence will chose an option similar to that of participants with somewhat low confidence and thus important variability will be lost.

While these data do generate concerns, they do not, on their own, answer the question as to whether the scale is appropriate to use in a self-sourced misinformation feedback study. This question is best addressed by the paired comparison data. These data provide the most direct answer as to whether the proposed manipulated of two scale points in either direction would be recognized by participants as feedback that suggested they were more or less confident than their original response. The results from this task clearly demonstrate that participants are not uniformly recognizing a two-point difference in the scale. Participants average correct judgments were well below the 80% benchmark set at the beginning of the study. Thus, if false feedback informed participants that they were two scale points higher in their identification than they originally selected than this would not be seen by many, or in some cases even most, participants as misinformation suggesting more confidence than their original report.

One possible explanation for these results is that the paired comparison judgments removed the scale from its original context. Participants only saw two individual responses. If participants had initially seen the full scale, this may have facilitated more accurate comparison judgments. Individuals tend to be less variable in their assignments of numeric probabilities to verbal statements when they are presented in ascending, rather than random, order (Hamm, 1991). However, in the proposed studies, the feedback would also be presented without the context of the surrounding response options, so it is important to identify whether participants can accurately discriminate in these circumstances.

The scale used here was a modified version of that validated and used in past research (Weber et al., 2008; Windschitl & Wells, 1996). While our results do provide some indication that participants understand the general order of the scale, it suggests that on smaller scale increments, discriminability is lost. However, one reason for this may have been the modifications made from the original scale. The unipolar nature of the scale tested here may have made the task more difficult for participants and harder to discriminate between small increments. Because of this, a second pilot study was conducted in which the original bipolar nature of the scale originally tested by Windschitl and Wells (1996) was retained to investigate whether this lead to improved results.

Pilot Study B

Method

Overview and Purpose

The purpose of this second pilot study was to test whether using a bipolar prompted verbal scale resulted in improved comprehension by participants than the unipolar scale tested in Study A. When using free report, participants tend not to explain their confidence in terms of degrees of uncertainty (Behrman & Richards, 2005). However, perhaps, when using a prompted verbal scale, response options that range from uncertainty to certainty facilitate better understanding of the scale as a whole. To test this, we used the original scale developed by Windschitl and Wells (1996). The only change made was to the scale endpoints to retain the scale's symmetric nature (see Appendix A).

Participants

Participants (*N* = 75) were recruited using TurkPrime with the same procedure as in Study A. Participants were on average 36 years old (*SD* = 11.8) and mostly female (50.7%). The majority of participants identified as White/Caucasian (68.0%) with a minority identifying as Black/African-American (14.7%), Hispanic/Latino (6.7%), Asian/Asian-American (6.7%). Participants were highly educated with 47.9% earning at least a college degree.

Materials and Procedures

This study was conducted in an identical manner as in Study A except using the bipolar prompted verbal scale. Participants completed an informed consent, received study instructions, completed the paired confidence judgments, completed the numeric translation judgments, and finally answered demographic questions.

Results

Paired Comparison Judgments

The results for the paired confidence judgements for Study B were similar to that of Study A. Overall, the majority of participants correctly identified confidence statements higher on the scale as displaying more confidence than confidence statements lower on the scale (Table B.1). Rates of incorrect and same responses were also similar as to Study A.

However, the difference in correct responses between the upper and lower halves of the scale was much more pronounced than in Study A. For *uncertain* response options, less than half of participants correctly discriminated between the paired statements. Correspondingly, rates of incorrect and same judgments also increased. On the other half of the scale, rates of correct responses were at their highest with over 70% of participants correctly differentiating between the confidence statements and only 11.5% of participants responding that the two responses displayed the same confidence.

Evidence for the difference between the *uncertainty* and *certainty* halves of the scale can also be seen by directly comparing individual questions. As the scale is symmetrical, participants evaluated statements containing the same two adverbs both with the *certainty* and *uncertainty* base words. For instance, only 50.7% of participants correctly identified *rather uncertain* as displaying higher certainty that *extremely uncertain* but 84% of participants correctly identified *extremely certain* as displaying higher confidence than *rather certain*. Comparing against the midpoint of *as certain as uncertain*, only 34% of participants correctly recognized *rather uncertain* as being below the midpoint while 78.7% identified *rather certain* as being above the midpoint. This discrepancy between halves of a symmetric scale has been found using other probability judgments as well (Reagan, Mosteller, & Youtz, 1989).

Numeric Translation

The numeric translation judgements in Study B showed some improvement over Study A. Overall, rank ordering participants' median responses matched perfectly with the scale order (see Table B.2). Moreover, the median responses for the scale endpoints (4, 100) revealed participants believed these scale endpoints more closely represented the ends of a numeric scale than in Study A.

However, results again show significant variability in responding. While median responses do match up with scale order, there are only an average of 13.9 percentage points separating each two-point gap across the scale. Again, this difference is primarily driven by the *uncertainty* half of the scale with the differences between a two-point gap larger on the *certainty* than *uncertainty* half of the scale.

Discussion

The bipolar prompted verbal scale in Study B had some advantages and disadvantages over the unipolar scale in Study A. For the *certainty* side of the scale, participants often accurately discriminated between response options and made numeric translations of the scale points consistent with proposed scale order. These results indicated an improvement over Study A. However, responses to the *uncertainty* side of the scale showed the most confusion for participants in both the paired comparison and numeric translation judgments. This finding is consistent with the reasoning proposed in Study A that the meta-judgments required for assessing uncertainty may be significantly more difficult than for judgments of certainty.

Overall, results suggest that prompted verbal scales like those tested here are unsuited for use as an item to be manipulated as self-sourced misinformation. Across both halves of the scales used in these two studies, no more than 70% of participants correctly identified a two-point gap in the scale. This means that even in the best-case scenario, when provided with misinformation that they were two scale points higher than what they originally said, 30% of participants would not view this feedback as indicating more confidence than their original response.

These results highlight several important points for the use of prompted verbal scales in eyewitness confidence research in general. It is important to note, these findings are not meant to indicate prompted verbal scales are unsuitable for use in research as a replacement for a numeric confidence statement after an identification. In fact, the results provide further validation that the order of the adverbs in this prompted verbal scale are consistent with how participants view varying levels of confidence. Removing response options from their surrounding context likely contributed to the low accuracy in the paired comparison judgments. However, these studies do suggest more attention should be paid to the bipolar or unipolar nature of prompted verbal scales. Participants struggle more with responses using a base term of *uncertain* than *certain* and so future research in this area may want to focus on developing and validating a unipolar confidence scale.

However, given the nature of the current studies, in that the goal is to specifically provide misinformation about a participant's prior confidence statement, the results from the pilot studies establish that a prompted verbal scale is unsuited. Because of this, Study 1 and 2 will use a 0-100% sliding numeric scale for participants to report their confidence in their identification. While less ecologically valid, numeric scales have several advantages in the methodology of the main studies. Manipulating a numeric confidence report will result in a manipulation of equivalent size for all participants at all points along the confidence scale. Participants are much more likely to judge 40% certain as being more confident than 20% certain than they are to judge *quite certain* as being more confident than *fairly certain*. The two main studies are the first to test misinformation as a form of PIF. The more straightforward manipulation of the numeric scale allows for the cleanest test when testing this new paradigm.

Study 1

Method

Overview and Purpose

Witness' memory for their confidence in their identification and other aspects of their witnessed experience are malleable and can be influenced by post-event information that informs them that they have correctly identified the suspect from the lineup (Wells & Bradfield, 1998). Study 1 investigates whether the same effects occur when using self-sourced misinformation as feedback rather than a statement about the accuracy of the witness' identification. The study uses a two-session procedure in which participant's confidence is obtained both after an initial double-blind lineup and at a second time point one week later. Study 1 aims to gather a pre-manipulation and post-manipulation measure of confidence to identify whether feedback given after a double-blind lineup can have similar detrimental effects on remembered confidence as feedback given before a first confidence report is gathered. Moreover, this study contributes to the memory literature about the ripple effects of misinformation beyond changing memory for the suggested information. Documenting the downstream consequences of misinformation can help establish that false memories formed by this kind of manipulation differ from false beliefs or are the result of demand characteristics (Smeets et al., 2005).

To that end, participants were recruited for a two-part study. In session one, participants watched a mock crime video, answered questions about their memory for the video (including identifying the suspect), and gave their confidence in these answers. At session two, participants were provided with a reminder of their confidence and asked to elaborate on this judgment. However, the reminder for some participants was 20 points lower or higher than their original rating. Following this, participants completed a posttest questionnaire that probed their memory about aspects of their witnessed experience including their retrospective confidence at the time of the identification.

Participants

Participants were recruited via TurkPrime to take part in a two-session study. Session one of the study took approximately 15 minutes to complete and participants received $0.50 as compensation. Session two of the study took approximately 10 minutes to complete and participants received $0.70 as compensation. Of the 607 participants that completed session one, 523 completed session two resulting in an 86.2% response rate. Seven participants were removed from analysis due to problems with the video, one participant was removed for responding to all free response questions in Spanish, and two participants were removed for withdrawing their consent after debriefing leaving a final sample of 513.

The sample was mostly female (55.3%) with an average age of 37.1 years (*SD* = 12.4). Participants were mostly White/Caucasian (73.3%) with a minority identifying as Black-African-American (7.2%), Hispanic/Latino (5.8), or Asian/Asian-American (9.7%). Participants were highly educated with 61.6% having completed at least a college degree.

Materials

Video. The mock crime video used in the current study was borrowed from Murphy and Greene (2016). In this video, a woman enters a room and searches through several items on a messy desk. During the video, she steals several objects including a laptop. One limitation of the PIF literature is that the vast majority of studies use one of two videos as stimulus material (Steblay et al., 2014). The video for the current study was selected as to expand the exploration of PIF to include new stimulus material. The video used here is newer, displaying a more modern scene and includes sound as well as images. The video was also selected as it showed a good distribution between correct and filler identifications as well as resulting in confidence judgments not near the scale endpoints.

Lineup. The lineup used in the current study consisted of five black and white photographs with the target being in the second position (see Appendix B). The lineup contained unbiased instructions with language similar to that recommended by the National Institute of Justice (Technical Working Group for Eyewitness Evidence, 1999) and participants were given the option to reject the lineup. If participants did not make an identification from the initial lineup, they received a second lineup. This lineup contained the same five photographs in the same order and participants were forced to make an identification (see Appendix B).

The purpose of this study was primarily to identify how self-sourced misinformation feedback would impact participants' memory for aspects of their identification. Because of this, the focus was primarily on lineup choosers. Biased instructions could have been used to force all participants to make an identification and this would have maximized sample size. However, this creates other issues. Nearly all law enforcement agencies report using unbiased instructions (Police Executive Research Forum, 2013). Thus, biased instructions do not match the real-world conditions that this applied studying is aiming to mirror. Moreover, making the lineup forced choice obscures important data. Participants that would have chosen to reject the lineup might conceivably respond in different ways when forced to make a choice such as exhibiting lower confidence or being less likely to attend to feedback. Because of this, it was important to distinguish between those participants that would and would not make an identification if given the option.

The follow-up lineup procedure used here was developed in order to both use unbiased instructions as well as obtain data for the research question of interest for each participant. By giving participants the option to reject the initial lineup and also forcing an identification on the second lineup, each participant contributes data both on whether they would have rejected the lineup if given the option and who they would have chosen if lineup rejection was not an option.

This type of follow-up lineup procedure is novel and has not yet been tested. Because of this, the results from non-choosers (i.e., participants that make an initial lineup rejection) were considered exploratory. For all main analyses, sensitivity tests will be conducted first with the restricted sample of only choosers (i.e., participants that made an identification at the initial lineup) and then with the full sample. In addition to exploring the pattern of results from this kind of secondary lineup procedure, this kind of analysis will test the robustness of the effects of the manipulation.

Procedure

Session one. Participants began the study by completing an informed consent. The informed consent explained that the study involved two sessions over a one-week period and that they should only complete session one if they would also agree to complete session two. The cover story for the study described that the purpose of the experiment was to investigate how people perceived and evaluated others. In order to ensure the video loaded properly, session one could not be completed on a mobile device. After completing the informed consent, participants watched the mock crime video (Murphy & Greene, 2016).

After the video, participants completed filler tasks consistent with the experiment's cover story. These tasks included reading several short stories and responding to questions about the characters and plot of the stories.

Following these filler tasks, instructions informed participants that the true purpose of the study was actually to investigate eyewitness memory and that they would next answer questions about the video similar to that asked of eyewitnesses to real crimes. The task consisted of seven multiple choice questions about events in the video. These questions asked about relevant topics such as the item stolen from the desk in order to disguise the focus of the study on only the identification. After each of these filler multiple-choice questions, participants rated their confidence in their answers on a sliding scale from 0-100%.

Finally, participants attempted to identify the suspect from a five-person, target-present lineup. If participants made an identification from the initial lineup, they then reported their confidence in this answer on a 0-100% sliding scale. If participants rejected the lineup, they were not shown this confidence question. Instead, they immediately responded to the second, follow-up lineup. They then reported their confidence in this answer on a 0-100% scale. Participants did not report their confidence in their non-identification so as to avoid confusion by giving their confidence in two aspects of the identification task. Before exiting the study, participants answered demographic questions and received a reminder about session two.

Session two. One week after completing session one, session two became available to participants. At this time, participants received a reminder email that asked to complete the study that day. They also received a follow-up email that afternoon. For each remaining day for five days, participants who had not yet completed the study received a reminder email in the morning. No participants were allowed to complete the study after this time had elapsed. The average time between completion of the two sessions was 7 days (*SD* = 0.8). Only 7.4% of participants completed session two more than 7 days after session one.

At the beginning of session two, participants read a brief reminder about the mock crime video. Instructions informed them that the purpose of this session was to learn more about their memory of the video. In the first task, participants were asked to elaborate on their answers to the memory test in session one. Participants were reminded of their answer and their confidence in that answer and then wrote why they felt they were this confident. This reminder was a true repetition of the confidence they reported during session one. For instance, if a participant said in session one that they were 80% confident that a laptop was stolen from the office, the prompt read: “In session one, you saw the female thief search through several items on the desk. Before she left, she stole something off the desk and left with it. Last session, we asked you about your memory for the theft. You reported that the item she stole was a laptop and that on a 0-100 scale you were 80% certain in that answer. Can you tell us more about why you felt you were 80% certain that she stole a laptop? Please try to be as detailed as possible in your response.” Participants elaborated on the reasons for their confidence using a free response box. Following this, a second filler elaboration question was asked about the color of the purse on the desk in the video. The purpose of these two elaboration questions was to acclimate participants to the task and build trust prior to the manipulation.

At this point, participants were randomly assigned to one of three conditions: control, confidence increase (CI), and confidence decrease (CD). Modified random assignment was used in which participants were twice as likely to be in one of the manipulated conditions as in the control condition. The purpose of modified random assignment was to maximize power in the manipulated condition where an unknown number of participants would detect the manipulation. In the control condition, after completing the two filler elaboration questions, participants moved on to the remaining tasks in the study.

In the CI and CD conditions, after the initial two filler elaborations, participants were asked to elaborate on their confidence in their identification. The prompt for this question was similar to that of the prior two. The prompt read: “Near the end of session one, you were shown several photographs and asked to identify the female thief. You were then asked to rate your certainty in this choice on a 0-100 scale. You told us you were X% certain in your identification choice. Can you tell us more about why you were X% certain in your answer? Please try to be as detailed as possible in your response."

In the CI condition, the number displayed in the question prompt was 20 points higher than what the participant originally reported in session one. In the CD condition, the number displayed in the prompt was 20 points lower than what the participant originally reported in session one.

Participants in the CI and CD conditions that gave certainty judgments near the end points of the scale represented a unique subset of cases. Participants in the CI condition that gave a confidence judgment of 81 or higher and participants in the CD condition that gave a confidence judgment of 19 or lower could not be manipulated a full 20 points as this would result in a response beyond the range of the scale.

Several solutions were considered for this subset of participants. These participants could have been assigned to one of the two other conditions as has been done in past choice blindness studies (Merckelbach, Jelicic, & Pieters, 2011). However, this solution has several problems here. First, it systematically biases random assignment. Second, there are several reasons why learning about this subgroup in their assigned condition may be particularly interesting. It provides evidence as to whether participants near the end points of the scale are as susceptible to PIF as participants closer to the midpoint of the scale. Because of these reasons, this subset of participants was not moved to the control group. Rather, participants in the CI group that gave a confidence judgment of 81 or more were assigned to receive feedback that they rated their confidence as a 98. For the CD group, participants that originally gave a confidence judgment of

19 or less were assigned to receive feedback that they originally rated their confidence as a 2. Two and 98 were chosen rather than 0 and 100 as these numbers are particularly salient responses and using them may have artificially increased detection. Extant choice blindness research has also encountered this problem in that all participants do not receive the same magnitude of the manipulation. Detection status has not been found to be affected by the magnitude of the manipulation (Hall et al., 2013). Sensitivity analyses were planned to investigate whether including these participants affected the overall pattern of results.

This did create another unique situation for participants in the CD group that rated their confidence above 98 and participants in the CD group that rated their confidence lower than 2. For these participants, the manipulation actually moved them in the opposite of the intended manipulation. However, due to the small degree of change, it is believed that this is unlikely to affect responses. Nonetheless, these participants are excluded from analysis of the main results. This happened to only 0.2% of participants in the manipulated groups.

From this point onwards, all participants completed the same remaining tasks. After the elaboration task, participants answered a posttest questionnaire (see Appendix C). This questionnaire was based off of that used by Smalarz and Wells (2014) and Wells and Bradfield (1998) and was designed to assess the witness' retrospective memory and witnessed experience. Three main categories were assessed: qualities of the witnessed experience, qualities of the identification task, and summative qualities of the witnessed experience (Wells & Bradfield, 1998).

Finally, participants completed a funneled debriefing procedure with the goal of assessing whether they detected the manipulation (Appendix D). For participants in the CI and CD conditions, debriefing began with vague questions that probed whether the participant felt anything odd occurred during the study. The questions became increasingly specific, then asking if there was anything strange specifically about the elaboration task. Then the full manipulation was explained, and participants reported whether they believed they were in the manipulated condition or not. If participants reported believing they were in the manipulated condition, then they were asked to identify whether the manipulation increased or decreased their original confidence judgments. This funneled debriefing process mirrored that used in past research about choice blindness (Johansson et al., 2005).

Participants in the control condition did not answer the identification elaboration question and so the funneled debriefing questions would not apply to them. However, data from control participants about these types of questions can be informative as it provides an estimate as to how many participants would false alarm as detectors even to a non-manipulated item. Because of this, participants in the control condition received a modified funneled debriefing. This funneled debriefing asked the same questions in the same order as for the CI and CD conditions. However, the questions about the manipulation implied that the confidence manipulation had occurred for the question about the item stolen from the office.

Finally, all participants were fully debriefed about the true purpose of the study. At this point, participants were given the option to withdraw consent and have their data deleted.

Measures

Concurrent detection. Concurrent detection was assessed by two independent research assistants. Coders, blind to condition, read participants’ responses when they elaborated on their confidence in their identification. During pilot testing, problems with coding emerged in that participants would refer to a specific number for their confidence and it was unclear whether they were repeating the number displayed in the prompt or indicating detection by asserting their original confidence. Because of this, coders were also given information about the displayed confidence number for each participant to provide additional context. This number was not the participant's initial confidence, but rather the modified confidence displayed in the prompt. This allowed for coders to remain blind while distinguishing between participants confirming the manipulation or disputing the number displayed.

Retrospective detection. Retrospective detection was coded based on participants' responses to the funneled debriefing. Participants were coded as detectors if, after being informed of the study's manipulation, they believed they were in the manipulated condition (see Appendix D for a full description of retrospective coding).

Confidence. Participants gave their initial measure about certainty in their identification at session one on a sliding scale from 0-100. At session two, during the posttest questionnaire, they again reported their certainty in their identification. The question at session two was retrospective in nature and asked participants to report how certain they were when they made their identification during session one ("At the time you answered the memory questions in session one, how certain were you that the person you identified from the photo lineup was the person you saw in the video?"). For analyses, a difference score was calculated that subtracted participants' certainty at session one from their certainty at time two. The difference score thus standardizes change over time regardless of what the participant's initial confidence was.

Hypotheses

Detection. Past research has shown that rates of concurrent detection vary widely from study to study. Higher detection rates tend to be found in studies that manipulate a categorical, rather than a continuous, variable, in studies that have a longer time between the initial report and the manipulation, and in studies that manipulate more serious, or self-relevant variables

(Cochran et al., 2016; Hall, Johansson, & Strandberg, 2012; Sauerland et al., 2013). As the manipulation used in the current studies had a long gap between the initial confidence report and the manipulation and manipulated a continuous outcome, it was hypothesized that a small minority of participants would concurrently detect the manipulation. As retrospective detection tends to be a more liberal measure, it was hypothesized that more participants would be coded as retrospective detectors than as concurrent detectors.

Confidence change. The main purpose of Study 1 was to assess whether presenting participants with a manipulated version of their own confidence ratings would cause later changes in their retrospective memory for their confidence in their identification. Given the research on the effect of typical confirming feedback (Steblay et al., 2014), it was hypothesized that participants in the CI condition would show significant confidence inflation from session one to session two relative to those in the control condition. Similarly, compared to those in the control condition, it was hypothesized that those in the CD condition would show confidence deflation over time. The effect of the CD manipulation was predicted to be smaller than the effect of the CI manipulation based on meta-analyses demonstrating that confirming feedback has a greater impact that disconfirming feedback (Steblay et al., 2014). The effect of the manipulation on confidence was predicted to be driven primarily by non-detectors. Past research has shown that it is participants that fail to detect the manipulation that are most susceptible to the effects of self-sourced misinformation (Cochran et al., 2016).

Posttest questionnaire. Consistent with research on the effects of typical feedback and on research about the consequences of choice blindness manipulations, it was predicted that presenting participants with self-sourced misinformation about their confidence would have ripple effects on their memory for their witnessed experience (Cochran et al., 2016; Steblay et al., 2014). Specifically, it was predicted that participants in the CI condition would report having a significantly better witnessed experience than participants in the control condition and that participants in the CD condition would report having a significant worse witnessed experience than participants in the control condition. Of the thirteen items on the posttest measure, eleven of them have been found to be affected by PIF (Steblay et al., 2014). The other two items (distance and time) are not typically impacted by PIF. These items were used to demonstrate divergent validity between the newly proposed misinformation feedback and typical PIF.

Results

Detection

Manipulation. One hundred and one participants were assigned to the control condition, 204 to the CI condition, and 208 to the control condition. Overall, 12.7% of participants in the CI condition and 19.2% of participants in the CD condition did not receive the full manipulation. That is, they were within 20 points of the scale endpoint in the direction of the manipulation they were assigned to. The larger group of participants that did not receive the full manipulation in the CD group compared to the CI group is mainly driven by the low confidence of participants that initially rejected the lineup. For the CI condition, the feedback provided a confidence number an average of 8.3 points away from the participant's initial response and for the CD condition the feedback was an average of 7.2 points away from the participant's initial response.

Concurrent detection. After removing outliers[1], the average participant wrote 21.9 words (*SD* = 18.7) when elaborating on their identification confidence and spent an average of 67.4 seconds on the page (*SD* = 46.9; see Table 1.1 for example responses). No significant differences emerged between the CI and CD conditions in the amount of words written during

[1] Outliers were defined as responses more than three standard deviations from the mean.

elaboration, $t(401) = 0.18$, $p = .856$, $d = .017$, or on time spent on the page, $t(378) = 0.75$, $p = .455$, $d = .076$.

Of the 412 total participants in the CI and CD conditions, seven were coded as concurrent detectors[2] (see Table 1.2 for example responses). Four of these participants were in the CI group and three in the CD group. Their initial confidence ranged from 4-81. Two participants initially made a non-identification and four initially made a correct identification. Thus, this small sample of concurrent detectors does not seem to differ in important ways from the overall sample.

Retrospective detection. Consistent with our hypothesis, more participants retrospectively detected the manipulation than concurrently detected with 27.0% of participants in the CI condition and 37.0% of participants in the CD condition coded as retrospective detectors. For comparison, 16.8% of participants in the control condition believed their response to a control question was manipulated.

A 2 (condition: CI, CD) x 2 (retrospective detection status: detector, non-detector) chi-square test revealed significant differences in detection by condition, χ^2 (1, $N = 412$) = 4.79, $p = .029$, $\varphi = .108$, with retrospective detection more commonly occurring in the CD condition. To investigate if choosers and non-choosers differed in rates of detection, a 2 (condition: CI, CD) x 2 (identification: chooser, non-chooser) chi-square test was conducted. Rates of detection did not differ between choosers and non-choosers, χ^2 (1, $N = 411$) = 0.87, $p = .352$, $\varphi = .046$. Rates of detection also did not differ between those making a correct and filler identification, χ^2 (1, $N = 411$) = 0.13, $p = .908$, $\varphi = .006$.

Lineup

[2] Of the seven concurrent detectors, five were identified by both coders. For the remaining two, only one coder identified the response as a concurrent detector. These disagreements were resolved via discussion between the two coders after the initial coding was completed. Reliability analyses were not calculated given the very small number of detectors.

On the initial lineup, 48.9% of participants identified the target, 25.6% identified a filler, and 25.5% made a non-identification. Filler choices were roughly evenly distributed across the four fillers (7.8%, 6.2%, 4.3%, 7.2%). Of participants who initially made a non-identification, 51.1% went on to identify the target on the second lineup with filler identifications mostly evenly distributed (9.9%, 12.2%, 15.3%, 11.5%).

Initial Confidence

Participants displayed a moderate level of confidence in their initial identification (*M* = 48.3, *SD* = 27.1; see Figure 1.1). Initial confidence varied based on lineup rejection status, *t*(267.76) = 8.43, *p* < .001, *d* = 0.83. Participants that made an initial identification were significantly more confident (*M* = 53.4, *SD* = 26.7) than participants that initially rejected the lineup and gave their confidence in their identification from the second lineup (*M* = 33.1, *SD* = 22.3). To parse potential differences in confidence between those who initially made an identification, an independent sample t-test was conducted amongst just initial choosers. Results revealed that participants that made a target identification reported higher confidence (*M* = 55.8, *SD* = 26.4) than participants that made a filler identification (*M* = 48.9, *SD* = 26.5), *t*(264.24) = 2.40, *p* = .017, *d* = .26.

Confidence Change

To provide the most direct test of the hypothesis, only choosers from lineup one that received the full manipulation (hereafter called the restricted sample) were included in this first set of analyses. Following this, sensitivity analyses are reported to investigate whether these effects replicate with the full sample.

To initially explore confidence change over time, three paired sample t-tests were conducted comparing session one confidence to session two confidence separately for each condition. As predicted for the CI condition, participants reported higher confidence at session two (*M* = 53.9, *SD* = 26.1) than at session one (*M* = 44.2, *SD* = 20.6), *t*(177) = 6.68, *p* < .001, *d* = 0.50. Also, as predicted, participants in the CD condition reported lower confidence at session two (*M* = 42.4, *SD* = 29.0) than at session one (*M* = 55.7, *SD* = 42.4), *t*(167) = 8.9, *p* < .001, *d* = 0.69. Participants receiving no feedback did not show confidence change from session one (*M* = 53.4, *SD* = 27.3) to session two (*M* = 56.4, *SD* = 24.0), *t*(74) = 1.18, *p* = .241, *d* = 0.14.

To assess change over time between groups, a difference score was calculated for each participant that subtracted their initial confidence at session one from their retrospective confidence at session two. Higher numbers indicated participants' confidence increased at session two relative to session one. A one-way ANOVA was conducted with condition (control, CI, CD) serving as the independent variable and the difference score serving as the dependent variable. The ANOVA showed a significant main effect for condition, *F*(2, 418) = 59.93, *p* < .001, η_p^2 = 0.89. Post hoc Bonferroni comparisons revealed significant differences between the control and CD conditions, *p* < .001 and between the control and CI conditions, *p* = .046 (see Figure 1.2).

It was predicted that the effect of the manipulation on confidence change would be greater for those in the CI condition than those in the CD condition. To test this, the difference scores for the CD groups were reversed in sign. Then an independent samples t-test was conducted with condition (CI, CD) serving as the independent variable and difference score serving as the dependent variable. Results revealed a marginally significant difference in the non-predicted direction, *t*(344) = 1.79, *p* = .074, *d* = .20. Results trended such that participants in the CD group (M_D = 13.4, SD_D = 19.4) were more affected by the manipulation than those in the CI group (M_D = 9.7, SD_D = 19.3).

Past research has typically found that the effect of typical confirming feedback is greater for target identifications than for filler identifications as those making a filler identification may have a weaker initial memory to begin with (Bradfield et al., 2002). To test whether this result was replicated in the current study, a 3 (condition: CI, CD, control) x 2 (identification outcome: target, filler) ANOVA was conducted. The two-way interaction between condition and identification outcome was non-significant, $F(2, 414) = 1.98$, $p = .140$, $\eta_p^2 = .009$. Condition, $F(2, 414) = 55.91$, $p < .001$, $\eta_p^2 = .213$, and identification outcome, $F(1, 414) = 7.67$, $p = .006$, $\eta_p^2 = .018$, both emerged as significant predictors. Collapsed across conditions, participants making a target identification showed a slight confidence decrease ($M_D = -2.4$, $SD_D = 22.6$) over time whereas participants making a filler identification showed a slight confidence increase ($M_D = 2.2$, $SD_D = 21.9$) over time. However, the effect of the manipulation on confidence change did not differ between correct and filler identifications.

During visual inspection of the descriptive information about the interaction, an unexpected trend emerged. Participants in the control condition that made a target identification showed little confidence change whereas participants that made a filler identification showed distinct confidence inflation over time (see Figure 1.3). An exploratory t-test was conducted with those in the control condition to test whether confidence differed significantly over time. Results revealed a significant difference between the confidence change of correct and filler identifications for control participants, $t(73) = 2.38$, $p = .020$, $d = .64$. This indicates that even when participants that make a filler identification receive no feedback, their confidence inflates naturally over time. This finding was called natural confidence inflation.

Sensitivity analyses. For the initial analyses, only choosers who received the full manipulation were included. This provides the most direct test of the hypotheses which aims to investigate confidence change for identifications. However, to test for the robustness of the results, analyses were conducted again using the full sample[3]. Unless specified below, all analyses conducted with the full sample resulted in the same conclusions as those using the restricted sample.

When using the full sample, the paired t-test for confidence at session one and session two for the control condition did reveal significant differences. Participants in the control condition rated their confidence as higher at session two (*M* = 53.2, *SD* = 25.0) than at session one (*M* = 47.7, *SD* = 28.4), *t*(100) = 2.38, *p* = .019, *d* = .24. This finding of natural confidence inflation is consistent with findings about natural confidence inflation for filler identifications in the restricted sample. This confidence inflation for the control condition in the full sample had some ripple effects on further analyses. Although the results of the earlier one-way ANOVA testing the effects of condition on confidence change remained the same, the pairwise comparison between the control and CI condition became non-significant, *p* = .805. This was due to the significant confidence inflation in the control group rather than a change in the effect of the manipulation on the CI group between the full and restricted sample. This was further shown in the significant two-way interaction between confidence and identification outcome, *F*(2, 495) = 3.02, *p* = .050, η_p^2 = .012. Simple main effects revealed the source of this interaction was the difference in confidence change between target identifications in the control condition and filler identifications in the control condition.

Overall, the results of the sensitivity analysis demonstrate consistent effects of the manipulation with both the full and restricted sample. In fact, the only significant differences

[3] Participants rating their confidence that rated their confidence as a 98-100 in the CI condition or as a 0-2 in the CD conditions were not included in the analyses for the full sample. While other participants did not receive the full 20-point manipulation, this subgroup received either no manipulation or a manipulation in the opposite direction as intended.

found in the sensitivity analysis were driven by changes in the control group. With the full sample, the control group showed significant natural confidence inflation even when collapsed across identification outcome. Again, results regarding the specific occurrence of natural confidence inflation for fillers was replicated.

Confidence change and retrospective detection. It was hypothesized that the effect of condition on confidence change would be primarily driven by non-detectors. To test this question, a 2 (retrospective detection status: detector, non-detector) x 2 (condition: CI, CD) ANOVA with the restricted sample was conducted. This analysis revealed a significant main effect of condition, $F(1, 342) = 91.97$, $p < .001$, $\eta_p^2 = .212$. However, this was qualified by a significant condition by detection status interaction, $F(1, 342) = 6.12$, $p = .014$, $\eta_p^2 = .018$. Simple main effects indicated this interaction was driven by significant differences in confidence change in the CD group between detectors and non-detectors, $p =.005$ (see Figure 1.4). A sensitivity analysis conducted with the full sample revealed no change in the pattern or significance of these results.

Posttest Questions

At the end of session two, participants answered 13 questions about their witnessed experience. It was hypothesized that participants in the CI condition would report a better witnessed experience than participants in the control condition and that participants in the CD condition would report a worse witnessed experience compared to participants in the control condition. As confidence was previously analyzed, the remaining 12 items were the focus of this section.

Initially, each of the questions for the restricted sample were analyzed separately in a one-way ANOVA with condition serving as the independent variable. For two of the variables (view and ability to make out features of the thief's face), results revealed significant differences amongst conditions in participants' retrospective memory of their witnessed experience (see Table 1.3). For the full sample, 10 of the 12 variables revealed significant differences amongst conditions (see Table 1.4). The two questions that revealed non-significant results in the full sample are the same two items that past meta-analysis has revealed are not affected by typical post-identification feedback (Steblay et al., 2014). This provides good divergent validity that the current manipulation similarly effects participant's responses as typical feedback. Pairwise comparison revealed the condition differences were primarily driven by the differences in the CD and control group.

A composite score was created for the 10 questions predicted to be affected by feedback (Cronbach's α = .859) with higher numbers indicating the participant had a better witnessed experience. A one-way ANOVA with the restricted sample with condition as the independent variable and the composite posttest score as the dependent variable revealed a significant main effect for condition, $F(2, 418) = 3.10$, $p = .046$, $\eta_p^2 = .015$. Post hoc Bonferroni pairwise comparisons showed that the control ($M = 5.7$, $SD = 1.4$) condition did not differ significantly from the CI condition ($M = 5.3$, $SD = 1.6$), $p = .428$. However, consistent with our hypothesis, participants in the CD condition ($M = 5.1$, $SD = 1.7$) did report a significantly worse witnessed experience than those in the control condition, $p = .042$. These results replicated with the full sample.

Subscales. Two subscales were created from the posttest questionnaire: qualities of the witnessed event (α = .789) and summative judgments (α = .827). A composite scale for qualities of the identification task was not created due to low reliability (α = .631).

Results of subscale analyses for the restricted sample revealed that the manipulation significantly impacted participants memories for qualities of the witnessed event. $F(2, 418) = 3.78$, $p = .009$, $\eta_p^2 = .022$. Bonferroni post-hoc comparisons replicated that significant differences occurred between the control and CD group, $p = .012$. A parallel analysis was conducted for the summative judgment subscale revealing marginally significant differences in the predicted direction, $F(2, 418) = 2.59$, $p = .076$, $\eta_p^2 = .012$, with the CD group reporting having lower scores on the summative judgment subscale than the control group, $p = .071$ (see Figure 1.5). Results replicated with the full sample[4].

Posttest questionnaire and retrospective detection. To investigate whether the effect of condition on differences in the posttest measure differed based on detection status, a 2 (condition: CI, CD) x 2 (retrospective detection status: detector, non-detector) was conducted. Results revealed a significant interaction between condition and detection status, $F(1, 342) = 5.88$, $p = .016$, $\eta_p^2 = .017$. As shown in Figure 1.6, this difference was primarily driven by the difference between detectors and non-detectors in the CD condition. Non-detectors in the CD condition reported having a significantly worse witnessed experience that detectors in the CD condition. Results replicated with the full sample.

Discussion

The results of Study 1 demonstrate that self-sourced misinformation about witnesses' prior confidence statements can influence their memory for their retrospective confidence and have ripple effects for other aspects of their witnessed experience. Consistent with our hypothesis, participants that received misinformation that their confidence was higher than originally reported later remembered having more confidence in their identification. Similarly,

[4] Results for the summative judgment subscale become significant with the full sample.

participants that received suggestion that their confidence was lower than originally reported later remembered having less confidence in their identification and having a worse witnessed experience.

Specifically for those in the CD condition, detection played a significant role in these results. Non-detectors who failed to notice the manipulation were more strongly influenced by the manipulation as seen both in their confidence change scores and in their reports on the posttest questionnaire. It is unclear why detection status played in role in the CD and not the CI condition. One reason for this could be the use of retrospective, rather than concurrent detection. Choice blindness studies investigating the downstream consequence of the manipulation nearly always assess blindness concurrently or in a combined measure of detection that includes both concurrent and retrospective detection (e.g. Johansson et al., 2014). This is the first study to demonstrate ripple effects of the manipulation using retrospective detection. Concurrent detection could not be used in this analysis as too few participants immediately reported noticing the manipulation. This was likely caused by the long retention interval the original confidence report in session one and the manipulation in session two. The one-week retention interval used in the current study is the longest in the choice blindness literature.

Based on past research on the effect of typical confirming and disconfirming feedback, it was predicted that the CI manipulation would have a stronger impact on confidence change than the CD manipulation. However, results did not support this hypothesis. There were no differences in overall confidence change between the CI and CD groups. Interestingly, there were some indications that it was actually the CD group in the current study that had a more powerful impact on participants' memories. The CD condition, and not the CI condition, differed significantly from the control condition on the posttest questionnaire. While past research has not investigated whether the ripple effects of choice blindness manipulations are affected by the direction of the manipulation, some studies have shown that detection is not affected by direction of the manipulation on continuous scales (Sauerland, Schell-Leugers, & Sagana, 2015).

One reason that typical confirming feedback has such a powerful influence is because it activates individuals' needs to be component and accurate in their decision making (Charman et al., 2010; Wells & Quinlivan, 2009). Typical confirming feedback is self-relevant and ego inflating; it tells the participant they got the right answer, and this can play into people's strong desire to view themselves in a positive light. Conversely, typical disconfirming feedback is less relevant to people's need to maintain a positive self-image and so may be less salient. Misinformation feedback does not activate these self-presentation goals in the same way and this may help explain why CI and CD feedback do not follow the same pattern as typical confirming and disconfirming feedback. In fact, the high rate of non-detectors suggest that many participants do not even notice the feedback.

Another potential explanation of these pattern of results can come from the SCIF. This framework proposes that witnesses with weak internal memory cues search for external cues to develop their confidence assessment (Charman et al., 2010). One of the claims of this second stage is that people differentially seek out and accept information that supports their preexisting beliefs and subject disconfirming information to more scrutiny than confirming information. However, as seen by the low rates of detection, misinformation feedback is not easily explicitly recognized and so this biased search for information may be less affected by misinformation relative to typical feedback.

One unexpected result that emerged from Study 1 regarded the control condition. In the full sample, participants that received no feedback showed significant confidence inflation over time. For both the full and restricted sample, this effect occurred for those who made a filler identification. During a follow-up test one week after the initial identification, participants who made a filler identification but received no feedback remembered being significantly more confident in their identification at session two than they were at session one.

Most studies examining PIF, and most eyewitness studies more broadly, only collect confidence from witnesses at a single time point. No other PIF studies examined in the literature review gathered a confidence statement from participants at two separate time points. Instead, the effect of feedback is compared to the effect of no feedback at the same time. In a one-time point study design, natural confidence inflation cannot be detected. This new finding has important implications for the discussion of confidence malleability. The focus of this conversation has mainly centered around reducing the suggestion and feedback witnesses are exposed to. However, these results suggest that suggestion and feedback are only one aspect of confidence malleability. Even under pristine circumstances, confidence inflation seems to occur naturally over time. This supports the continual-malleability hypothesis which states that retrospective judgments like confidence are continually subject to change over time and might never fully set at a fixed value (Quinlivan et al., 2010).

In summary, the main hypotheses of Study 1 were supported. While the rate of concurrent detectors was lower than predicted, the manipulation did affect witness' remembered confidence and aspects of their witnessed experience. Study 1 highlighted some important discrepancies in the pattern of results typically found in PIF studies and that found here. Specifically, whether confirming and disconfirming feedback have similar effects. This spurs the question of whether the differences results are caused by the different kinds of feedback or the different methodology used here. In this study, participants gave their confidence at two time points. No past PIF study has done this. So, the results about misinformation feedback could differ from that of typical feedback not because of the new kind of feedback, but because the participant gave a confidence report prior to the manipulation.

Study 2

Method

Overview and Purpose

The goal of Study 2 is to directly compare the effects of typical and misinformation feedback in a two-session study in which participants report their confidence both pre- and post-manipulation. Not only does this directly compare the size and direction of the effects of the two types of feedback, it also allows for the exploration of new research questions regarding typical feedback. Specifically, it tests whether typical feedback given at a delay after an initial double-blind lineup results in confidence change or whether the initial double-blind lineup serves as a protective factor. A secondary goal of Study 2 was to replicate the new finding of natural confidence inflation found in Study 1.

To that end, participants were recruited to take part in a two-session study. The procedure for Study 2 was similar to that of Study 1. In addition to the three conditions used in Study 1, two new conditions were added typical confirming feedback and typical disconfirming feedback.

Participants

Participants for Study 2 were recruited in the same manner as in Study 1. In session one, 863 participants completed the survey. Of these, 714 also completed session two leading to an 82.7% response rate. Six participants were removed for analysis for withdrawing their consent after the final debriefing and two were removed for responding to all free response questions in Spanish. This resulted in a final sample of 706.

Participants were on average 37 years old (*SD* = 12.6, 53.5% female). Most participants identified as White/Caucasian (71.0%) with a minority identifying as Black/African-American

(12.0%), Asian/Asian-American (7.8%), or Hispanic/Latino (4.1%). Participants were highly educated with 66.5% completing at least a college degree.

Procedure

Session one. The procedure for session one of Study 2 was identical to that of session one of Study 1. Participants completed an informed consent, watched the video, completed filler tasks, and then finished with the memory test including making an identification from the same five-person target-present lineup.

Session two. At the beginning of session two, all participants again completed an informed consent. Following this, they were reminded about the video from session one and asked to elaborate on their confidence in their answers to two of the memory test questions from that session. All participants completed the two filler memory elaboration questions. These questions reminded participants of their confidence in their answer to both the question about the item stolen from the office and the color of the purse on the desk and asked participants to elaborate on why they were this confident.

At this point, participants were randomly assigned to one of five conditions (control, CI, CD, typical confirming feedback (TCF), typical disconfirming feedback (TDF)). Modified random assignment was not used in Study 2 as one of the goals of the study was to explore natural confidence inflation in the control condition. In the control condition, participants saw no further information after the two filler elaboration questions and went on to complete the posttest questionnaire. The CI and CD conditions in Study 2 mirrored that of Study 1. For participants in these conditions (together referred to as the misinformation conditions), after answering the two filler memory elaboration questions, they were reminded of their confidence in their identification at session one and asked to elaborate on why they were this confident. For participants in the CI condition, the reminder in the question prompt indicated a confidence judgment that was 20 points higher than what the participant actually reported in session one. For participants in the CD condition, the reminder was 20 points lower than what participants originally said in session one. Participants near the end points of the scale that could not be manipulated a full 20 points were handled in the same manner as Study 1.

In the two new typical feedback conditions, participants were not reminded of their confidence in their initial identification and did not elaborate on their identification confidence. Instead, after completing the two filler memory elaboration questions, they received further instructions that read: "Near the end of session one, you were shown several photographs and asked to identify the female thief. You selected one of the women in the lineup and then rated how certain you were that she was the thief on a 0-100 scale. Many participants are interested in whether the person they picked was the actual thief from the video." For participants in the TCF condition, these instructions were followed by the statement "Good job, you correctly identified the actual thief from the video." For participants in the TDF condition, the statement instead read "Sorry, you did not correctly identify the actual thief from the video." The survey required participants to stay on this page for a short period before advancing to ensure they did not simply click through the feedback.

Following this, participants completed the same posttest questionnaire as in Study 1 (see Appendix C). Participants in the CI and CD conditions then completed the same funneled debriefing as participants in the CI and CD conditions in Study 1 (see Appendix D). Participants in the TCF, TDF, and control conditions completed the same funneled debriefing as control participants in Study 1.

Measures

Detection. Concurrent detection was assessed by two independent research assistants, blind to condition. Participants that wrote in their elaboration that they felt there was something strange with the number presented to them or gave other indications of recognizing the manipulation were coded as concurrent detectors. Retrospective detection was assessed by participants' responses to the funneled debriefing at the end of session two. After being informed about the design of the study, participants that indicated they believed they were in the misinformation condition were coded as retrospective detectors.

Certainty. Participants gave their initial measure of certainty after the identification in session one on a sliding scale from 0-100. At session two, during the posttest questionnaire, they again reported their certainty in their identification. For analyses, a difference score was calculated that subtracted participants' certainty at session one from their certainty at session two to assess change over time. Higher numbers indicate confidence inflation over time.

Posttest questions. The posttest questionnaire contained three main judgment categories: qualities of the witnessed event, qualities of the identification task, and summative judgements (see Appendix C).

Hypotheses

Detection. Given the low detection rate in Study 1, it was predicted that a very small number of participants in Study 2 would concurrently detect the manipulation. Retrospective detection was predicted to be more common with a small minority of participants hypothesized to retrospectively detect the manipulation.

Confidence change. Overall, it was hypothesized that the four feedback groups would differ significantly from the control group in their confidence change from session one to session two. It was hypothesized that the CI and TCF conditions would show significant confidence inflation over time relative to the control group. It was predicted that participants in the CD and TDF conditions would show significant confidence deflation over time relative to the control group. The effect of misinformation feedback on confidence was expected to be stronger for non-detectors than for detectors.

One of the main novel research question this study was aiming to explore focuses on the difference between misinformation and typical feedback. It was predicted that the effect of TCF feedback would be greater than that of TDF feedback, but the effect of CI feedback would not be greater than CD feedback.

One unexpected finding from Study 1 was the natural confidence inflation shown by participants in the control group. In this study, it was hypothesized that control participants would again show natural confidence inflation and that this effect would occur primarily for filler identifications.

Posttest Questions

The effect of condition on participants' responses to the posttest questions was predicted to occur in the same pattern as their responses to the confidence question. That is, participants in the CI and TCF conditions were predicted to report having a better witnessed experience at posttest relative to the control condition and participants in the CD and TDF conditions were predicted to report having a worse witnessed experience relative to the control condition.

Results

Detection

Manipulation. In total, 144 participants were assigned to the control condition, 136 to the CD condition, 145 to the CI condition, 138 to the TDF condition, and 143 to the TCF condition. Of the 281 participants in the misinformation conditions, 9.6% of participants in the

CI condition and 20% of participants in the CD condition did not get the full manipulation. The larger group of participants that did not receive the full manipulation in the CD group compared to the CI group is mainly driven by the low confidence of participants that initially rejected the lineup. Participants in the CI group that did not receive the full manipulation received feedback that was an average of 4.6 points away from their original answer and for participants in the CD condition feedback was an average of 6.6 points away from their original response.

Concurrent detection. Seven participants in the manipulated conditions were coded as concurrently detecting the manipulation,[5] two in the CI group and five in the CD group (see Table 2.1). All seven of the concurrent detectors correctly identified the target from the initial lineup. Their initial confidence ranged from 20-100.

Participants in the misinformation conditions wrote an average of 28.0 words (*SD* = 18.6) when elaborating on their confidence statement and spent an average of 72.6 seconds on the page (*SD* = 52.8)[6]. Two independent sample t-tests were conducted to determine whether time to page submit or words written during elaboration differed based on condition. Time to page submit, *t*(273) = 1.54, *p* = .126, *d* = .186, and word count, *t*(248.62) = 0.23, *p* = .815, *d* = .027, did not differ between the CI and CD conditions.

Retrospective detection. In the CD condition, 37.5% of participants retrospectively detected the manipulation compared to 35.9% of participants in the CI group. To investigate whether detection differed based on the direction of the manipulation, a 2 (condition: CI, CD) x 2 (retrospective detection status: detector, non-detector) chi-square test was conducted. Results revealed no significant differences in detection status between conditions, χ^2 (1, *N* = 281) = 0.81,

[5] Similar to Study 2, five participants were identified by both coders as concurrent detectors while the other two were identified by only one coder. Disagreements were resolved by discussion between coders. Reliability statistics were not calculated given the low sample size of detectors.

[6] Outliers (defined as more than three standard deviations beyond the mean) were removed.

p = .776, φ = .017. For comparison, 27.1% of participants in the control condition, 27.5% of participants in the TDF condition, and 21.7% of participants in the TCF condition thought their confidence had been manipulated in the prompt of one of the filler elaboration questions.

To evaluate whether retrospective detection differed based on lineup rejection status, a 2 (retrospective detection status: detector, non-detector) x 2 (lineup rejection status: chooser, non-choosers) chi-square test was conducted. Participants that initially rejected the lineup were significantly less likely to detect the manipulation (22.4%) compared to participants that initially made a choice from the lineup (41.1%), χ^2 (1, N = 281) = 7.71, p = .005, φ = .166. Amongst choosers, no differences in detection occurred between target and filler identifications, χ^2 (1, N = 214) = 0.14, p = .699, φ = .026.

Lineup

On the initial lineup, 50.4% of participants identified the target, 28.1% of participants identified a filler, and 21.5% of participants made a non-identification. For participants who initially made a non-identification, 50.7% identified the target at the follow-up lineup.

Initial Confidence

Participants reported a moderate amount of confidence in their lineup identification at session one (M = 52.8, SD = 27.6; see Figure 2.1). Participants that initially made a correct identification were on average 59.1% confident, participants that made an initial filler identification were 53.8% confident, and participants that made a non-identification were 36.8% confident in their lineup selection. Initial confidence differed based on identification decision, $F(2, 707) = 38.65$, $p < .001$, $\eta_p^2 = .099$. Pairwise Bonferroni comparison indicated that those making a non-identification had significantly lower confidence than those making a correct identification, $p < .001$, and those making a filler identification, $p < .001$. The difference in confidence between correct and filler identifications was marginally significant, *p* = .070. For participants who initially made a non-identification, those who went on to correctly identify the target in the second lineup had a higher confidence in their identification (41.7%) compared to those that picked a filler in the second lineup (31.7%), *t*(150) = 2.23, *p* = .027, *d* = .361.

Confidence Change

Analysis plan. Participants that made a non-identification from the initial lineup were excluded from all further analyses. In Study 1, these participants were included in sensitivity analyses to test whether including them affected the overall pattern of results. However, in Study 2 they were excluded as the follow-up lineup had a different connotation for some conditions here. In the typical feedback conditions, the effect of information about identification accuracy has a different meaning based on whether the lineup was initially rejected. A participant that receives feedback that they incorrectly identified the suspect from the first lineup would likely find that feedback more compelling and salient than a participant that was told they made an incorrect identification after initially rejecting the first lineup. That participant already had low confidence and was forced to make an identification from the second lineup and so the feedback would not be as surprising or meaningful for them. Similarly, the ego enhancing aspect of TCF feedback might be stronger for participants that made an initial identification than those who did not. Because of these confounds, and that the study was intended to focus on identifications and not non-identifications, participants who made a non-identification were excluded from confidence change and posttest analyses.

Participants that did not receive the full 20-point manipulation in the misinformation conditions were handled in a similar manner as in Study 1. They were removed from initial analyses and then sensitivity tests were conducted to explore whether their inclusion in the sample changed the overall pattern of results[7]. In order to be consistent across conditions, participants in the TCF and TDF conditions that would have been excluded if they were in the misinformation conditions were also removed from initial analyses. For example, a participant that rated their confidence as 90% in the TCF condition will also be excluded as, if they were assigned to the CI condition, they would not have been able to receive the full manipulation. This ensures that the typical feedback conditions contain the same subsample as the misinformation conditions. When sensitivity analyses are conducted, these participants will also be added back in. When participants near the endpoints are added back to the sample, this group will be referred to as the full sample.

Confidence change by condition. Initially, a series of paired sample t-tests were conducted to explore whether confidence changed significantly from session one to session two within each condition. Replicating natural confidence inflation, participants in the control condition reported significantly more confidence at session two than they had at session one, $t(114) = 3.005$, $p = .003$, $d = 0.285$. Results for the misinformation condition were also in the predicted direction such that participants in the CD condition reported less confidence at session two relative to session one, $t(96) = 8.60$, $p < .001$, $d = 0.876$, and participants in the CI condition reported more confidence at session two relative to session one, $t(85) = 9.28$, $p < .001$, $d = 1.00$. Results for the TCF group were also in the predicted direction with participants reporting greater confidence at session two than session one, $t(91) = 8.80$, $p < .001$, $d = 0.919$. However, contrary to our hypotheses, the TDF manipulation did not significantly impact participants' confidence over time, $t(101) = 1.02$, $p = .309$, $d = 0.102$.

[7] Participants that rated their confidence as a 98-100 in the CI condition or as a 0-2 in the CD conditions were removed from all analyses and not included in the sensitivity analyses. While other participants did not receive the full 20-point manipulation, this subgroup received either no manipulation or a manipulation in the opposite direction as intended. As in Study 1, this happened to only a very small number of participants (*N* = 14).

Next, a difference score was calculated that subtracted participants session one confidence from their session two confidence. Higher numbers indicated confidence inflation with participants reporting greater confidence at session two than they had at session one. For the initial test of the main hypothesis, a one-way ANOVA was conducted with condition serving as the independent variable and the difference score serving as the dependent variable. Results showed a significant main effect of condition, $F(4, 487) = 46.05$, $p < .001$, $\eta_p^2 = .274$.

Planned pairwise comparisons that compared the four feedback conditions to the control condition were then performed (see Figure 2.2). It was predicted that the four feedback conditions would differ significantly from the control condition. Results supported this hypothesis with the CD, $p < .001$, and TDF condition, $p = .004$, differing significantly from the control condition in the predicted direction. The CI, $p = .003$, and the TCF, $p < .001$, condition also differed significantly from the control condition in the predicted direction.

It was predicted that confirming feedback would have a stronger impact on confidence change than disconfirming feedback, but only for typical, and not misinformation, conditions. To test this prediction, independent sample t-tests were conducted between the CI and CD conditions and between the TCF and TDF conditions. The signs of the CD and TDF conditions were reversed for this analysis. For the misinformation feedback conditions, there were no significant differences in the extent of confidence change between the CI and CD group, $t(181) = 0.05$, $p = .956$, $d = .006$. On the other hand, the effect of typical feedback on confidence change did differ between confirming and disconfirming group with confirming feedback having a stronger effect, $t(192) = 5.57$, $p < .001$, $d = .801$.

To test for the interaction between identification accuracy and condition, a 2 (identification outcome: target, filler) x 5 (condition: CI, CD, control, TDF, TCF) ANOVA was conducted revealing no significant interaction, $F(4, 482) = 1.08$, $p = .36$, $\eta_p^2 = .009$. The effect of feedback on confidence change did not differ based on whether the participant made a correct identification. The control group was of particular interest in this study due to the natural confidence inflation found in filler identifications in Study 2. An independent sample t-test of the control condition tested whether correct and filler identifications differed in their confidence change. No significant differences were found in confidence change between correct and incorrect identifications, $t(113) = 0.52$, $p = .607$, $d = .105$. Thus, while the natural confidence inflation found in Study 1 replicated here, the effects were not different between target and filler identifications which differed from Study 1.

Sensitivity analysis. The pattern of results described here did not change when participants near the endpoints of the scale were included back into the sample.

Confidence change and retrospective detection. It was hypothesized that the effect of condition on confidence change in the misinformation conditions would be primarily driven by non-detectors. A 2 (retrospective detection status: detector, non-detector) x 2 (condition: CI, CD) ANOVA demonstrated this hypothesis was not supported. The analysis revealed only a significant main effect of condition, $F(1, 179) = 142.98$, $p < .001$, $\eta_p^2 = .444$. The effect of the manipulation on change scores did not differ between detectors and non-detectors (see Figure 2.3). Results replicated with the full sample.

Posttest Questionnaire

Twelve one-way ANOVAs were conducted to investigate the effect of condition on each of the posttest questions in the restricted sample (see Table 2.2). For nine of the 12 questions, condition had a significant effect on participant’s self-reported witness experience (see Figure 2.4). To assess the overall effect of condition on participants’ retrospective memory, a composite score was created of the ten posttest questions typically used in the PIF literature ($\alpha = .888$). A one-way ANOVA was conducted with the composite posttest measure serving as the dependent variable and condition serving as the independent variable. Results revealed a significant main effect of condition, $F(4, 487) = 6.64$, $p < .001$, $\eta_p^2 = .050$ (see Figure 2.5). It was predicted that the four feedback groups would differ significantly from the control group. Planned pairwise comparisons revealed that, consistent with our hypothesis, participants in the CD condition, $p = .006$, and the TDF condition, $p = .018$, reported having a significantly worse witnessed experience than those in the control condition. Inconsistent with our hypothesis, participants in the TCF condition did not report having a better witnessed experience than those in the control condition, $p = .142$. Also inconsistent with our hypothesis, participants in the CI condition reported having a significantly worse witnessed experience than those in the control condition, $p = .014$.

Subgroup analyses were conducted to investigate whether condition affected different aspects of the witness' retrospective memory (see Figure 2.5). The subscale for qualities of the witnessed experience ($\alpha = .800$) differed significantly based on condition, $F(4, 487) = 4.09$, $p = .003$, $\eta_p^2 = .033$. Planned pairwise comparisons between the control and four feedback conditions revealed no significant differences between the control condition and either the CI, $p = .274$, or TCF, $p = .102$, condition. Participants in both the CD, $p = .073$, and TDF, $p = .048$ conditions reported having a worse witnessed experience than those in the control condition. This mostly matches the pattern of results for the overall composite measure.

The subscale for qualities of the identification did not meet sufficient reliability and so was not analyzed ($\alpha = .522$). The subscale for summative judgments ($\alpha = .854$) revealed significant differences amongst conditions, $F(4, 487) = 5.12$, $p < .001$, $\eta_p^2 = .040$. Planned pairwise comparisons between the control and four feedback conditions revealed significant differences between the control and CD condition, p = .009, and a marginally significant difference in the opposite of the predicted direction between the control and CI condition, p = .078.

Overall, the analysis demonstrate support for only some of the hypotheses. The effect of the manipulation on the CD and TDF conditions were confirmed. However, no significant differences occurred for the TCF condition and the results for the CI condition often occurred in the opposite of the predicted direction.

Sensitivity analysis. For the sensitivity analysis, participants near the endpoints of the scale were re-included. The pattern of results for the ANOVA on the effect of condition on the overall composite measure remained the same, $F(4, 539) = 11,02$, $p < .001$, $\eta_p^2 = .076$. The results of the pairwise comparisons with the full sample more fully supported the hypothesis. The CD and TDF conditions had a significantly lower score on the composite measure compared to the control condition. Participants in the TCF condition had a significantly higher score on the composite measure compared to controls. The CI and control conditions did not significantly differ from each other. This same pattern occurred for the subscales for qualities of the witnessed event and summative judgments. For both subscales the CD, TCF, and TCD conditions differed significantly from control in the predicted direction and the CI condition did not differ significantly from the control.

Thus, the results of the sensitivity and main analysis both similarly support the hypothesis that participants in the CD and TDF conditions would report having a worse witnessed experience compared to that of the control condition. The main difference in the results between the full and restricted sample are in the CI condition. In the restricted sample, the CI condition did not report having a better witnessed experience than the control condition. In fact, results revealed that participants in the CI condition reported having a worse witnessed experience. This was not true for the full sample in which the CI and control conditions did not differ. For the TCF condition, no significant differences were found relative to controls for the restricted sample, but in the full sample the TCF condition showed inflated confidence relative to controls.

The difference between the full and restricted samples for the CI and TCF conditions was that, for the restricted sample, participants at the highest level of confidence (81-100%) were excluded. Thus, it is possible that the effect of confirming feedback on participants' memory for their witnessed experience is greater for participants that were initially high in confidence and weaker for participants initially lower in confidence. To test this explanation, the ANOVAs for the full sample were run separately for participants above and below 50% on initial confidence. Results supported this post-hoc explanation (see Figure 2.6). Only for participants above 50% confidence on the initial lineup did the CI condition differ significantly in the predicted direction for both the qualities of the witnessed experience and the summative judgment subscale.

Discussion

The results of Study 2 mostly replicated the results of Study 1. Consistent with the first study, nearly no participants concurrently detected the manipulation. This provides further indication that choice blindness manipulations that presented manipulated version of participants' own memory reports relatively far in time from the initial reporting are highly unlikely to be immediately noticed and reported by participants. As predicted, retrospective detection occurred more frequently. However, retrospective detection does not provide as precise of a measure as concurrent detection. During the funneled debriefing, the questions become increasingly specific until the entire manipulation is explained. By this point, significant demand characteristics have been created for participants. This can most easily be seen in the meaningful number of participants in the non-misinformation conditions that believe the confidence displayed to them in one of the filler elaboration questions had been manipulated. This variability may help explain why retrospective detection did not significantly interact with condition in Study 2 when it did in Study 1.

For the results regarding the effect of misinformation feedback on confidence change, Study 2 did provide a replication of Study 1. Participants in the CI condition reported inflated confidence and participants in the CD condition reported deflated confidence relative to controls. As predicted, the CI and CD condition did not differ in overall extent of confidence change. This provides additional evidence that although misinformation and typical feedback lead to similar effects, they are somewhat distinct phenomenon. Misinformation feedback does not seem to activate the self-presentation aspect of typical feedback. These results are also consistent with the second stage of the Selective Cue Integration framework (Charman et al., 2010).

Independently of the misinformation feedback conditions, the typical feedback conditions provide important information about the extent to which feedback impacts memory after an initial double-blind lineup. Results demonstrate that typical feedback, particularly when confirming in nature, has a significant, large effect on witness memory even after pristine lineup procedures are used. Choosers' average confidence in the TCF condition inflated from 47.7% from the initial lineup to 70.4% after confirming feedback.

These findings make a novel contribution to the PIF literature as they are the first to assess confidence across two time points. Although many theorized that confidence is an unstable judgment that can change over time, this is the first study to demonstrate how typical PIF impacts memory after a lineup that uses best practices procedures has been conducted.

The comparison between the misinformation and typical feedback conditions revealed the differential effect of disconfirming feedback. While typical disconfirming feedback did not significantly impact confidence change over time, misinformation disconfirming (i.e., CD condition) feedback did. One reason for the differences between these conditions is the elaboration task. Participants in the misinformation feedback conditions spent time thinking about and explaining their reasons for their manipulated confidence. This elaboration and reflection on the misinformation may have facilitated the memory change results. Participants in the typical feedback condition did not engage in elaboration as this is not part of the typical PIF paradigm. Past research has shown that elaborating on misinformation items increases false memory for these items (Drivdahl & Zaragoza, 2001).

The purpose of this study was to test whether the procedure typically used in misinformation choice blindness studies results in similar impacts on participants' memory for their witnessed experience compared to the procedure studied in typical PIF research. It was for this reason that participants in the typical feedback conditions did not complete an elaboration task. Future research is needed to tease apart whether the differential effects are caused by exposure to misinformation, elaboration on the misinformation, or a combination of the two.

The unexpected finding regarding natural confidence inflation in Study 1 was replicated in Study 2. Participants that received no feedback showed greater confidence one week after their initial identification than they had at the time of the lineup. This phenomenon has not yet been identified in the literature to date. Studies about confidence malleability generally document confidence at only one time point. Studies using this methodology inherently cannot study natural confidence inflation.

One past study that has gathered a confidence statement from witnesses at two time points was Cochran et al. (2016). To test whether natural confidence inflation occurs in different contexts, the data from Cochran et al. (2016) was re-analyzed. In this study, participants watched a mock crime slideshow and made an identification from a target absent lineup with biased instructions. Participants then gave their confidence on a scale from 1-11 with each scale point labelled from 0% to 100%. After completing filler tasks, some participants were exposed to a choice blindness manipulation. Following this, a second lineup and confidence judgment were collected from all participants later in the study. A paired sample t-test revealed that participants in the control condition showed higher confidence at lineup two (M = 6.0, SD = 2.7) than at lineup one (M = 5.6, SD = 2.5), t(125) = 3.14, p = .002, d = .280. This provides a replication of natural confidence inflation using different lineup materials, using a different confidence scale, and using a different sample of college undergraduates.

The finding regarding natural confidence inflation has several important implications. It suggests that even when pristine procedures are used, confidence will inflate over time without any outside influences. This is an important point as a no feedback condition typically serves as the control to compare against the effects of feedback. If natural confidence inflation is occurring, then the effects of suggestion are not being compared against no confidence change, but rather compared against significant confidence increases that happen naturally without suggestion. In regard to policy, this reinforces the recommendation to not only make sure the lineup is double-blind but also stresses the importance of only considering the initial confidence statement as all others are subject to not only suggestion but also the effects of simply elapsed time. The study of how jurors consider multiple confidence statements is scarce and has resulted in contradictory results (Jones et al., 2008; Paiva et al., 2011). These findings highlight the need for future research in this area as it demonstrates that, even with pristine conditions, confidence change over time will occur. Thus, it is critical to understand how actors in the legal system view and make decisions about variable confidence.

Results for the posttest measure tell a more complicated story than in Study 1. For disconfirming feedback, the hypotheses were mostly supported with participants in the CD and TDF conditions reporting having a worse witnessed experience than controls. However, for this analysis in particular, results changed when including the full sample. This led to the hypothesis that the reason for the change in results in the sensitivity analysis was that the ripple effects of the confidence manipulation would be strongest for confirming feedback for those who were initially high in confidence. This highlights a potential future avenue of PIF research focusing on not just the effects of feedback but whether specific subgroups are more susceptible to feedback than others. Work on this topic has typically focused on the differential effect of feedback for correct and incorrect identifications. However, like the research on confidence, subgroup analyses can help elucidate the full picture about the overall effects of feedback.

General Discussion

Two studies investigated the effect of misinformation choice blindness feedback as a type of PIF. In Study 1, participants were given a misleading reminder about their confidence in their lineup identification. Only a very small number of participants detected the change between the number presented to them in session two and the number they chose during session one. The manipulation had substantial effects on participants' later memory. Participants that received misinformation that they were more confident than they really were in their lineup choice later remembered greater confidence relative to participants that received no feedback. This feedback also had ripple effects on their memory for other aspects of their witnessed experience. Unlike typical feedback, disconfirming misinformation feedback had effects on a similar magnitude as confirming feedback. In Study 2, the effect of misinformation feedback in the current paradigm was compared with typical feedback. Both misinformation and typical feedback affected how confident witnesses remembered being in their initial identification and had ripple effects on their memory for other aspects of the events.

In addition to extending the literature on PIF to a new time point and a new kind of feedback, these studies contribute to the literature on the downstream consequences of misinformation. In the misinformation conditions, only a slight manipulation was used. Participants received a single reminder that their confidence was 20 points higher or lower than what they originally said. This small manipulation had distinct consequences for their later memory and beliefs. It not only affected their remembered confidence, but also impacted their memory for memory acquisition judgements and global assessments of memory and identifications in general. This powerful indicates that the confidence change results are not just the result of demand characteristics or changed memory for just the information implied by the misinformation. Rather, the small manipulation had a global effect on a variety of related measures.

Limitations and Future Directions. There are several important limitations to consider in the current studies. Both studies were online experiments using MTurk participants. Online research has been found to replicate the results of in-lab studies in general and in cognitive psychology and false memory studies in particular (Zwaan et al., 2017). While MTurk participants do represent a more diverse sample than college undergraduates, they are still non-representative in meaningful ways. MTurk samples have been critiqued for involving participants that may attend less to study materials or be non-naïve about important aspects of a study. As workers on the platform can communicate each other using messaging boards and complete similar tasks in other experiments, there are concerns about the non-naiveté of participants. However, survey research suggests that only 13% of workers read blogs and only half of those reported ever reading about the contents of a social science experiment (Chandler, Mueller, & Paolacci, 2014). Moreover, the issue of non-naïve participants is particularly problematic for studies that use very common stimuli such as the trolley problem. In video in the current study has only been used in a small number of other studies and has not been used with U.S. participants. This reduces the chances that these participants had ever been exposed to the stimuli before. In addition, TurkPrime was used to prevent participants from enrolling in the study if they had ever taken a study by the lead researcher in the past. Many workers on MTurk follow requesters who post interesting research (Chandler et al., 2014). Thus, it was important to try to prevent participants who had previously been exposed and debriefed about misinformation studies from enrolling in the current studies.

Another limitation of these experiments was the use of a computer to provide the manipulation rather than another person. Feedback may be more effective when provided by a person rather than a computer, particularly if there is rapport between the participant and the feedback-giver as may be the case in police cases. However, past research has shown that choice blindness experiments can be effective both online and in-person (Cochran et al., 2016; Hall, Johansson, Tärning, Sikström, & Deutgen, 2010). The findings of typical PIF studies have also been replicated using feedback given in an online study (Key et al., 2017). On the other hand, the online context may have made participants feel more comfortable in reporting detection as it removes the uncomfortable nature of having to report an error to a person involved in the research. To date, no choice blindness studies have manipulated whether the feedback is provided by a person or a computer. Such a study would help explain whether feedback source differentially affects concurrent or retrospective detection.

In the two main experiments in the dissertation, confidence was gathered numerically on a 0-100% scale. Before pilot testing, the goal of the studies was to gather confidence using a confidence scale closer to the verbal free report used by police. However, the two pilot studies demonstrated that a prompted verbal scale was not suitable for use in providing misinformation feedback. Past choice blindness research has manipulated responses to a Likert-type scale (Merckelbach et al., 2011). However, this scale only had five response points and so a manipulation of two response options represented a more significant overall change than the two-point manipulation of the 11-point scale tested in the pilot studies. To address this, the manipulation could have provided feedback that was three or four points away from the participants’ original response. But, this would have compounded the issue of participants being too close to the scale endpoints to receive the full manipulation. While not as ecologically valid, pilot results showed the numeric scale used here was the best option. It allowed for a direct test of the new feedback that was understood by all participants.

Conclusion.

In the current studies, the paradigms for studying PIF, misinformation, and choice blindness were merged. Using a choice blindness paradigm, participants were provided with a modified version of their original memory report one week after an initial double-blind lineup. This new method of providing feedback revealed that the effects found with typical feedback replicate at a new time point and with a subtler kind of feedback. While results of the new misinformation and typical feedbacks were largely similar, disconfirming misinformation feedback had a stronger effect on witness memory than confirming feedback. This indicates that misinformation feedback may operate differently than typical feedback, perhaps because participants often fail to detect it. Overall, the results support recommendations to only rely on witness' initial confidence statement. Confidence statements gathered at any later point are affected by feedback from officers, suggestive questions, and simply the effects of time.

References

Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior*, *29*(3), 279–301. http://doi.org/10.1007/s10979-005-3617-y

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (6th ed., pp. 1–62). New York: Academic Press.

Berkowitz, S. R., Laney, C., Morris, E. K., Garry, M., & Loftus, E. F. (2008). Pluto behaving badly: False beliefs and their consequences. *The American Journal of Psychology*, *121*(4), 643. http://doi.org/10.2307/20445490

Bernfstein, D. M., Laney, C., Morris, E. K., & Loftus, E. F. (2005). False memories about food can lead to food avoidance. *Social Cognition*, *23*(1), 11–34. http://doi.org/10.1521/soco.23.1.11.59195

Borchard, E. M. (1961). *Convicting the innocent*. Рипол Классик.

Bradfield, A. L., Wells, G. L., & Olson, E. A. (2002). The damaging effect of confirming feedback on the relation between eyewitness certainty and identification accuracy. *Journal of Applied Psychology*, *87*(1), 112–120. http://doi.org/10.1037/0021-9010.87.1.112

Bradfield, A., & McQuiston, D. E. (2004). When does evidence of eyewitness confidence inflation affect judgments in a criminal trial? *Law and Human Behavior*, *28*(4), 369–387. http://doi.org/10.1023/B:LAHU.0000039331.54147.ff

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*(3), 390–404. http://doi.org/10.1016/0749-5978(88)90036-2

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. http://doi.org/10.1177/1745691610393980

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–30. http://doi.org/10.3758/s13428-013-0365-7

Charman, S. D., Carlucci, M., Vallano, J., & Gregory, A. H. (2010). The selective cue integration framework: a theory of postidentification witness confidence assessment. *Journal of Experimental Psychology. Applied*, *16*(2), 204–218. http://doi.org/10.1037/a0019495

Cochran, K. J., Greenspan, R. L., Bogart, D. F., & Loftus, E. F. (2016). Memory blindness: Altered memory reports lead to distortion in eyewitness memory. *Memory & Cognition*, *44*(5), 717–726. http://doi.org/10.3758/s13421-016-0594-y

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1990). Juror sensitivity to eyewitness identification evidence. *Law and Human Behavior*, *14*(2), 185–191. http://doi.org/10.1007/BF01062972

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, *4*(4), 243–260. http://doi.org/10.1007/BF01040617

Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting eyewitness expressions of confidence: The featural justification effect. *Law and Human Behavior*, *39*(3), 266–280. http://doi.org/10.1037/lhb0000120

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology*, *30*(1), 113–125. http://doi.org/10.1002/acp.3178

Douglass, A. B., & Jones, E. E. (2013). Confidence inflation in eyewitnesses: Seeing is not believing. *Legal and Criminological Psychology*, *18*(1), 152–167. http://doi.org/10.1111/j.2044-8333.2011.02031.x

Drivdahl, S. B., & Zaragoza, M. S. (2001). The role of perceptual elaboration and individual differences in the creation of false memories for suggested events. *Applied Cognitive Psychology*, *15*(3), 265–281. http://doi.org/10.1002/acp.701

Dysart, J. E., Lawson, V. Z., & Rainey, A. (2012). Blind lineup administration as a prophylactic against the postidentification feedback effect. *Law and Human Behavior*, *36*(4), 312–319. http://doi.org/10.1037/h0093921

Garry, M., Manning, C. G., Loftus, E. F., & Sherman, S. J. (1996). Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin & Review*, *3*(2), 208–14. http://doi.org/10.3758/BF03212420

Geraerts, E., Bernstein, D. M., Merckelbach, H., Linders, C., Raymaekers, L., & Loftus, E. F. (2008). Lasting false beliefs and their behavioral consequences. *Psychological Science*, *19*(8), 749–753. http://doi.org/10.2139/ssrn.1270110

Goldstein, A. G., Chance, J., & Schneller, G. (1989). Frequency of eyewitness identification in criminal cases: A survey of prosecutors. *Bulletin of the Psychonomic Society*, *27*(I), 71–74. http://doi.org/10.3758/BF03329902

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS One*, *7*(9). http://doi.org/10.1371/journal.pone.0045457

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, *117*(1), 54–61. http://doi.org/10.1016/j.cognition.2010.06.010

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PloS One*, *8*(4). http://doi.org/10.1371/journal.pone.0060554

Hamm, R. M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes, 48*(2), 193–223. http://doi.org/10.1016/0749-5978(91)90012-I

Johansson, P., Hall, L., Gulz, A., Haake, M., & Watanabe, K. (2007). Choice blindness and trust in the virtual world. *Human Interface, 9*(2), 83–86.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–9. http://doi.org/10.1126/science.1111709

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change: You will like this paper better if you (believe you) chose to read it! *Journal of Behavioral Decision Making*, *27*(3), 281–289. http://doi.org/10.1002/bdm.1807

Jones, E. E., Williams, K. D., & Brewer, N. (2008). “I had a confidence epiphany!”: Obstacles to combating post-identification confidence inflation. *Law and Human Behavior, 32*(2), 164–176. http://doi.org/10.1007/s10979-007-9101-0

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1304–1316. http://doi.org/10.1037/0278-7393.22.5.1304

Key, K. N., Wetmore, S. A., Cash, D. K., Neuschatz, J. S., & Gronlund, S. D. (2017). The effect of post-ID feedback on retrospective self-reports in showups. *Journal of Police and Criminal Psychology*. http://doi.org/10.1007/s11896-017-9228-y

Laney, C., & Loftus, E. F. (2017). False memories matter: The repercussions that follow the development of false memory. In R. A. Nash & J. Ost (Eds.), *False and distorted memories*. Abbington, UK: Psychology Press.

Laney, C., Morris, E. K., Bernstein, D. M., Wakefield, B. M., & Loftus, E. F. (2008). Asparagus, a love story: Healthier eating could be just a false memory away. *Experimental Psychology, 55*(5), 291–300. http://doi.org/10.1027/1618-3169.55.5.291

Leippe, M. R., Wells, G. L., & Ostrom, T. M. (1978). Crime seriousness as a determinant of accuracy in eyewitness identification . *Journal of Applied Psychology*, *63*(3), 345–351. http://doi.org/10.1037/0021-9010.63.3.345

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. http://doi.org/10.1177/1529100612451018

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. http://doi.org/10.3758/s13428-016-0727-z

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519–528. http://doi.org/10.3758/s13428-014-0483-x

Loftus, E. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*(4), 361–6. http://doi.org/10.1101/lm.94705

Loftus, E. F., & Greenspan, R. L. (2017). If I'm certain, is it true? Accuracy and confidence in eyewitness memory. *Psychological Science in the Public Interest*, *18*(1), 1–2. http://doi.org/10.1177/1529100617699241

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology. Human Learning and Memory*, *4*(1), 19–31. http://doi.org/10.1037/0278-7393.4.1.19

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 585–589. http://doi.org/10.1016/S0022-5371(74)80011-3

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. http://doi.org/10.1037//0022-3514.37.11.2098

Luus, C. A. L., & Wells, C. L. (1994). The malleability of eyewitness confidence: Co-witness and perseverance effects. *Journal of Applied Psychology*, *79*(October 1994), 714–723. http://doi.org/10.1037/0021-9010.79.5.714

Maclean, C. L., Brimacombe, C. A. E., Allison, M., Dahl, L. C., & Kadlec, H. (2011). Post-identification feedback effects: Investigators and evaluators. *Applied Cognitive Psychology*, *25*(5), 739–752. http://doi.org/10.1002/acp.1745

Merckelbach, H., Jelicic, M., & Pieters, M. (2011). Misinformation increases symptom reporting: A test - retest study. *JRSM Short Reports*, *2*(10), 75.

http://doi.org/10.1258/shorts.2011.011062

Mickes, L., Clark, S. E., & Gronlund, S. D. (2017). Distilling the confidence-accuracy message: A comment on Wixted and Wells (2017). *Psychological Science in the Public Interest*, *18*(1), 6–9. http://doi.org/10.1177/1529100617699240

Murphy, G., & Greene, C. M. (2016). Perceptual load affects eyewitness accuracy and susceptibility to leading questions. *Frontiers in Psychology*, *7*(AUG), 1–10. http://doi.org/10.3389/fpsyg.2016.01322

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: The National Academies Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. http://doi.org/10.1037//1089-2680.2.2.175

Niel v. Biggers, 409 U.S. 188 (1972).

Paiva, M., Berman, G. L., Cutler, B. L., Platania, J., & Weipert, R. (2011). Influence of confidence inflation and explanations for changes in confidence on evaluations of eyewitness identification accuracy. *Legal and Criminological Psychology*, *16*(March), 266–276. http://doi.org/10.1348/135532510X503340

Paterson, H. M., & Kemp, R. I. (2006). Comparing methods of encountering post-event information: The power of co-witness suggestion. *Applied Cognitive Psychology*, *20*(8), 1083–1099. http://doi.org/10.1002/acp.1261

Patihis, L., Frenda, S. J., Leport, A. K. R., Petersen, N., Nichols, R. M., Stark, C. E. L., … Loftus, E. F. (2013). False memories in highly superior autobiographical memory individuals. *Proceedings of the National Academy of Sciences of the United States of America*, (20). http://doi.org/10.1073/pnas.1314373110

Petty, R. E., Haugtvedt, C. P., & Smith, S. M. (1995). Elaboration as a determinant of attitude strength: Creating attitudes that are persistent, resistant, and predictive of behavior. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 93–130). Mahwah, NJ. Retrieved from http://www.psy.ohio-state.edu/petty/documents/1995PettyHaugtvedtSmithElaboration.pdf

Pickrell, J. E., Heima, M., Weinstein, P., Coolidge, T., Coldwell, S. E., Skaret, E., … Milgrom, P. (2007). Using memory restructuring strategy to enhance dental behaviour. *International Journal of Paediatric Dentistry*, *17*(6), 439–448. http://doi.org/10.1111/j.1365-263X.2007.00873.x

Police Executive Research Forum. (2013). A national survey of eyewitness identification procedures in law enforcement agencies, 126. Retrieved from http://www.policeforum.org/assets/docs/Free_Online_Documents/Eyewitness_Identification/a national survey of eyewitness identification procedures in law enforcement agencies 2013.pdf

Quinlivan, D. S., Neuschatz, J. S., Douglass, A. B., Wells, G. L., & Wetmore, S. A. (2011). The effect of post-identificption feedback, delay, and suspicion on accurate eyewitnesses. *Law and Human Behavior*, 1–11. http://doi.org/10.1007/s10979-011-9277-1

Quinlivan, D. S., Neuschatz, J. S., Jimenez, A., Cling, A. D., Douglass, A. B., & Goodsell, C. A. (2009). Do prophylactics prevent inflation? Post-identification feedback and the effectiveness of procedures to protect against confidence-inflation in ear-witnesses. *Law and Human Behavior*, *33*(2), 111–121. http://doi.org/10.1007/s10979-008-9132-1

Quinlivan, D. S., Wells, G. L., & Neuschatz, J. S. (2010). Is manipulative intent necessary to mitigate the eyewitness post-identification feedback effect. *Law and Human Behavior*, *34*(3), 186–197. http://doi.org/10.1007/s10979-009-9179-7

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology*, *74*(3), 433–442. http://doi.org/10.1037//0021-9010.74.3.433

Renooij, S., & Witteman, C. (1999). Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, *22*(3), 169–194. http://doi.org/10.1016/S0888-613X(99)00027-4

Sagana, A., Sauerland, M., & Merckelbach, H. (2013). Witnesses' blindness for their own facial recognition decisions: A field study. *Behavioral Sciences & the Law*, *31*(5), 624–636. http://doi.org/10.1002/bsl.2082

Sagana, A., Sauerland, M., & Merckelbach, H. (2014). 'This is the person you selected': Eyewitnesses' blindness for their own facial recognition decisions. *Applied Cognitive Psychology*, *28*(5), 753–765. http://doi.org/10.1002/acp.3062

Sauerland, M., Schell-Leugers, J. M., & Sagana, A. (2015). Fabrication puts suspects at risk: Blindness to changes in transgression-related statements. *Applied Cognitive Psychology*, *29*(4). http://doi.org/10.1002/acp.3133

Sauerland, M., Schell, J. M., Collaris, J., Reimer, N. K., Schneider, M., & Merckelbach, H. (2013). "Yes, I have sometimes stolen bikes": Blindness for norm-violating behaviors and implications for suspect interrogations. *Behavioral Sciences & the Law*, *31*(2), 239–55. http://doi.org/10.1002/bsl.2063

Scoboria, A., Mazzoni, G., Jarry, J. L., & Bernstein, D. M. (2012). Personalized and not general suggestion produces false autobiographical memories and suggestion-consistent behavior. *Acta Psychologica*, *139*(1), 225–232. http://doi.org/10.1016/j.actpsy.2011.10.008

Semmler, C., Brewer, N., & Wells, G. L. (2004). Effects of postidentification feedback on eyewitness identification and nonidentification confidence. *Journal of Applied Psychology, 89*(2), 334–346. http://doi.org/10.1037/0021-9010.89.2.334

Skagerberg, E. M., & Wright, D. B. (2009). Susceptibility to postidentification feedback is affected by source credibility. *Applied Cognitive Psychology*, *23*(4), 506–523. http://doi.org/10.1002/acp.1470

Smalarz, L. (2015). *Pre-feedback eyewitness statements: Proposed safeguard against feedback effects on evaluations of eyewitness testimony*. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/250422.pdf

Smalarz, L., & Wells, G. L. (2014). Post-identification feedback to eyewitnesses impairs evaluators' abilities to discriminate between accurate and mistaken testimony. *Law and Human Behavior*, *38*(2), 194–202. http://doi.org/10.1037/lhb0000067

Smeets, T., Merckelbach, H., Horselenberg, R., & Jelicic, M. (2005). Trying to recollect past events: Confidence, beliefs, and memories. *Clinical Psychology Review*, *25*(7), 917–934. http://doi.org/10.1016/j.cpr.2005.03.005

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*(3), 315–327. http://doi.org/10.1037/0033-2909.118.3.315

Steblay, N. K., Wells, G. L., & Douglass, A. B. (2014). The eyewitness post identification feedback effect 15 years later: Theoretical and policy implications. *Psychology, Public Policy, and Law*, *20*(1), 1–18. http://doi.org/10.1037/law0000001

Stille, L., Norin, E., & Sikstro, S. (2017). Self-delivered misinformation- Merging the choice blindness and misinformation effect paradigms. *PLOS One*, *12*(3), e0173606.

http://doi.org/10.1371/journal.pone.0173606

Stovall v. Deno, 388 US 29 (1967).

Taya, F., Gupta, S., Farber, I., & Mullette-Gillman, O. A. (2014). Manipulation detection and preference alterations in a choice blindness paradigm. *PloS One*, *9*(9), e108515. http://doi.org/10.1371/journal.pone.0108515

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. Retrieved from http://www.ojp.usdoj.gov.ny

Tousignant, J. P., Hall, D., & Loftus, E. F. (1986). Discrepancy detection and vulnerability to misleading postevent information. *Memory & Cognition*, *14*(4), 329–338. http://doi.org/10.3758/BF03202511

United States v. Wade, 388 U.S. 230, 288 (1967).

Weber, N., Brewer, N., & Margitich, S. (2008). The confidence-accuracy relation in eyewitness identification: Effects of verbal versus numeric confidence scales. In K. H. Kiefer (Ed.), *Applied Psychology Research Trends* (pp. 103–118). Nova Publishers.

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology*, *36*(12), 1546–1557. http://doi.org/10.1037/0022-3514.36.12.1546

Wells, G. L., & Bradfield, A. L. (1998). “Good, you identified the suspect”: Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*(3), 360–376. http://doi.org/10.1037/0021-9010.83.3.360

Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitnesses’ recollections: Can the postidentification-feedback effect be moderated? *Psychological Science*, *10*(2), 138–144. http://doi.org/10.1111/1467-9280.00121

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, *66*(6), 688–696. http://doi.org/10.1037/0021-9010.66.6.688

Wells, G. L., & Murray, D. M. (1983). What can psychology say about the Neil v. Biggers criteria for judging eyewitness accuracy? *Journal of Applied Psychology*, *68*(3), 347–362. http://doi.org/10.1037/0021-9010.68.3.347

Wells, G. L., Olson, E. A., & Charman, S. D. (2003). Distorted retrospective eyewitness reports as functions of feedback and delay. *Journal of Experimental Psychology. Applied*, *9*(1), 42–52. http://doi.org/10.1037/1076-898X.9.1.42

Wells, G. L., & Quinlivan, D. S. (2009). The eyewitness post-identification feedback effect: What is the function of flexible confidence estimates for autobiographical events? *Applied Cognitive Psychology*, *23*(8), 1153–1163. http://doi.org/10.1002/acp.1616

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*(6), 603–647. http://doi.org/10.1023/A:1025750605807

Wesson, C. J., & Pulford, B. D. (2009). Verbal expressions of confidence and doubt. *Psychological Reports*, *105*(1), 151–160. http://doi.org/10.2466/PR0.105.1.151-160

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology*, *2*(4), 343–364. http://doi.org/10.1037/1076-898X.2.4.343

Wise, R. A., Pawlenko, N. B., Safer, M. A., & Meyer, D. (2009). What US prosecutors and defence attorneys know and believe about eyewitness testimony. *Applied Cognitive*

Psychology, *23*(9), 1266–1281. http://doi.org/10.1002/acp.1530

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. http://doi.org/10.1177/1529100616686966

Wright, D. B., & Skagerberg, E. M. (2007). Postidentification feedback affects real eyewitnesses. *Psychological Science, 18*(2), 172–178. http://doi.org/10.1111/j.1467-9280.2007.01868.x

Wylie, L. E., Patihis, L., McCuller, L. L., Davis, D., Brank, E., Loftus, E. F., & Bornstein, B. (2014). Misinformation effect in older versus younger adults: A meta-analysis and review. In M. P. Toglia, D. F. Ross, J. Pozzulo, & E. Pica (Eds.), *The elderly eyewitness in court* (pp. 38–66). UK: Psychology Press. http://doi.org/10.4324/9781315813936

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2017). Participant Nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin and Review,* 1–5. http://doi.org/10.3758/s13423-017-1348-y

Appendix A

Pilot Study A: Confidence Scale

1. Not at all certain
2. Slightly certain
3. Somewhat certain
4. Fairly certain
5. Rather certain
6. Quite certain
7. Very certain
8. Extremely certain
9. Almost totally certain
10. Completely certain

Pilot Study B: Confidence Scale

1. Completely uncertain
2. Extremely uncertain
3. Quite uncertain
4. Rather uncertain
5. Somewhat uncertain
6. As certain as uncertain
7. Somewhat certain
8. Rather certain
9. Quite certain
10. Extremely certain
11. Completely certain

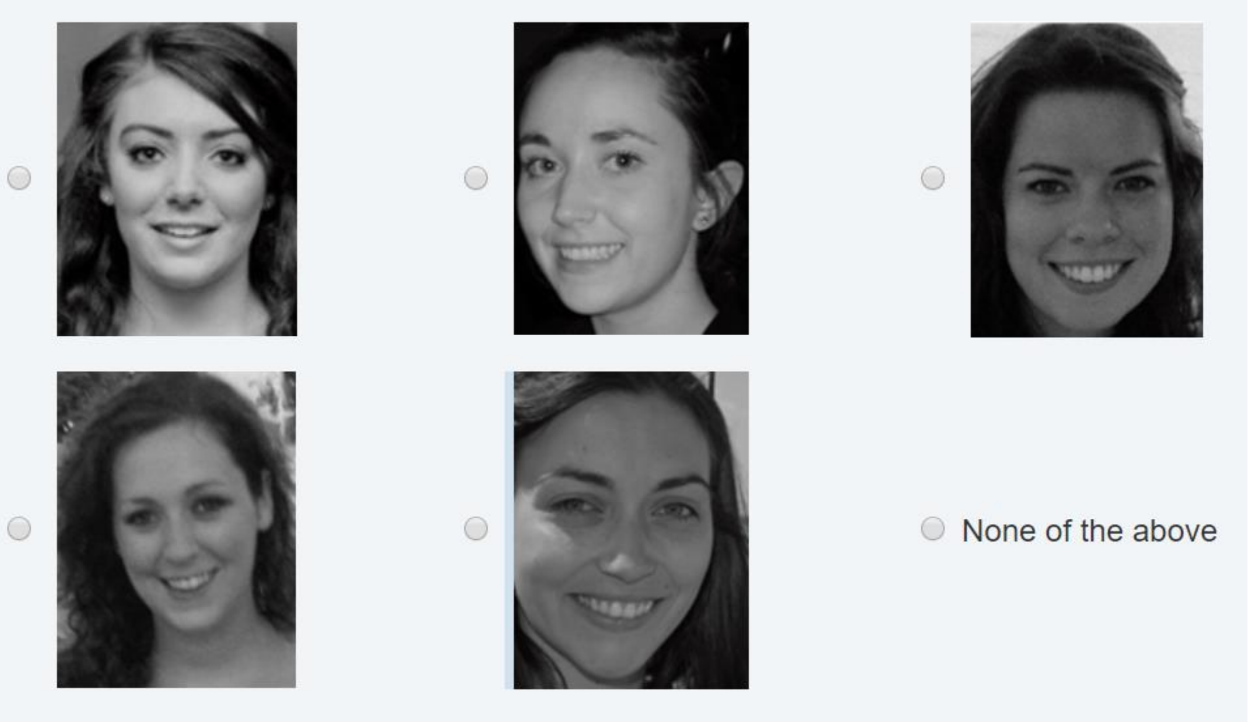
Note: Although witness confidence is generally discussed in the research literature using that nomenclature, c*ertain* was used as the base word of the scales rather than *confident. Certain* is a word more commonly used in daily lexicon than *confident* and so it was chosen as to be both more easily understood by participants and more likely to be the kind of word used by participants during free response. The iWeb corpus was used to confirm the difference in use in daily lexicon between *certain* and *uncertain* (https://corpus.byu.edu/iweb/). This corpus uses a sample of over 14 billion words online. The corpus showed that *certain* occurred with a frequency of 2,513,675 and *confident* occurred with a frequency of 457,120.

Appendix B

Initial Lineup Instructions- Study 1

Below, several photographs are displayed. Your task is to identify which, if any, of these photographs is the thief from the video.

The thief may or may not be included in these photographs. It's just as important to clear innocent persons as to identify the guilty. Keep in mind that the thief may not appear exactly as she did in the video as her hair and clothing may have changed over time.



If participants chose none of the above then they received a second lineup with the instructions: "If you had to choose, which of these photographs is the thief from the video?" The second lineup was identical to the first with the *none of the above* option removed.

Source: Murphy and Greene (2016)

Appendix C

Posttest questionnaire used in Study 1 and Study 2

Instructions: In the next section, we have some final questions for you about the theft. Please read each question carefully and respond based on your memory of the video you saw in session one.

View[1]: How good of a view did you get of the thief?

- 0 (very poor) – 10 (very good)

Face[1]: How well were you able to make out specific features of the thief's face from the video?

- 0 (not at all) – 10 (very well)

Attention[1]: How much attention were you paying to the thief's face while watching the video?

- 0 (none) – 10 (my total attention)

Distance: How much distance would you estimate there was between the camera and the thief?

- 0 (not very much distance) – 10 (a lot of distance)

Time: How much time would you estimate that you saw the thief's face for?

- 0 (not very much time) – 10 (a lot of time)

Basis[3]: To what extent do you feel that you had a good basis (enough information) to identify the thief from the photo lineup?

- 0 (no basis at all) – 10 (a very good basis)

Ease[2]: How easy or difficult was it for you to figure out which person in the photo lineup was the thief from the video?

- 0 (extremely difficult) – 10 (extremely easy)

*IDTime**[2]: How much time do you estimate it took for you to make an identification?

- 0 (I needed almost no time) – 10 (I had to look at all the photos for a long time)

Image[2]: How clear is the image you have in your memory of the thief you saw in the video?

- 0 (not at all clear) – 10 (very clear)

Willing[3]: On the basis of your memory of the thief from the video, how willing would you have been to testify in court that the person you identified was the same person from the video?

- 0 (not at all willing) – 10 (totally willing)

StealConf: At the time you answered the memory questions in session one, how certain were you that you correctly identified the item the thief stole from the office?

- 0 (not at all certain) – 100 (completely certain)

ColorConf: At the time you answered the memory questions in session one, how certain were you that you correctly identified the color of the purse on the desk in the office?

- 0 (not at all certain) – 100 (completely certain)

Certainty: At the time you answered the memory questions in session one, how certain were you that the person you identified from the photo lineup was the person you saw in the video?

- 0 (not at all certain) – 100 (completely certain)

Strangers[3]: Generally, how good is your recognition memory for faces of strangers you have met only once before?

- 0 (very poor) – 10 (excellent)

Trust[3]: Think about another participant in this study that had the same view of the thief in the video as you did. Do you think an identification by this eyewitness ought to be trusted?

- 0 (definitely should not be trusted) – 10 (definitely should be trusted)

Note: * indicates variable was reverse coded. “StealConf” and “colorconf” questions were asked in order to disguise the focus on the identification question. The remaining 13 questions constitute the typical posttest questions used in post-identification feedback studies. [1] indicates subscale of qualities of the witnessed event. [2] indicates subscale of qualities of the identification task. [3] indicates summative judgments subscale.

Appendix D

Funneled debriefing questionnaire used in Study 1 and Study 2

Do you have any questions or comments for us about the study?

- Free response answer

What do you think the study was about?

- Free response answer

Did you find anything odd or unusual about the study?

- Yes
    - If yes, then: What did you find odd or unusual about the study?
        - Free response answer
- No

Think back to the beginning of this session. We reminded you of how certain (on a 0-100 scale) you were in some of your answers to the memory test in session one. Did you feel there was anything strange about this process?

- Yes
    - If yes, then: What did you find strange about the process?
        - Free response answer
- No

At this point, we want to tell you more about the true purpose of the study. Please read the following information carefully.

Our main interest in this study was how confident witnesses are when they are asked questions about their memory of a crime. Earlier in this session, we asked you about some of your answers to the memory questions we asked in session one. We then asked you to explain more about your certainty in these answers.

One of these questions asked you about the person you identified in session one from the photo lineup. You were also reminded about your certainty in that choice. For instance, in session one, you might have said you were 80% certain, and in session two we asked you why you felt you were 80% certain in your choice.

However, for some participants, the number they saw in that question was not what they had originally picked. Rather, it was replaced with a different number. So if you said you were 80% certain in your answer, the question prompt gave you a different number. Do you think that you were one of the participants this happened to?

- Yes
- No
    - If no: Do you have any final questions or comments for us about the study?
        - Free response

- *After this, participants received the full debriefing and were again asked to consent to participate in the study. They did not see any of the remaining questions.*

- I’m not sure
  - If I’m not sure: If you had to guess, do you think that the reminder we gave you in session two had a different number than what you originally said in session one?
    - Yes
    - No
      - If no: Do you have any final questions or comments for us about the study?
        - Free response
          - *After this, participants received the full debriefing and were again asked to consent to participate in the study. They did not see any of the remaining questions.*

You said that you think you were one of the participants who was shown a different number on the certainty scale than what you originally picked. How was the response you were reminded of in session two different from what you said in session one?

- The number shown to me for my confidence in my identification in session two was **higher** than what I originally said in session one.
- The number shown to me for my confidence in my identification in session two was **lower** than what I originally said in session one.
- I’m not sure
  - If I’m not sure: If you had to guess, which of these best describes what you experienced?
    - The number shown to me for my confidence in my identification in session two was **higher** than what I originally said in session one.
    - The number shown to me for my confidence in my identification in session two was **lower** than what I originally said in session one.

When did you realize that the certainty statement we showed you at session two was different than what you originally said in session one?

- Free response answer

How did you realize that the certainty statement we showed you at session two was different than what you originally said in session one?

- Free response answer

Do you have any final questions or comments for us about the study?

- Free response answer

After this, participants received the full debriefing and were again asked to consent to participate in the study. They did not see any of the remaining questions.

Note: Participants were coded as retrospective detectors if they chose “yes” to the question “Do you think that you were one of the participants this happened to?”. Participants that chose “I’m not sure” were not automatically coded as detectors. Only participants that went on to answer “yes” and then correctly assessed which direction their response was manipulated in were coded as detectors. This more conservative coding scheme was implemented in order to avoid problems with past measures of retrospective detection that tend to overestimate the number of true detectors due to guessing (Taya et al., 2014).

Table A.1

Paired Confidence Judgments- Pilot Study A

Confidence Pair	**Correct**	**Incorrect**	**Equal**
1-2	75.3	15.6	9.1
1-3	80.5	9.1	10.4
2-4	61.0	11.7	27.3
3-5	53.2	23.4	23.4
4-6	61.0	20.8	18.2
5-7	72.7	15.6	11.7
6-8	75.3	13.0	11.7
7-9	26.0	53.2	20.8
8-10	39.0	37.7	23.4
9-10	72.7	18.2	9.1
Average (Top 5)	*66.2*	*16.1*	*17.7*
Average (Bottom 5)	*57.1*	*27.5*	*15.3*
Total Average:	*61.7*	*21.8*	*16.5*

Note: Confidence pair numbers correspond to the confidence scale in Appendix A. Numbers in the correct, incorrect, and same column are percentages.

Table A.2

Numeric Translation of Verbal Confidence Scale- Pilot Study A

Scale Point	***M***	***Mdn***	***SD***
Not at all certain	27.1	14.0	31.0
Slightly certain	48.1	51.0	23.4
Somewhat certain	57.6	60.0	20.7
Fairly certain	59.5	63.0	19.5
Rather certain	65.4	69.0	18.4
Quite certain	71.6	75.0	16.1
Very certain	79.7	82.0	15.1
Extremely certain	85.1	92.0	18.4
Almost totally certain	79.9	83.0	16.4
Completely certain	86.4	95.0	18.7
Total Average:	*66.0*	*68.4*	*19.77*

Note: Confidence pair numbers correspond to the confidence scale in Appendix A.

Table B.1

Paired Confidence Judgments- Pilot Study B

Confidence Pair	**Correct**	**Incorrect**	**Same**
1-2	36.0	28.0	36.0
1-3	49.3	26.7	24.0
2-4	50.7	37.3	12.0
3-5			
4-6	34.7	28.0	37.3
5-7	80.0	5.3	14.7
6-8	78.7	13.3	8.0
7-9	76.0	10.7	13.3
8-10	84.0	9.3	6.7
9-11	74.7	18.7	6.7
10-11	41.3	36.0	22.7
Average (Top 5)	*42.7*	*30.0*	*27.3*
Average (Bottom 5)	*70.9*	*17.6*	*11.5*
Total Average:	*60.5*	*21.3*	*18.1*

Note: Confidence pair numbers correspond to the confidence scale in Appendix A. Numbers in the correct, incorrect, and same column are percentages.

Table B.2

Numeric Translation of Verbal Confidence Scale- Pilot Study B

Scale Point	***M***	***Mdn***	***SD***
Completely uncertain	29.5	4.0	41.3
Extremely uncertain	25.0	6.0	34.5
Quite uncertain	32.7	21.0	29.0
Rather uncertain	34.4	27.0	23.7
Somewhat uncertain	33.2	34.0	17.0
As certain as uncertain	45.7	50.0	20.2
Somewhat certain	58.7	59.0	19.3
Rather certain	67.4	74.0	20.4
Quite certain	72.8	80.0	23.3
Extremely certain	87.4	95.0	23.0
Completely certain	88.3	100.0	26.0
Total Average:	*52.3*	*50.0*	*25.2*

Note: Confidence pair numbers correspond to the confidence scale in Appendix A.

Table 1.1

Example Confidence Elaborations- Study 1

Response	Initial Confidence	Displayed Confidence
I was 70% certain in my answer because I had just seen the female thief and felt that photo looked most like her. However, I am not great at identifying faces and tend to think many people look alike.	50	70
I said I was 30 percent sure because wasn't entirely certain that I picked the right photograph from the pictures shown.	10	30
I was fairly certain that the person I selected was the thief. I couldn't say for sure though because there was another woman in the lineup that looked similar to who I remembered doing it.	45	65
Well, I am pretty good with faces. I am almost sure now that she was the one. I have a real good memory for faces.	77	97
I answered the question as being 20% certain in my identification because I was not certain at all. The photos of female thieves that we were shown had some similarities in their appearance including their hair and the way that they wore it, which made it more difficult to be certain as to my identification.	40	20
I recall that none of the photographs really seemed to identify the thief. I thought that there might be about a 1 in 3 chance that I had identified the correct photograph, or, for that matter, that any of the photographs identified the thief.	50	30

Note: Slight modifications to responses made for spelling, length and clarity.

Table 1.2

Concurrent Detector Example Responses- Study 1

Response	Initial Confidence	Displayed Confidence
If I were only 20% certain about my choice, I must've not thought the girl was in the line-up at all or perhaps, I selected the wrong photo. I usually don't rate my choice so low unless something has happened.	40	20
Did I? 100%? I remember a nice-looking girl with grey eyes and dark hair. I do not remember selecting 100%, honestly.	80	100
I guess that at the moment, I remembered the girl by some identifiable feature. Honestly, I don't remember might response and I am surprised that I was 98% sure because if I saw the girl today, I probably would not recognize her if she slapped me in the face. If I said that I was 98% that she wasn't any of the girls pictured, that might make more sense to me. I sort of remember her having her back to the camera most of the time.	81	98
Wow, I must have been feeling confident that day... I believe there was more than one woman who I thought could have been the thief, but the one I chose (checking my notes from that day, I scribbled down both 1 and 4, I think I chose 4) seemed to stir the most gut reaction.	70	90

Note: Slight modifications to responses made for spelling, length and clarity.

Table 1.3

Posttest Questions- Study 1 Restricted Sample

Question	*F*	*p*	η_p^2
View*	4.36	.013	.020
Face*	5.46	.002	.030
Attention	1.52	.219	.007
Distance	1.29	.275	.014
Time	1.26	.286	.006
Basis	2.16	.116	.010
Ease	1.34	.262	.006
IDTime	0.56	.560	.003
Image	0.42	.651	.002
Willing	2.27	.104	.011
Strangers	2.67	.071	.013
Trust	1.14	.320	.005

Note: Analyses conducted using the restricted sample. Full description of items in the question column are in Appendix C. *indicates significant difference between control and CD condition at $\alpha = .05$ level using Bonferroni correction.

Table 1.4

Posttest Questions- Study 1 Full Sample

Question	*F*	*p*	η_p^2
View*	5.48	.004	.021
Face*	12.59	<.001	.048
Attention	3.66	.026	.014
Distance	0.69	.503	.003
Time	2.05	.130	.008
Basis*	10.02	<.001	.039
Ease	6.76	.001	.026
IDTime	3.21	.041	.013
Image	2.99	.051	.012
Willing	3.15	.044	.012
Strangers*	4.44	.012	.017
Trust	5.74	.003	.022

Note: Analyses conducted using the full sample. Full description of items in the question column are in Appendix C. *indicates significant difference between control and CD condition at $\alpha = .05$ level using Bonferroni correction.

Table 2.1

Concurrent Detectors Example Responses- Study 2

Response	Initial Confidence	Displayed Confidence
I feel like I reported being more certain, but I was certain because I paid more attention to what she looked like rather than what she was stealing. I wasn't sure what I was supposed to be memorizing from the scene, so I focused a lot of my attention on remembering what she looked like.	60	40
I was fairly certain it was her, but all the people looked the same. To be honest my certainty may have been below 80%.	60	80
I must've had some doubt because a couple of the girls in the pictures had similar pictures.	100	80
I think there were a few key facial features that also helped potentially identify the thief. Indeed, I think if I hadn't just been told my certainty score was 56%, I would have said it should have been closer to 75% on the day of and would be 65-75% today as I take this second survey.	76	56
I don't believe I said I was 0% certain actually, but sometimes it is hard to identify people in the lineup because they often don't look the same as they did during the crime. Hair changes, facial changes and the like increase the level of doubt. In a lineup situation, you also feel more compelled to be correct which further increases self-doubt about what you remember seeing.	0	20

Note: Slight modifications to responses made for spelling and clarity.

Table 2.2

Posttest Questions- Study 2 Restricted Sample

Question	*F*	*p*	η_p^2
View	2.72	.029	.022
Face	4.84	.002	.033
Attention	3.08	.016	.025
Distance	3.22	.013	.026
Time	0.66	.624	.005
Basis	4.44	.002	.035
Ease	7.34	<.001	.057
IDTime	8.28	<.001	.064
Image	0.88	.474	.007
Willing	7.80	<.001	.060
Strangers	2.8	.051	.019
Trust	1.03	.392	.008

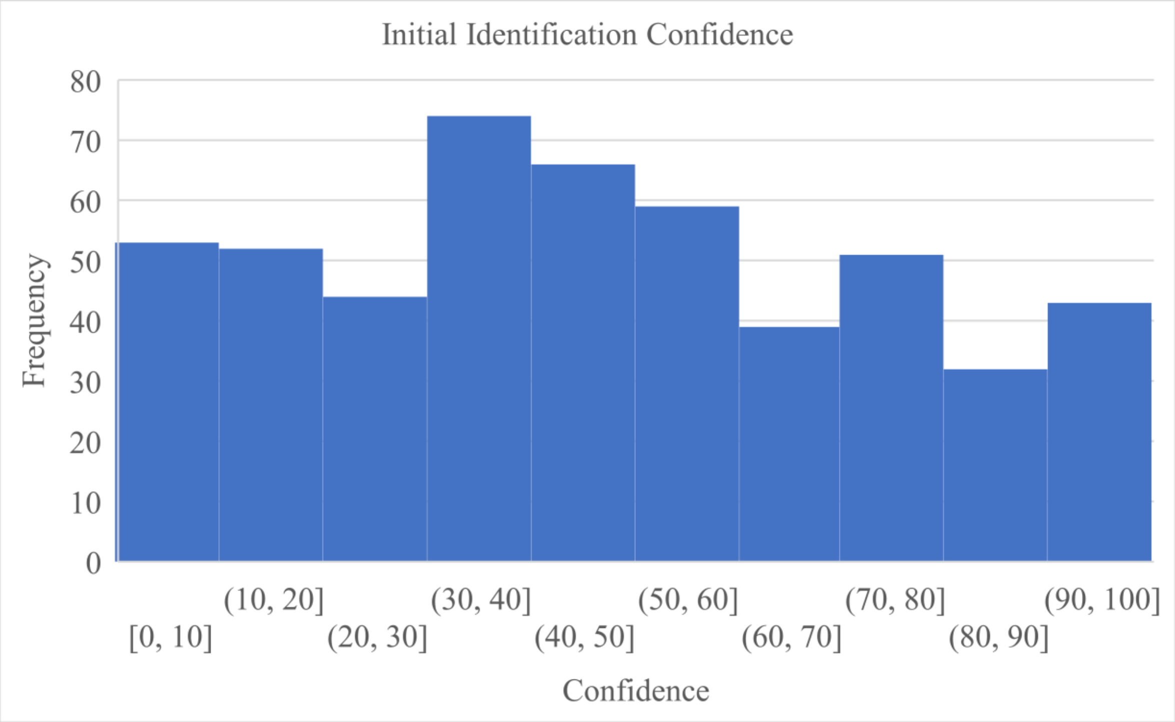


Figure 1.1. Initial identification confidence for the full sample for Study 1.

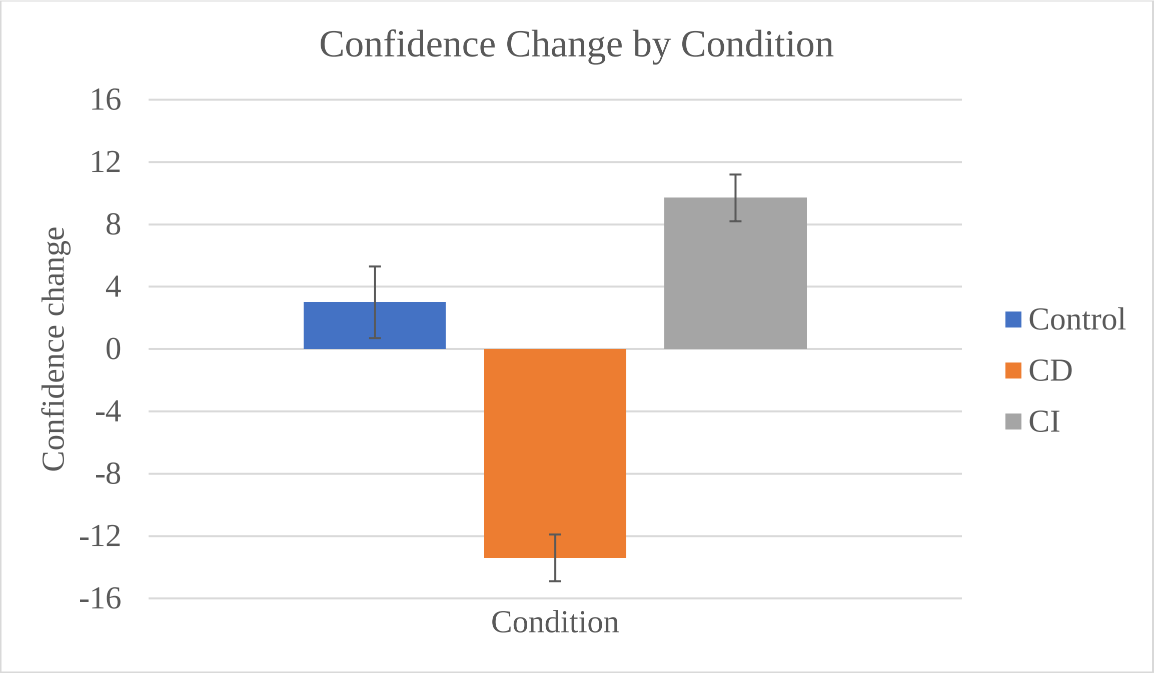


Figure 1.2. Confidence change over time in Study 1 for the restricted sample. Higher numbers on confidence change score indicate confidence inflation from session one to session two. Error bars are for standard errors.

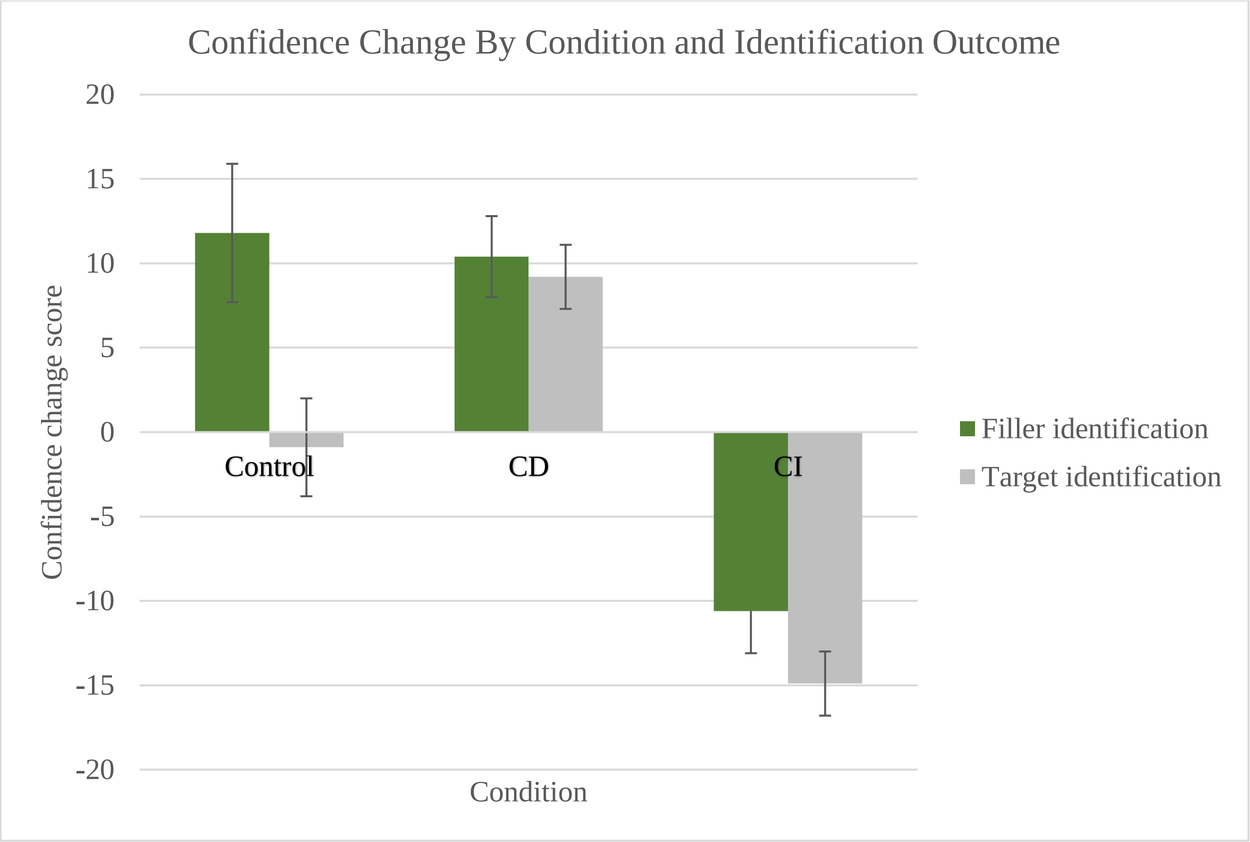


Figure 1.3. Confidence change by condition and identification outcome in Study 1, restricted sample. Positive numbers on the y-axis indicate confidence inflation and negative numbers indicate confidence deflation. Error bars represent standard errors.

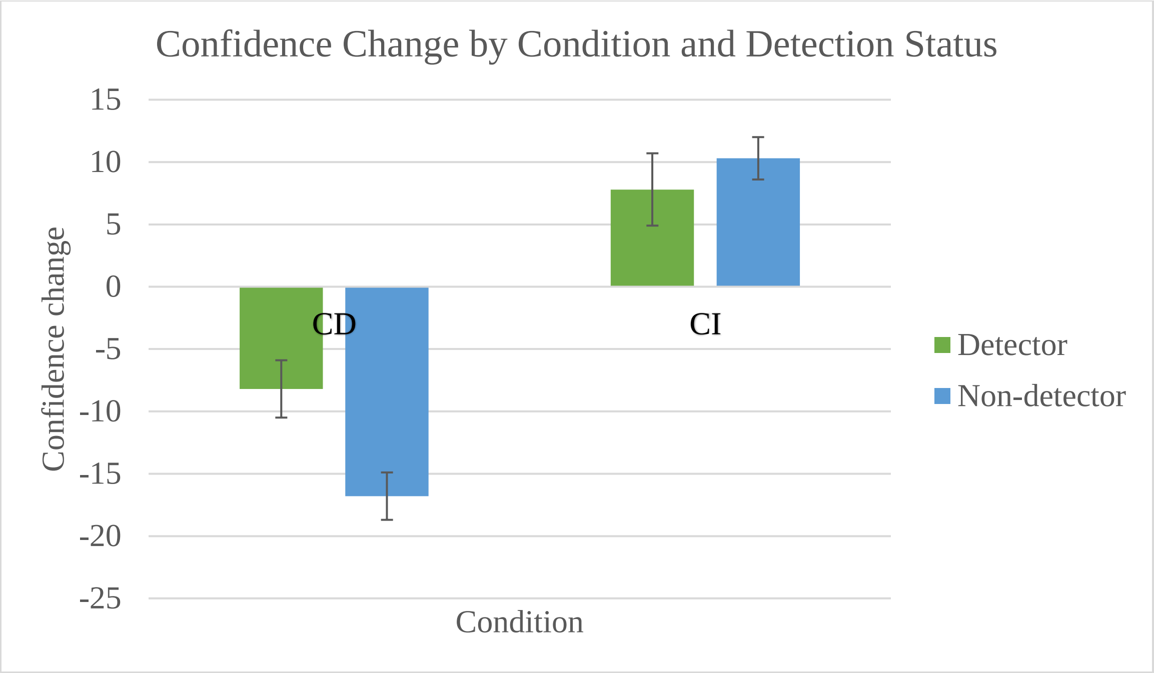


Figure 1.4. Confidence change over time by identification outcome and detection status in Study 1 for the restricted sample. Detection status is for the retrospective detection measure. Positive numbers on the y-axis indicate confidence inflation and negative numbers indicate confidence deflation. Error bars represent standard errors.

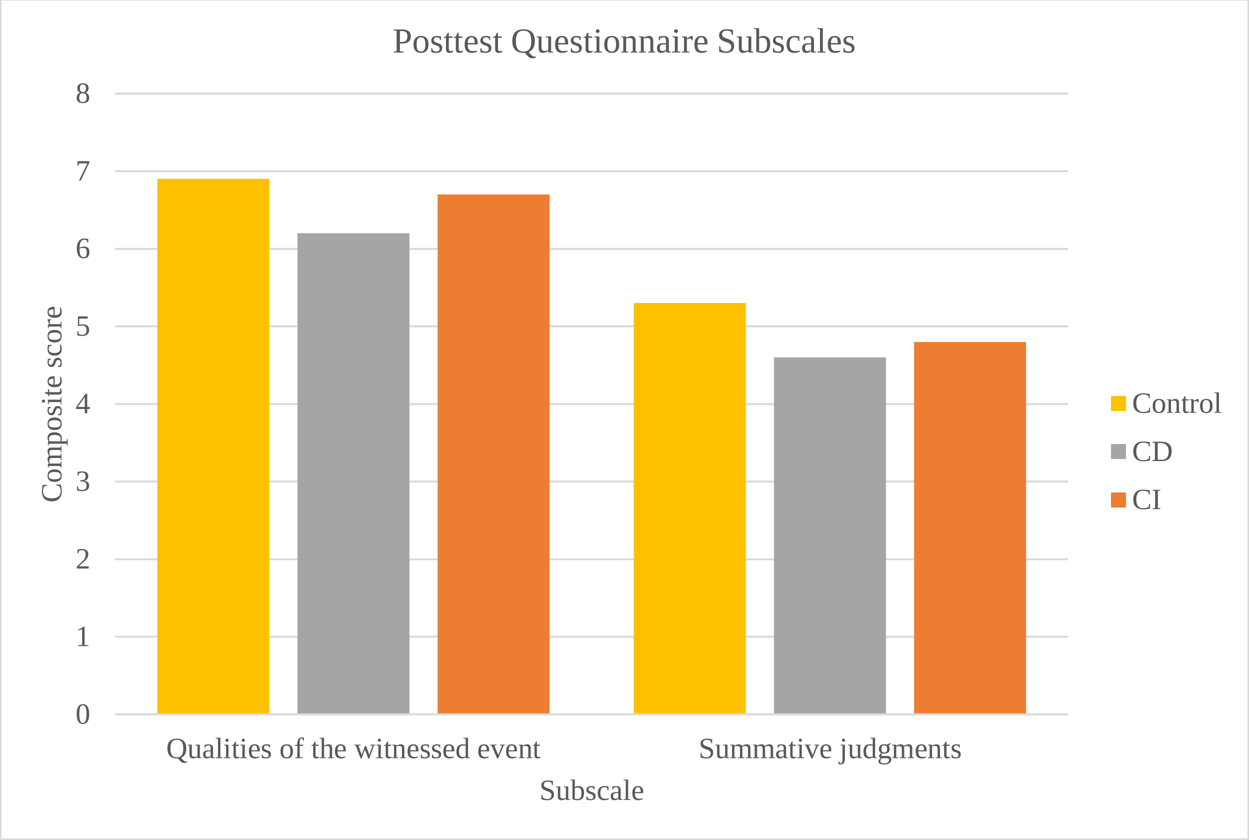


Figure 1.5. Results of posttest questionnaire subscales by condition for Study 1, restricted sample.

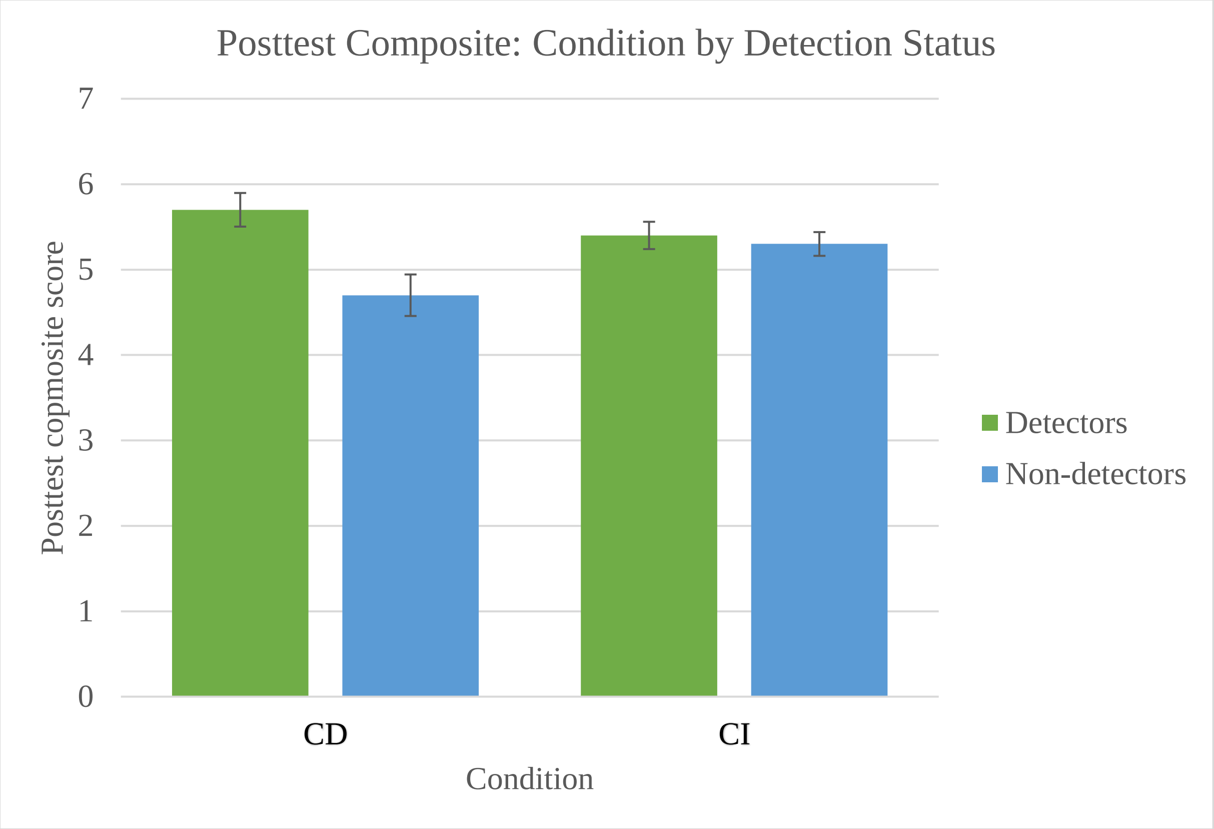


Figure 1.6. Posttest composite of witnessed experience questions by condition and detection status for the restricted sample in Study 1. Detection status refers to retrospective detectors. Error bars represent standard errors.

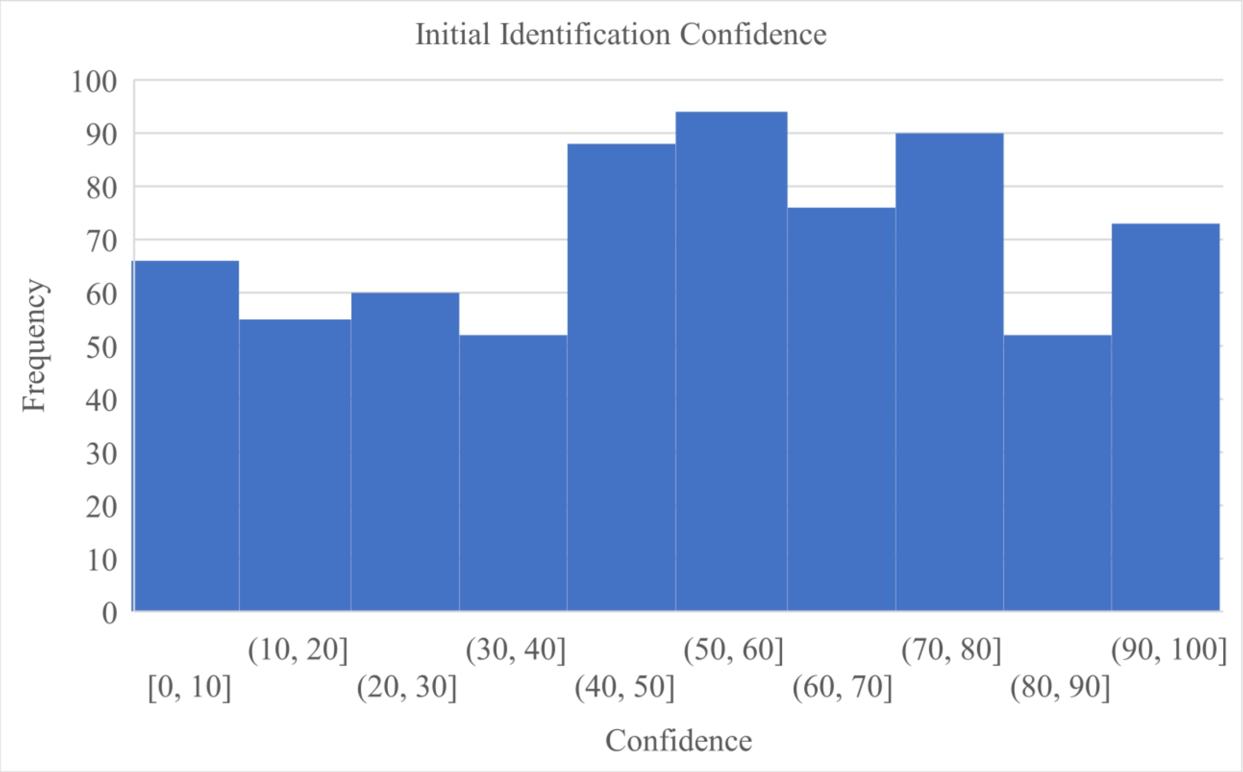


Figure 2.1. Initial identification confidence for the full sample for Study 2.

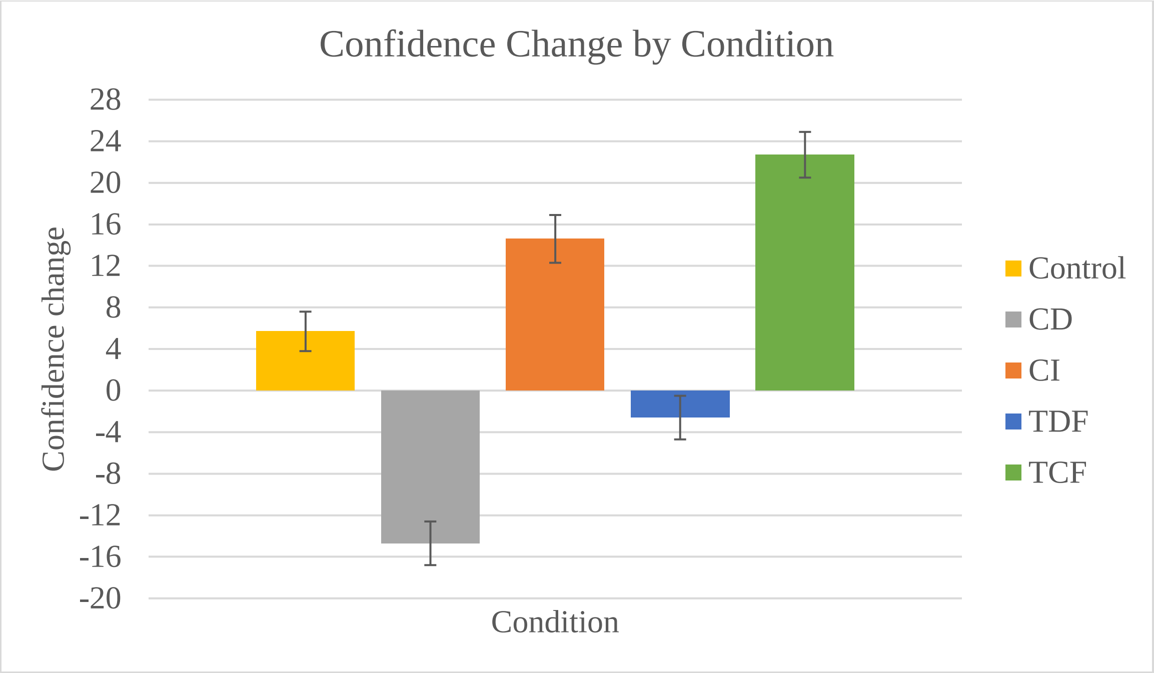


Figure 2.2. Confidence change over time by condition for Study 2, restricted sample. Error bars represent the standard errors.

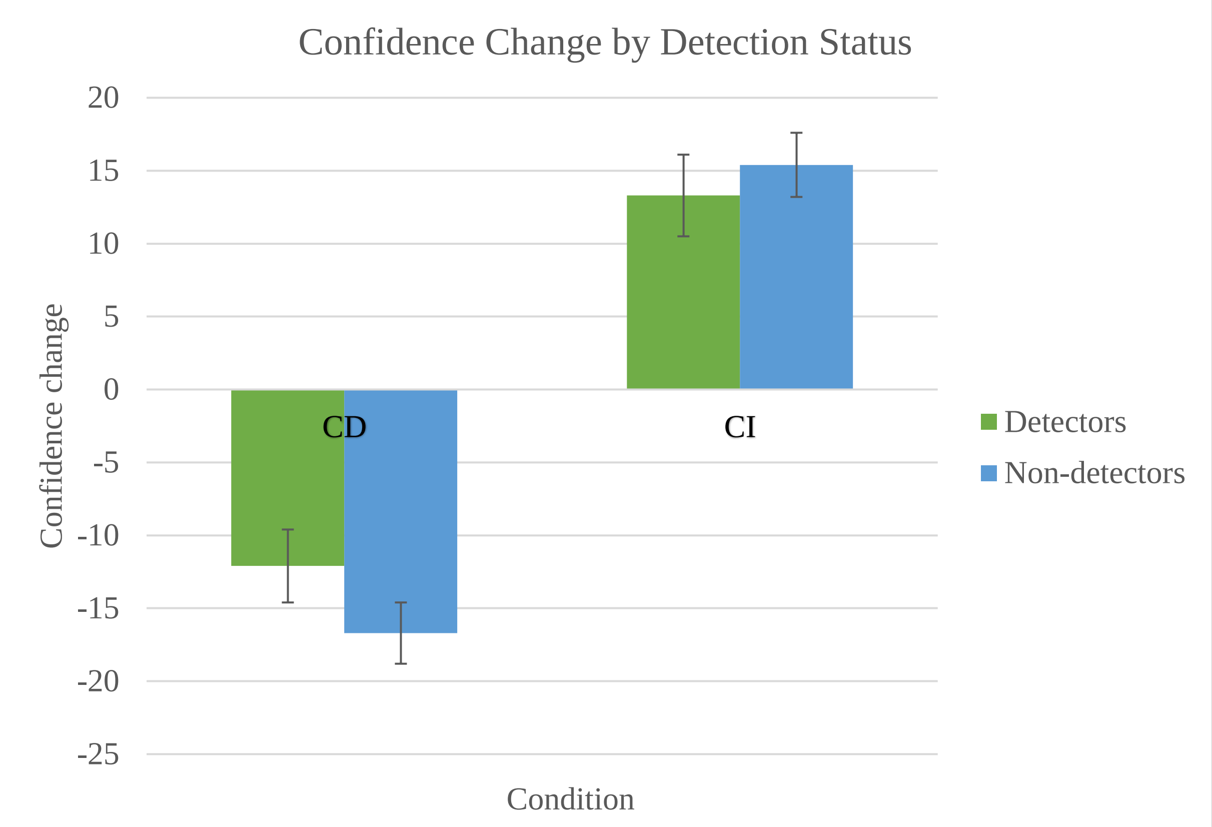


Figure 2.3. Detection status is for retrospective detectors. Confidence change over time by identification outcome in Study 2. Positive numbers indicate confidence inflation and negative numbers indicate confidence deflation. Error bars represent standard errors.

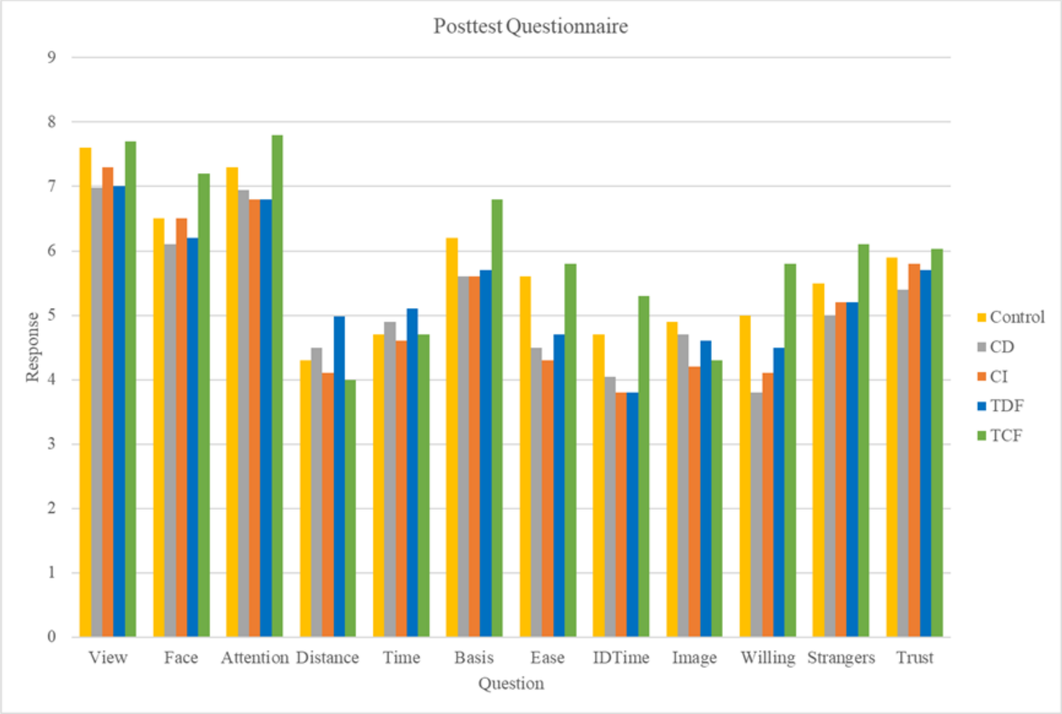


Figure 2.4. Results of posttest questionnaire items by condition for Study 2, restricted sample.

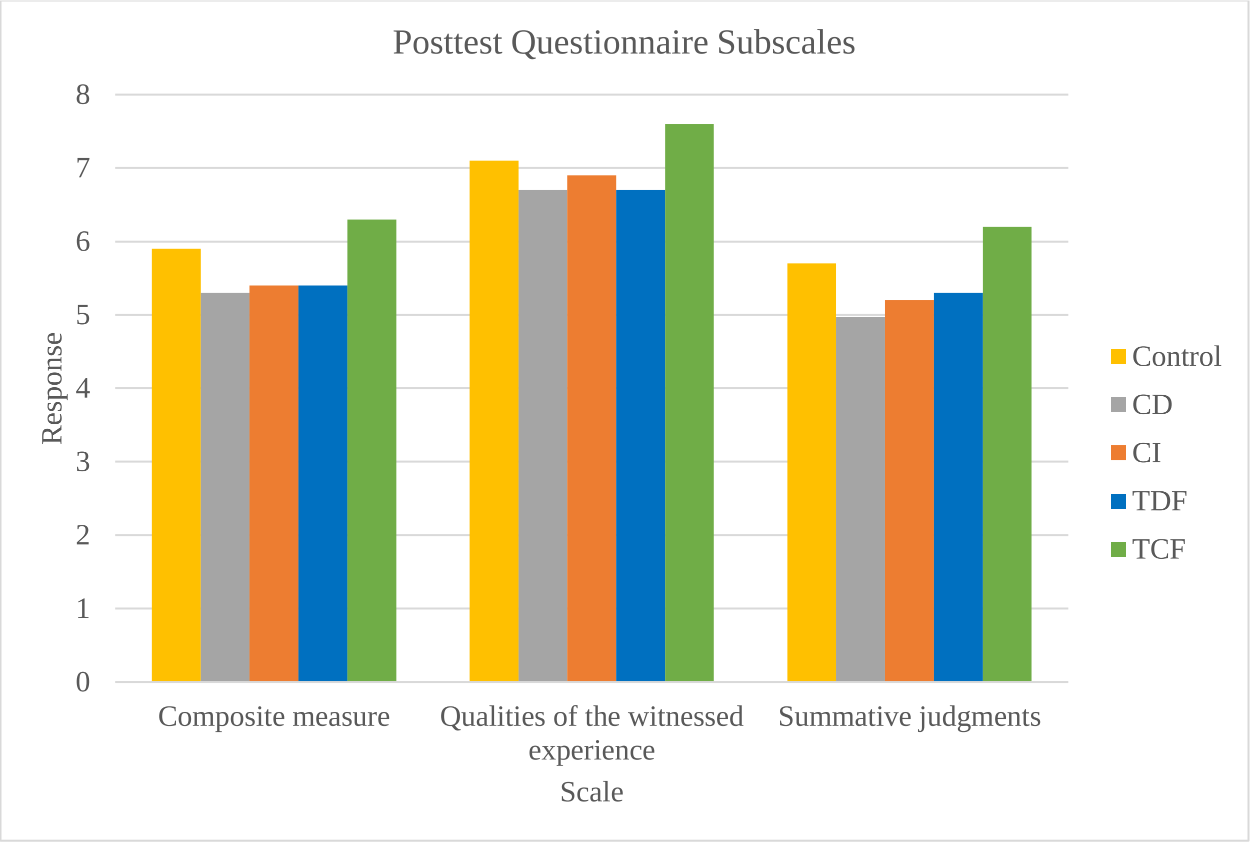


Figure 2.5. Results of posttest composite measures by condition for Study 2, restricted sample.

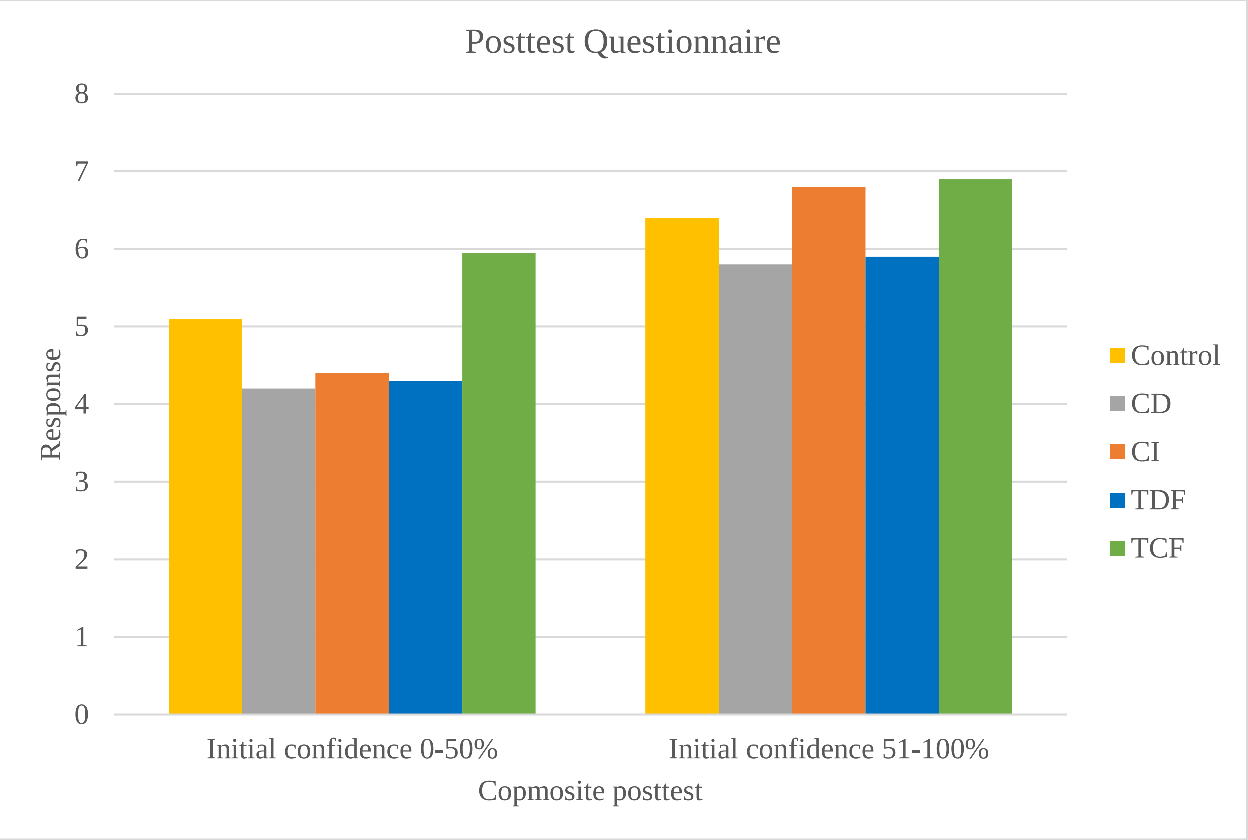


Figure 2.6. Difference in the posttest composite measure for participants above and below 50% on initial confidence for the full sample in Study 2.