UNIVERSITY OF CALIFORNIA, SAN DIEGO

Monitoring recognition memory: A signal detection analysis of internally and externally generated influences on discriminability and response bias

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Psychology

by

Brent M. Wilson

Committee in charge:

Professor John T. Wixted, Chair
Professor Nicholas Christenfeld
Professor Michael Cole
Professor Scott Desposato
Professor Piotr Winkielman

2017

The dissertation of Brent M. Wilson is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2017

DEDICATION

For Edmund Fantino

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I thank my coauthors John Wixted, Edmund Fantino, Stephanie Stolarz-Fantino, Laura Mickes, Matthew Evrard, and Travis Seale-Carlisle.

Chapter 2, in full, is a reprint of materials as it appears in Wilson, B. M., Mickes, L., Stolarz-Fantino, S., Evrard, M., & Fantino, E. (2015). Increased false-memory susceptibility after mindfulness meditation. *Psychological Science*, *26*(10), 1567-1573. The dissertation author was the primary investigator and author of this manuscript.

Chapter 3, in full, has been submitted for publication as Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. "The Effects of Verbal Descriptions on Performance in Lineups and Showups." The dissertation author was the primary investigator and author of this manuscript.

Chapter 4, in full, has been submitted for publication as Wilson, B. M. & Wixted, J. T. "The Cross-Race Effect in Eyewitness Identification: Reduced Discriminability Does Not Necessarily Imply Reduced Reliability." The dissertation author was the primary investigator and author of this manuscript.

VITA

Education

Ph.D. in Psychology, University of California, San Diego 2017

M.A. in Social Sciences, University of Chicago 2010

B.A., Summa Cum Laude, in Economics and Psychology, University of Illinois at Urbana-Champaign 2008

Publications

Wilson, B. M., Mickes, L., Stolarz-Fantino, S., Evrard, M., & Fantino, E. (2015) Increased False-Memory Susceptibility After Mindfulness Meditation. *Psychological Science*, *26(10)*, 1567-1573. doi: http://dx.doi.org/10.1177/0956797615593705

Selected media coverage: *Pacific Standard / The Daily Beast* / *Yahoo Health* / *The Telegraph* / *Shape Magazine* / *Daily Mail / UCSD Guardian / APS*

Wilson B. M., Stolarz-Fantino S., & Fantino E. (2013) Regulating the Way to Obesity: Unintended Consequences of Limiting Sugary Drink Sizes. *PLoS ONE*, *8(4)*, e61081. doi:10.1371/journal.pone.0061081

Selected media coverage: *CBS News / Smithsonian / Los Angeles Times (1) / Los Angeles Times (2) / New York Daily News / Washington Post / Forbes / The Today Show / Montreal Gazette / Daily Mail / Huffington Post / Fox News / Yahoo News / The Washington Times / UCSD Guardian*

Abstracts and Conference Presentations

Wilson, B. M., Vo, K., & Wixted, J.T. (2016) The Cross-Race Effect in Eyewitness Identification: Reduced Discriminability Does Not Necessarily Imply Reduced Reliability. *Abstracts of the Psychonomic Society*, Volume 21, 294

Wilson, B. M., Fantino, E., & Mickes, L. (2015) Re-examining the Verbal Overshadowing Effect With Showups. *Abstracts of the Psychonomic Society*, Volume 20, 255

Wilson, B. M., Mickes, L., Evrard, M., Stolarz-Fantino S., & Fantino, E. (2014) False Memory Susceptibility Following Mindfulness Meditation. *Abstracts of the Psychonomic Society*, Volume 19, 255

Wilson, B. M., Fantino, E., Stolarz-Fantino, S., & Mickes, L. (2013) Identifying the Guilty Suspect From a Lineup: Does Crime Severity Matter? *Abstracts of the Psychonomic Society*, Volume 18, 124

Professional Affiliations
Association for Psychological Science
Psychonomic Society

Teaching Experience
Behavior Modification (TA: Summer 2012)
Behavioral Neuroscience (TA: Summer 2012)
Choice and Self-Control (TA: Spring 2013)
Criminology (TA: Fall 2012, Fall 2013)
Control and Analysis of Human Behavior (TA: Summer 2013, Spring 2014)
Cognitive Psychology (Instructor of Record: Summer 2016, Winter 2017; TA: Winter 2015)
Introduction to Psychology (TA: Winter 2014, Fall 2014)
Interpersonal Relationships (TA: Summer 2015)
Memory and Amnesia (TA: Winter 2016)
Psychology of Food and Behavior (TA: Summer 2013, Summer 2014, Summer 2015)
Psychology and the Law (TA: Spring 2016, Spring 2017)
Sensory Neuroscience (TA: Winter 2013)
Social Psychology (TA: Fall 2011, Spring 2014, Fall 2016)

Professional Service
Ad Hoc Reviewer: *International Journal of Obesity, Mindfulness, PLoS ONE, Psychological Science*

ABSTRACT OF THE DISSERTATION

Monitoring recognition memory: A signal detection analysis of internally and externally generated influences on discriminability and response bias

by

Brent M. Wilson

Doctor of Philosophy in Psychology

University of California, San Diego, 2017

Professor John T. Wixted, Chair

This dissertation contributes to a growing body of research that attempts to bridge the chasm between basic and applied memory research. Its basic approach is to use signal detection theory to analyze higher-level cognitive components that influence recognition memory. The work consists of three research papers that examine the effects of internal and external sources of memory on discriminability and response bias in order to better understand memory in the real world. Paper one provides new evidence supporting a basic assumption of reality-monitoring theory that certain cognitive operations are important for knowing whether or not a memory comes from an internal source.

This research demonstrates that people are more susceptible to false memories after completing mindfulness training because their reality-monitoring accuracy is reduced. Paper two examines the verbal overshadowing effect (where people are worse at correctly identifying someone from a police lineup after providing a verbal description of a face) with receiver operating characteristic (ROC) analysis to determine if that well-known effect is due to differences in actual memorability rather than differences in response bias. This research indicates that internally-generated information can be confused with an external memory source when the internally-generated information is not sufficiently detailed. Paper three examines the cross-race effect wherein memory is worse for a person of a different race than a person of the same race. This research indicates that although memory is worse in terms of discriminability, high-confidence identifications are just as reliable for a cross-race face as for a same-race face.

CHAPTER 1

Introduction to the Dissertation

Memory research is often sharply divided into two broad types: basic and applied. Basic memory research attempts to understand how memory works and focuses on testing specific models of memory. Applied memory research, on the other hand, focuses on how memory operates in the real world but often disregards well-developed memory models from basic research. My dissertation attempts to bridge the chasm between basic and applied memory research by using signal detection theory to examine higher-level cognitive components that influence memory. Its main contribution consists of three research papers that use a model testing approach based on signal detection theory to better understand memory in the real world.

Even though the use of signal detection theory to interpret recognition memory data is often absent from applied research that examines memory in the real world, it is standard practice in basic memory research. Signal detection theory is important for ensuring that a particular measure of accuracy is actually answering the question researchers want to know (Rotello, Heit, & Dubé, 2015). In particular, whereas intuitively plausible measures of accuracy typically conflate discriminability (the ability to distinguish between two states of the world) and response bias (the tendency to choose one response option over the other), signal detection theory separates these separate processes into separate measures.

When discriminability and response bias are not separately assessed, an experimental manipulation that affects response bias without affecting

discriminability can create the false impression that memory differs across the two conditions. For example, a difference in response bias can be observed when an eyewitness is asked to identify a suspect in a police lineup. If eyewitnesses in one condition are led to believe that it is important to avoid falsely identifying an innocent person, this will likely result in a more conservative response bias compared to the other condition. The change in response bias will affect an intuitive accuracy measure (e.g., percent correct will differ across conditions) without actually affecting the ability of eyewitnesses to discriminate innocent from guilty suspects. A signal detection analysis would reveal that response bias was affected by the experimental manipulation whereas discriminability remained unchanged.

Usually the question of interest is whether or not an experimental manipulation influences discriminability. Inducing people to change their response bias in one direction or the other (using instructions) is usually easier than actually improving discriminability. Improving discriminability (not changing response bias) is the goal of most applied research on memory. The reason is that if discriminability is increased, then the two types of errors that occur on recognition memory tasks – misidentifying an item as having been seen before when it was not and failing to identify an item as having been seen before when it was – can be minimized. Different measures of accuracy, however, often conflate these two aspects of memory performance making it difficult or impossible to tell exactly what effect a particular manipulation is having.

In the context of memory, discriminability theoretically refers to the overlap of two distributions that represent the memory strengths associated with the test items. In most situations (and in all of those in my dissertation) these two distributions are the target distribution and the foil distribution. These memory-strength distributions represent the subjective "evidence values" upon which recognition decisions are based. The target distribution is the distribution of evidence values for items that were actually encountered on an earlier study list. The foil distribution is the distribution of evidence values for items that were *not* encountered on an earlier study list. If these distributions happened to be completely overlapping in a particular condition, the target and foil distributions would overlap and discriminability would be zero. An example can be seen in Figure 1.1. At the other extreme, if these distributions were fully separated with virtually no overlap between the target and foil distributions, discriminability would be extremely high (i.e., responding would be essentially error-free). An example of near-perfect discriminability is illustrated in Figure 1.2.

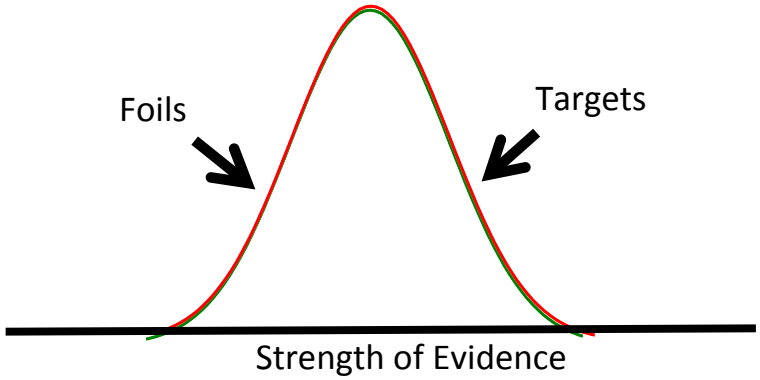


Figure 1.1. Distribution of evidence values for zero discriminability.

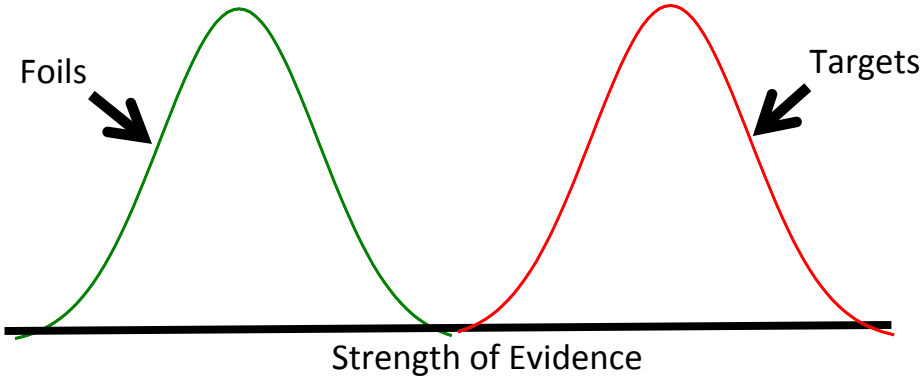


Figure 1.2. Distribution of evidence values for infinitely high discriminability.

For a given test item, a decision about whether or not the item is a target or a foil is made by first setting a decision criterion on the strength-of-evidence axis. The concept of discriminability is not dependent on where a person places a particular decision criterion. The placement of the decision criterion is not fixed but is determined by the participant's response bias. In some situations people

may have a liberal response bias (where the decision criterion is far to the left). Figure 1.3 shows an example of a liberal decision criterion for one particular level of discriminability. In other situations people may have a conservative response bias (where the decision criterion is far to the right). Figure 1.4 shows an example of a conservative decision criterion for the same level of discriminability in Figure 1.3.

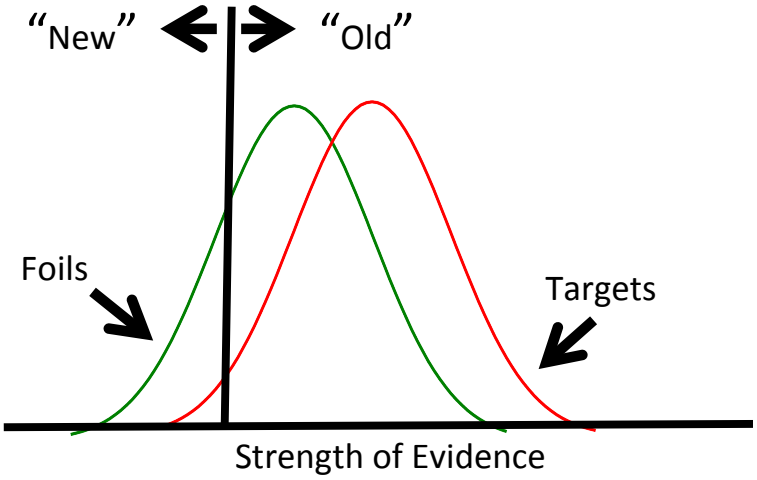


Figure 1.3. Example of a liberal decision criterion for one particular level of discriminability.

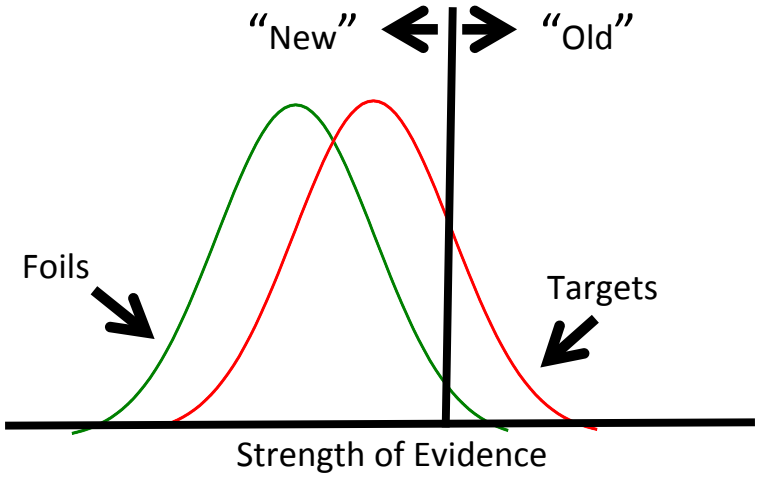


Figure 1.4. Example of a conservative decision criterion for the same level of discriminability shown in Figure 1.3.

Signal detection theory has been used for decades to conceptualize recognition memory decisions on laboratory tasks designed to test cognitive models, and it is equally relevant for recognition decisions on laboratory tasks designed to test questions about memory in the real world. In my dissertation, I use signal detection theory to better understand real-world memory. My first paper demonstrates that people are more susceptible to false memories after completing mindfulness training because their reality-monitoring accuracy is reduced. My second paper examines the verbal overshadowing effect (where people are worse at correctly identifying someone from a police lineup after providing a verbal description) with receiver operating characteristic (ROC) analysis to see if the effect in previous research is due to differences in actual memorability (i.e., discriminability) rather than differences in response bias. My

third paper examines the cross-race effect wherein memory is worse for a person of a different race than a person of the same race. Although memory is worse, high-confidence identifications may be equally reliable for a cross-race and same-race faces.

As noted above, my first paper uses signal detection theory to understand how mindfulness meditation affects "reality-monitoring" (Johnson & Raye, 1981; Johnson, Raye, Foley, & Foley, 1981; Lindsay, 2008). Reality monitoring refers to the processes people rely on to determine if a memory exists because the experience actually happened in the real world or because it was imagined as happening. Mindfulness meditation focuses attention on the present moment without judgment or evaluation (Bishop et al., 2004). Categorizing and evaluating information during encoding is one of the ways people are able to later differentiate internal and external memories. Mindfulness, therefore, seems like a promising way of attenuating one of the important cues that allows people to know that something was not actually encountered in the real world, namely cognitive operations.

Theoretically, when people themselves generate information (when they imagine it, rather than actually encountering it in the real world) they leave a trace record that helps them later know this information was internally generated. This trace record reflects the outcome of a judgment process (the judgment being "this is information I just imagined"), which might be eliminated through the use of mindfulness meditation. The essence of mindfulness meditation is to

simply observe whatever comes to mind without judgment (i.e., without performing cognitive operations).

Signal detection theory provides a model to clearly conceptualize what memories from the real world should look like and what memories not from the real world (i.e., internally generated) should look like. Rather than being an all-or-nothing process, as high-threshold models assume (see Green & Swets, 1966; Macmillan & Creelman, 2005), a signal detection account of reality monitoring assumes that memories exist on a continuum ranging from memories that generate a strong sense of having been externally generated to memories that generate a strong sense of having been internally generated. If memories were clearly tagged as being one or the other, it would be unlikely that mindfulness meditation would have the ability to entirely change such a strong difference. However, if memories instead have various traces associated with them that make them look more or less internal or external, a manipulation like mindfulness meditation may be able to have the effect of changing the traces associated with memories just enough that it is harder to tell in which category a particular memory belongs.

After exploring how people determine whether or not a memory is actually from the real world (i.e., externally generated), I turn in my second paper to an important line of research that results in potential confusion between actual real-world experience and later internal generations of that experience. Previous research has examined whether or not verbally describing the face of a

perpetrator impairs later memory for that perpetrator. A verbal description that is internally generated by an eyewitness may interfere with the actual, real-world memory of the perpetrator. However, previous research (e.g., Schooler & Engstler-Schooler, 1990; Alogna et al. 2014) was not conducted using a signal detection approach, and the measures of accuracy that were used conflated discriminability and response bias. In order to have evidence that internally-generated descriptions of a perpetrator interfere with the memory representation of the perpetrator actually encountered in the real world, discriminability would need to be lower after providing a verbal description.

To determine if providing an internally-generated verbal description of a face makes the target and foil distributions harder to differentiate, I use receiver operating characteristic (ROC) analysis. ROC analysis clearly differentiates between differences in discriminability and response bias unlike the measures of accuracy used in previous research investigating this topic. In addition to examining the theoretical implications of providing a verbal description using ROC analysis, I explore the applied implications by calculating how likely a high-confidence identification is to be accurate if a verbal description has or has not been provided.

One of the advantages of using signal detection theory to inform research is that it helps to clarify if the result being reported actually addresses the question of interest. Signal detection theory does not only apply a set of tools in which to plug data and calculate a measure of discriminability; it also provides a

way to conceptualize how distributions can move independently of the decision criterion. Even previous research that used signal detection theory to measure discriminability made erroneous conclusions about the effect of an experimental manipulation because the reported measures do not appropriately answer the actual question of interest. In other words, although discriminability is usually the measures of interest, in some key cases, it is not the measure of interest. This is the primary problem I address in my third paper.

My third paper examines whether or not people are able to appropriately adjust their high-confidence decision criterion in a situation where their memory of a perpetrator is likely to be worse. Much previous research has examined the cross-race effect wherein people have worse memory for faces of different races than faces of the same race (Malpass & Kravitz, 1969; Meissner & Brigham, 2001; Sporer, 2001). In signal detection terms, this means the target and foil distributions overlap more for cross-race identifications than same-race identifications. Previous research has stopped there and has failed to answer the primary real-world question. Are identifications made with high confidence less accurate (less reliable) for cross-race than same-race faces? A measure of discriminability does not answer this question, but the field has believed otherwise for decades.

Signal detection theory provides the insight that the decision criterion can move independently of the distributions. This means that even if the target and foil distributions overlap more for cross-race identifications than same-race

identifications, high-confidence accuracy will not necessarily be lower for cross-race identifications than same-race identifications. This is a situation where the question of interest involves integrating the knowledge that distributions can move and the knowledge that the decision criterion can also move, and it illustrates the fact that signal detection theory is most important for framing the question of interest (i.e., it is not simply a toolbox of statistics).

These three papers in total enhance understanding of memory in the real world by using signal detection theory to clarify and inform the appropriate question of interest. I first examine how people are able to figure out if something is actually from the real world and how internally-generated representations can make this challenging. I then examine a specific real-world challenge for the legal system wherein an internally-generated verbal representation can interfere with the real-world representation of a face. Finally, I examine if people are able to set their decision criterion appropriately high in situations where target and foil distributions are more overlapping because the memory is of a cross-race face.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., & Birt, A. R., . . .Zwaan, R.A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556-579.

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., . . . Devins, G. (2004). Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice*, *11*, 230–241. doi:10.1093/clipsy/bph077

Green, D. M., & Swets J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*, 67–85.

Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *The American Journal of Psychology*, *94*, 37–64.

Lindsay, D. S. (2008). Source monitoring. In J. Byrne (Series Ed.) & H. L. Roediger III (Vol. Ed.), *Learning and memory: A comprehensive reference. Vol. 2. Cognitive psychology of memory* (pp. 325–348). Oxford, England: Elsevier. doi:10.1016/B978-012370509-9.00175-3

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, *13*(4), 330-334.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3-35.

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*(4), 944-954.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22,* 36-71.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law*, *7*(1), 36-97.

CHAPTER 2

Increased False-Memory Susceptibility After Mindfulness Meditation

Research Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
2015, Vol. 26(10) 1567–1573

DOI: 10.1177/0956797615593705
pss.sagepub.com
SAGE

Increased False-Memory Susceptibility After Mindfulness Meditation

Brent M. Wilson[1], Laura Mickes[2], Stephanie Stolarz-Fantino[1], Matthew Evrard[1], and Edmund Fantino[1]
[1]Department of Psychology, University of California, San Diego, and [2]Department of Psychology, Royal Holloway, University of London

Abstract
The effect of mindfulness meditation on false-memory susceptibility was examined in three experiments. Because mindfulness meditation encourages judgment-free thoughts and feelings, we predicted that participants in the mindfulness condition would be especially likely to form false memories. In two experiments, participants were randomly assigned to either a mindfulness induction, in which they were instructed to focus attention on their breathing, or a mind-wandering induction, in which they were instructed to think about whatever came to mind. The overall number of words from the Deese-Roediger-McDermott paradigm that were correctly recalled did not differ between conditions. However, participants in the mindfulness condition were significantly more likely to report critical nonstudied items than participants in the control condition. In a third experiment, which tested recognition and used a reality-monitoring paradigm, participants had reduced reality-monitoring accuracy after completing the mindfulness induction. These results demonstrate a potential unintended consequence of mindfulness meditation in which memories become less reliable.

Keywords
false memories, mindfulness, Deese-Roediger-McDermott (DRM) paradigm, source monitoring, reality monitoring, signal detection theory

Received 10/22/14; Revision accepted 6/9/15

The concept of mindfulness is pervasive in both popular culture and academic research. Oprah Winfrey, Deepak Chopra, and Dr. Oz (The Dr. Oz Show, 2013) have all extolled the merits of being mindful, and scholarly studies have investigated the benefits of this phenomenon. Mindfulness-based interventions for both physical and psychological disorders have been reported, and these include reduced pain intensity for patients with chronic pain (Reiner, Tibi, & Lipsitz, 2013), improved psychological well-being (Brown & Ryan, 2003), reduced levels of stress and anxiety (Astin, 1997; Jain et al., 2007; Rosenzweig, Reibel, Greeson, Brainard, & Hojat, 2003; Shapiro, Schwartz, & Bonner, 1998), and decreased depression in older adults (Geschwind, Peeters, Drukker, van Os, & Wichers, 2011). Mindfulness meditation focuses attention on the present moment in an accepting and nonjudgmental manner (Baer, Smith, & Allen, 2004; Brown & Ryan, 2003; Kabat-Zinn, 2013). Each thought, feeling, and sensation is acknowledged and accepted without judgment or evaluation (Bishop et al., 2004; Kabat-Zinn, 2013; Segal, Williams, & Teasdale, 2012; Teasdale, 1999). As Kabat-Zinn (2013) noted, "the practice involves suspending judgment and just watching *whatever* [emphasis in original] comes up" (p. 23).

In contrast to the myriad benefits of mindfulness, it may also increase false-memory susceptibility by affecting the cognitive operations needed to distinguish between internal and external sources of information. According to the source-monitoring framework, false memories occur because of a failure to distinguish the origin of a memory (Johnson, Hashtroudi, & Lindsay, 1993; Lindsay, 2008). When the origin of a memory is misattributed, information from one context is falsely remembered as having

Corresponding Author:
Brent M. Wilson, Department of Psychology, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093
E-mail: b6wilson@ucsd.edu

been part of a different context. Source-monitoring errors can arise as a consequence of confusing memory sources. Confusion can occur between two external sources as well as between an internally generated source and an external one (Johnson et al., 1993).

Reality monitoring is the process of discriminating between internally generated and external memory sources (Johnson & Raye, 1981). Information that people generate themselves is usually associated with cognitive operations (i.e., mental processes involved in the generation of information) that leave a trace and later provide cues that the information was internally generated rather than actually encountered in the external world (Johnson, Raye, Foley, & Foley, 1981; Lindsay, 2008). If focusing mindful attention without judgment results in the suspension of cognitive operations (and thus the elimination of the trace records those operations would otherwise leave), people will have greater difficulty differentiating internal and external sources of information. That is, mindfulness training might increase the risk for false memories because internally generated memories would lack the cues that are ordinarily used to help identify them as having been internally generated.

In the first two experiments, we examined the effect of mindfulness meditation on false-memory susceptibility using the Deese-Roediger-McDermott (DRM) paradigm (Roediger & McDermott, 1995). The DRM is the paradigm most widely used to test false memories (Brainerd & Reyna, 2005). The procedure involves presenting lists of closely related words and then testing memory with either recall or recognition. For each list, there is a word (the critical item) that is closely related to the words on the list but is not on the list. The critical item is strongly activated by the other words on the list, and it can be falsely remembered if people mistake this strong internal activation for an actual memory of the word. For example, the word list *garbage, waste, can, refuse, sewage, bag, junk, rubbish, sweep, scraps, pile, dump, landfill, debris*, and *litter* can activate the critical item *trash* (list from Roediger, Watson, McDermott, & Gallo, 2001).

In the third experiment, we used a reality-monitoring paradigm and extended the research to recognition memory. If increases in false memories after mindfulness training are due to reduced reality-monitoring abilities, participants will have reduced abilities to discriminate between words actually studied and words internally activated during study but not actually presented.

Experiment 1

Method

Participants. One hundred fifty-three undergraduate students (37 male, 116 female; mean age = 20.7 years, *SD* = 2.4) at the University of California, San Diego, participated in this experiment for course credit. We planned to recruit as many participants as possible before the end of the quarter.

Materials and procedure. Participants sat in individual sound-attenuated rooms and were randomly assigned to receive either a 15-min mindfulness induction or a 15-min mind-wandering induction. In the mindfulness induction, participants listened to a guided focused-breathing exercise recorded by Marilee Bresciani Ludvik at the Rushing to Yoga Foundation. This mindfulness induction was based on a script by Arch and Craske (2006) that had been adapted from work by Kabat-Zinn (1990). It instructed participants to focus attention on their breathing without judgment. The mind-wandering induction, also recorded by Marilee Bresciani Ludvik, instructed participants to think about whatever came to mind. Mind wandering has been used as a control condition in other mindfulness experiments to represent a neutral mental state (e.g., Hafenbrack, Kinias, & Barsade, 2014; Kiken & Shook, 2011).

All participants were then shown the DRM word list for the critical item *trash* (Roediger et al., 2001). Each word was presented in the center of the computer screen for 1.5 s. After all 15 words were presented, participants immediately typed as many words as they could remember.

Results

Participants in the mindfulness condition were significantly more likely to falsely remember seeing the word *trash*, 39%, 95% confidence interval, or CI = [29.15%, 49.46%], than those in the mind-wandering condition, 20%, 95% CI = [12.37%, 31.35%], $z = 2.48$, $p = .014$, Cohen's $d = 0.50$, 95% CI = [0.18, 0.82]. The mean number of correctly recalled words did not significantly differ between the mindfulness condition, 7.02, 95% CI = [6.68, 7.37], and the mind-wandering condition, 6.75, 95% CI = [6.35, 7.15], $t(152) = 1.02$, $p > .250$. For each list, we numbered the recalled words according to the order in which they were recalled. The average position number at which the critical item was reported did not differ significantly between the mindfulness condition (6.3) and the mind-wandering condition (6.1), $t(45) = 0.2$, $p > .250$. The average number of other words falsely recalled did not significantly differ between the mindfulness condition (0.34) and the mind-wandering condition (0.29), $t(152) = 0.45$, $p > .250$.

Experiment 2

Method

Participants. One hundred forty undergraduate students (40 male, 100 female; mean age = 21.5 years, *SD* = 4.3) at the University of California, San Diego,

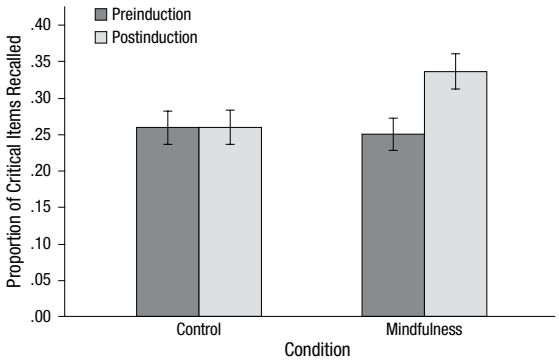


Fig. 1. Average proportion of critical items falsely recalled as being included on the preinduction and postinduction word lists in the mind-wandering (control) and mindfulness conditions. Error bars represent ±1 *SE*.

participated in this experiment for course credit. Using our effect size from Experiment 1, we estimated that we would need 128 participants to have 80% power to detect a statistically significant difference. We planned to recruit as many participants as possible before the end of the quarter, but the minimum was 128 participants.

Materials and procedure. Participants sat in individual sound-attenuated rooms. Six (preinduction) DRM word lists (critical items: *mountain, music, thief, doctor, cold, needle*) from Roediger et al. (2001) were presented in random order. Each word was presented in the center of the computer screen for 1.5 s. After viewing each list, participants immediately typed as many words as they could remember.

After the six lists were completed, the computer randomly assigned participants to either the mindfulness condition or the mind-wandering condition. The inductions were those used in Experiment 1. Participants then completed a different set of six postinduction DRM word lists (critical items: *lamp, trash, slow, wish, foot, window*) also from Roediger et al. (2001) presented in random order.[1] Each word was presented in the center of the computer screen for 1.5 s. Again, after viewing each list, participants immediately typed as many words as they could remember.

Results

In the within-subjects comparison, participants in the mindfulness condition were significantly more likely to falsely recall the critical items after the induction than before the induction, $t(67) = 2.75$, $p = .008$, Cohen's $d = 0.33$, 95% CI = [0.09, 0.58]. Participants in the mind-wandering condition showed no difference in critical-item recall on the preinduction and postinduction lists, $t(71) < 0.001$, $p > .250$, Cohen's $d = 0.00$, 95% CI = [0, 0]. The same results were also found in the between-subjects comparison. Participants in the mindfulness condition were significantly more likely to falsely recall the critical item, $M = .34$, 95% CI = [.29, .38], than were participants in the mind-wandering condition, $M = .26$, 95% CI = [.21, .31], $t(138) = 2.27$, $p = .025$, Cohen's $d = 0.38$, 95% CI = [0.05, 0.72]. This difference remained significant when we controlled for participants' baseline levels of false-memory susceptibility and memory performance using the average critical-item recall and proportion correct during preinduction, $F(1, 136) = 5.78$, $p = .018$. We performed a 2 × 2 analysis of variance and found a significant interaction between condition (mindfulness vs. mind-wandering) and time of recall (preinduction vs. postinduction), $F(1, 138) = 4.22$, $p = .042$. Figure 1 shows the average proportion of critical items falsely recalled as being included on the preinduction and postinduction word lists.

The average proportion of words correctly recalled did not differ significantly between conditions (mindfulness: $M = .46$, 95% CI = [.44, .49]; mind-wandering: $M = .45$, 95% CI = [.43, .48]), $t(138) = 0.66$, $p > .250$, Cohen's $d = 0.11$, 95% CI = [–0.22, 0.44]. The proportion correct was not significantly different even after we controlled for both correct identifications and critical-item recall on the preinduction lists, $F(1, 136) = 1.66$, $p = .200$. Participants in the two conditions did not significantly differ in critical-item recall ($p > .250$) or correct recall ($p > .250$) on the DRM lists completed before receiving the audio inductions. Again, for each list, we numbered the recalled words according to the order in which they were recalled. The average position number at which the critical item was reported did not differ significantly between the mindfulness condition (5.7) and the mind-wandering

condition (5.2), $t(119) = 1.60$, $p = .111$, and did not change significantly after participants completed the mindfulness induction (5.4 for preinduction and 5.7 for postinduction), $t(50) = 0.55$, $p > .250$. The average number of other words falsely recalled did not differ significantly between the mindfulness condition (0.22) and the mind-wandering condition (0.18), $t(138) = 0.96$, $p > .250$, and did not change after participants completed the mindfulness induction (0.22 for both preinduction and postinduction), $t(67) = 0$, $p > .250$.

Discussion

These results provide evidence that false-memory susceptibility increases after completing mindfulness training. The pretest-posttest design of this experiment (as opposed to the design of Experiment 1) also provides evidence that false-memory susceptibility is increased by mindfulness training rather than being decreased by mind wandering. In the next experiment, we extend this work to a reality-monitoring paradigm (Brainerd & Reyna, 2005) to better identify why false memories increase after mindfulness meditation training.

Experiment 3

Method

Participants. Two hundred fifteen undergraduate students (59 male, 156 female; mean age = 20.3 years, $SD = 2.9$) at the University of California, San Diego, participated in this experiment for course credit. On the basis of the effect size from our within-subjects comparison in Experiment 2, we estimated that we would need 75 participants to have 80% power to detect a statistically significant difference. We planned to recruit as many participants as possible before the end of the quarter, but the minimum was 75 participants.

Materials and procedure. Two hundred pairs of strongly associated words (e.g., *foot-shoe, sediment-fossil*) were constructed using databases of word associations (Palermo & Jenkins, 1964; Rotmistrov, 2014). One hundred word pairs were randomly selected for the preinduction study and test phase. The remaining 100 word pairs were then used for the postinduction study and test phase.

Participants sat in individual sound-attenuated rooms. During the preinduction study phase, 1 word from each pair was randomly selected and presented in the center of the computer screen for 1.5 s. The 100 words were presented in random order. After all the words had been presented to participants, the preinduction test phase began immediately. One word from each pair was randomly selected for the test phase and presented in the center of the computer screen. This procedure gave each word an equal probability of being a target or a lure. Participants identified whether the word had appeared on the word list ("old") or had not appeared on the word list ("new") and indicated their level of confidence in each answer.

All participants then listened to the 15-min mindfulness induction used in the first two experiments. After completing the mindfulness induction, participants began the postinduction study phase followed immediately by the postinduction test phase. The procedure was identical to that in the preinduction study and test phase.

Results

We used d' (Macmillan & Creelman, 2005) to compare how well participants were able to discriminate between externally presented (old or target) items and internally generated (new or lure) items. Accuracy (d') was significantly higher for the word lists studied and tested before the mindfulness induction ($M = 1.60$, $SD = 0.71$) than for the word lists studied and tested after the mindfulness induction ($M = 1.42$, $SD = 0.79$), $t(214) = 4.08$, $p < .001$, Cohen's $d = 0.28$, 95% CI = [0.14, 0.41]. With regard to the proportion of words declared to be "old," there was a significant interaction between the status of the word (internal vs. external) and condition (control vs. mindfulness), $F(1, 214) = 20.94$, $p < .001$. The false alarm rate increased significantly after participants completed the mindfulness induction (before: $M = .20$, $SD = .15$; after: $M = .25$, $SD = .18$), $t(214) = 4.49$, $p < .001$, Cohen's $d = 0.31$, 95% CI = [0.17, 0.44], but the hit rate did not change significantly (before: $M = .72$, $SD = .15$; after: $M = .71$, $SD = .16$), $t(214) = 1.55$, $p = .123$, Cohen's $d = 0.11$, 95% CI = [–0.03, 0.24]. Because null-hypothesis significance testing cannot provide evidence in favor of the null, we also calculated the Jeffreys-Zellner-Siow (JZS) Bayes factor for the nonsignificant change in the hit rate (Rouder, Speckman, Sun, Morey, & Iverson, 2009). This method gave 5.65:1 odds in favor of the null hypothesis.

We used c (Macmillan & Creelman, 2005) to measure response bias. Participants had a significantly more liberal response bias (i.e., more of the distribution exceeded the criterion line) after completing the mindfulness induction (before: $M = 0.15$, $SD = 0.40$; after: $M = 0.085$, $SD = 0.47$), $t(214) = 2.61$, $p = .0097$, Cohen's $d = 0.18$, 95% CI = [0.04, 0.31]. However, it is important to note that a change in measured bias does not necessarily entail a change in participants' decision strategy (Wixted & Stretch, 2000).

Discussion

The results of Experiment 3 are consistent with the results from Experiments 1 and 2 and provide additional evidence that mindfulness training increased false-memory susceptibility. Experiment 3 also extends the findings to

recognition memory and to a reality-monitoring paradigm. These findings support the idea that the increase in false memories is due to a reduction in reality-monitoring accuracy. Each word on the study list strongly activates its paired word. Participants are less accurate at discriminating between associated words (internally generated) and words actually studied (external memory source) after completing the mindfulness induction.

General Discussion

Our research adds to and connects the literature on mindfulness meditation and false memories. Whereas the preponderance of research on mindfulness has focused on the beneficial aspects of this phenomenon (Chiesa, Calati, & Serretti, 2011), our study examines a potential adverse effect. When meditators embrace judgment-free awareness and acceptance, their reality-monitoring accuracy may be impaired, increasing their susceptibility to false memories.

Information encountered in the external world is expected to leave a trace record that contains greater sensory detail than information that is internally generated, and this difference in sensory content is one factor that facilitates the discrimination between internally and externally generated information. Johnson et al. (1981) also noted the importance of a second factor: cognitive operations associated with the internal generation of information at the time of encoding. At retrieval, a trace record of those cognitive operations ordinarily helps to identify internally generated information as having been internally generated. However, the nonjudgmental aspect of mindfulness meditation may be expected to reduce this important cue. The essential idea of mindfulness meditation is to observe without judgment or reaction (rather than performing cognitive operations on) whatever comes to mind. The elimination of cognitive operations would therefore have the effect of also eliminating a trace record of such operations that might otherwise help to discriminate between internally and externally generated information on a later memory test. The result would be a decreased ability to discriminate between sources of information (Johnson & Raye, 1981), thereby increasing susceptibility to the DRM false-memory effect.

This argument can be illustrated using a simple signal detection model of a task in which the participant's goal is to discriminate between internally generated (new) and externally presented (old) information (as in Experiment 3). The *x*-axis in the model shown in Figure 2 ranges from strong evidence that a test item was internally generated (at the far left) to strong evidence that the test item was externally presented (at the far right). The distribution of evidence values for internally generated items in the control condition falls farther to the left than

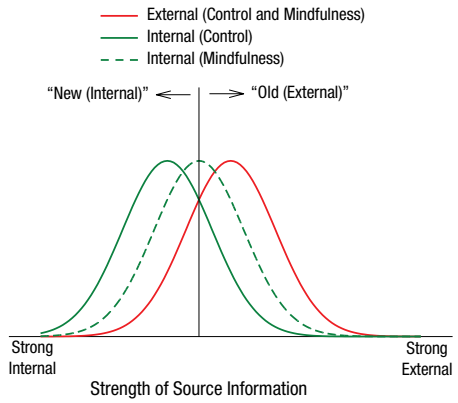


Fig. 2. Signal detection model representing how mindfulness meditation influences the distributions of source information for internally generated and externally presented items (relative to a control condition). According to this model, mindfulness meditation reduces the ability to discriminate between internally generated and externally presented memories by shifting the distribution of internally generated items to the right without influencing the distribution of externally presented items.

the distribution of evidence values for externally presented items. The difference between the two distributions is d'.

As noted earlier, this discrimination is facilitated both by the sensory content of the memory trace (more detailed for externally presented items than for internally generated items) and by the record of cognitive operations associated with the generation of internally generated items. Thus, for example, a test item that falls to the far left (a strong evidence trace for internal generation) might be associated with limited sensory content as well as a trace record of cognitive operations associated with the internal generation of that item. However, in the mindfulness condition, the trace record of cognitive operations is largely reduced. This reduction shifts the distribution associated with internally generated items to the right and increases the false alarm rate (i.e., the proportion of the internal distribution that falls above the decision criterion). A test item that falls to the far right, by contrast, might be associated with considerable sensory content and would also have no trace record of cognitive operations associated with internal generation (because the item was externally presented). Mindfulness, which selectively reduces cognitive operations, would therefore not change the representation of externally presented items, so the same external distribution would apply in both the control and the mindfulness conditions. If the decision criterion remains fixed across conditions, this

increase in the false alarm rate would not be accompanied by a change in the hit rate associated with externally generated items. Thus, the selective change in the false alarm rate would affect measured bias (more liberal in the mindfulness condition than in the control condition) even though the decision criterion remained unchanged.

Measured bias reflects the distance of the criterion line (i.e., the point at which participants switch from responding "new" to responding "old") from the point of intersection for the internal and external distributions. In Figure 2, the point of intersection for the internal and external distributions in the mindfulness condition is farther to the right than the point of intersection for the internal and external distributions in the control condition. This means that the relative position of the criterion line (indicated by the vertical line in the center of the figure) is farther to the left of the intersection of the internal and external distributions in the mindfulness condition than in the control condition. This change in the relative location of the criterion line is why measured bias (c) changes between conditions, even though the absolute location of the criterion line stays the same in this model. Thus, the model predicts that measured bias should be more liberal for the mindfulness condition than for the control condition because of this change in the relative location of the decision criterion (resulting from an increase in the mean of the internal distribution in that condition). This simple model accounts for all of the results observed in Experiment 3, and it explains why false-memory susceptibility increases after mindfulness meditation.

A simple criterion-shift model (in which the distributions remain in the same locations but the criterion line changes) cannot fully account for the Experiment 3 results. Not only did measured bias change between conditions, d' values also changed between conditions. The lower d' value in the mindfulness condition means that the internal and external distributions moved in a manner that resulted in greater overlap between the two distributions. A simple criterion-shift model can explain only the change in measured bias; it cannot explain the change in d' values observed between conditions.

Another possible model assumes that the effect occurs at retrieval rather than during encoding. Such a model can explain the change in d' values but cannot readily explain all of the Experiment 3 results. According to this retrieval-based interpretation, one might assume that participants in the mindfulness condition respond on the basis of familiarity without engaging in recollection of source information (whereas control participants do engage in recollection of source information). In the absence of recollection, the internal distribution in the mindfulness condition would be to the right (in the external direction) of the internal distribution in the control condition because recollection would not count as evidence against familiar-but-imagined items having appeared on the list. By contrast, the external distribution in the mindfulness condition would be to the left (in the internal direction) of the external distribution in the control condition because recollection would not add evidence in favor of target items having appeared on the list. Thus, d' values would be lower for the mindfulness condition, consistent with our results. However, the simplest version of this account would predict a difference in both hit and false alarm rates across conditions with no effect on measured bias, whereas we observed a selective effect on the false alarm rate and a clear effect on measured bias.

Mindfulness meditation appears to reduce reality-monitoring accuracy. By embracing judgment-free awareness and acceptance, meditators can have greater difficulty differentiating internal and external sources of information. As a result, the same aspects of mindfulness that create countless benefits can also have the unintended negative consequence of increasing false-memory susceptibility.

Author Contributions

The initial study concept came from B. M. Wilson and was developed by all authors. B. M. Wilson analyzed the data, and L. Mickes contributed to the data analysis. B. M. Wilson drafted the manuscript, and L. Mickes, S. Stolarz-Fantino, and E. Fantino edited it. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank John Wixted for valuable assistance with the signal detection model and data interpretation.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Note

1. These lists were not counterbalanced between preinduction and postinduction (which is not ideal for the within-subjects comparisons) because the preinduction lists were originally included to serve as covariates in the analysis of the postinduction word lists.

References

Arch, J. J., & Craske, M. G. (2006). Mechanisms of mindfulness: Emotion regulation following a focused breathing induction. *Behaviour Research and Therapy*, *44*, 1849–1858. doi:10.1016/j.brat.2005.12.007

Astin, J. A. (1997). Stress reduction through mindfulness meditation. *Psychotherapy and Psychosomatics*, *66*, 97–106.

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of mindfulness by self-report: The Kentucky Inventory of Mindfulness Skills. *Assessment*, *11*, 191–206. doi:10.1177/1073191104268029

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., . . . Devins, G. (2004). Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice*, *11*, 230–241. doi:10.1093/clipsy/bph077

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York, NY: Oxford University Press.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*, 822–848. doi:10.1037/0022-3514.84.4.822

Chiesa, A., Calati, R., & Serretti, A. (2011). Does mindfulness training improve cognitive abilities? A systematic review of neuropsychological findings. *Clinical Psychology Review*, *31*, 449–464. doi:10.1016/j.cpr.2010.11.003

The Dr. Oz Show. (2013, February 14). Mindfulness meditation can change brain waves [Web log post]. Retrieved from http://blog.doctoroz.com/in-the-news/study-demonstrates-mindfulness-meditation-can-change-brain-waves

Geschwind, N., Peeters, F., Drukker, M., van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial. *Journal of Consulting and Clinical Psychology, 79*, 618–628. doi:10.1037/a0024595

Hafenbrack, A. C., Kinias, Z., & Barsade, S. G. (2014). Debiasing the mind through meditation: Mindfulness and the sunk-cost bias. *Psychological Science*, *25*, 369–376. doi:10.1177/0956797613503853

Jain, S., Shapiro, S. L., Swanick, S., Roesch, S. C., Mills, P. J., Bell, I., & Schwartz, G. E. (2007). A randomized controlled trial of mindfulness meditation versus relaxation training: Effects on distress, positive states of mind, rumination, and distraction. *Annals of Behavioral Medicine*, *33*, 11–21.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*, 67–85.

Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *The American Journal of Psychology*, *94*, 37–64.

Kabat-Zinn, J. (1990). Full catastrophe living: Using the wisdom of your body and mind to face stress, pain, and illness. New York, NY: Dell.

Kabat-Zinn, J. (2013). Full catastrophe living: Using the wisdom of your body and mind to face stress, pain, and illness (Rev. ed.). New York, NY: Bantam Books.

Kiken, L. G., & Shook, N. J. (2011). Looking up: Mindfulness increases positive judgments and reduces negativity bias. *Social Psychological & Personality Science*, *2*, 425–431. doi:10.1177/1948550610396585

Lindsay, D. S. (2008). Source monitoring. In J. Byrne (Series Ed.) & H. L. Roediger III (Vol. Ed.), *Learning and memory: A comprehensive reference. Vol. 2. Cognitive psychology of memory* (pp. 325–348). Oxford, England: Elsevier. doi:10.1016/B978-012370509-9.00175-3

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Palermo, D. S., & Jenkins, J. J. (1964). *Word association norms: Grade school through college*. Minneapolis: University of Minnesota Press.

Reiner, K., Tibi, L., & Lipsitz, J. D. (2013). Do mindfulness-based interventions reduce pain intensity? A critical review of the literature. *Pain Medicine*, *14*, 230–242.

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

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*, 385–407.

Rosenzweig, S., Reibel, D. K., Greeson, J. M., Brainard, G. C., & Hojat, M. (2003). Mindfulness-based stress reduction lowers psychological distress in medical students. *Teaching and Learning in Medicine*, *15*, 88–92. doi:10.1207/S15328015TLM1502_03

Rotmistrov, Y. A. (2014). *Word associations network*. Available from http://wordassociations.net/

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. doi:10.3758/PBR.16.2.225

Segal, Z. V., Williams, J. M. G., & Teasdale, J. D. (2012). *Mindfulness-based cognitive therapy for depression*. New York, NY: Guilford Press.

Shapiro, S. L., Schwartz, G. E., & Bonner, G. (1998). Effects of mindfulness-based stress reduction on medical and premedical students. *Journal of Behavioral Medicine*, *21*, 581–599.

Teasdale, J. D. (1999). Metacognition, mindfulness and the modification of mood disorders. *Clinical Psychology & Psychotherapy*, *6*, 146–155.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, *107*, 368–376.

Chapter 2, in full, is a reprint of materials as it appears in Wilson, B. M., Mickes, L., Stolarz-Fantino, S., Evrard, M., & Fantino, E. (2015). Increased false-memory susceptibility after mindfulness meditation. *Psychological Science*, *26*(10), 1567-1573. The dissertation author was the primary investigator and author of this manuscript.

CHAPTER 3

The Effects of Verbal Descriptions on Performance in Lineups and Showups

Abstract

Verbally describing a face has been found to impair subsequent recognition of that face from a photo lineup, a phenomenon known as the verbal overshadowing effect (Schooler & Engstler-Schooler, 1990). Recently, a large direct replication study successfully reproduced that original finding (Alogna et al. 2014). However, in both the original study and the replication studies, memory was tested using only target-present lineups (i.e., lineups containing the previously-seen target face), making it possible to compute the correct ID rate (i.e., the hit rate) but not the false ID rate (i.e., the false alarm rate). Thus, the lower correct ID rate for the verbal condition could reflect either reduced discriminability or a conservative criterion shift relative to the control condition. In four verbal overshadowing experiments reported here, we measured both correct ID rates and false ID rates using photo lineups (Experiments 1 and 2) or single-photo showups (Experiments 3 and 4). The experimental manipulation (verbally describing the face or not) occurred either immediately after encoding (Experiments 1 and 3) or 20-minutes after encoding (Experiments 2 and 4). In the immediate condition, discriminability did not differ between groups, but in the delayed condition, discriminability was lower in the verbal description group (i.e., a verbal overshadowing effect was observed). A fifth experiment found that the effect of the immediate-vs.-delayed manipulation may be attributable to a change

in the content of verbal descriptions, with the ratio of diagnostic to generic facial features in the descriptions decreasing as delay increases.

Introduction

A police lineup is administered to victims and eyewitnesses to aid criminal investigations. The lineup is a collection of individuals, including the police suspect (who may be innocent or guilty) and a number of fillers (who are known to be innocent and resemble the perpetrator). Verbally reporting the details of a crime is a necessity in the investigative process. Whether or not the very act of reporting details about the perpetrator retrieved from memory impairs later memory for the perpetrator has been a topic of interest and debate for the last several decades. Interest in this topic was triggered by a finding reported by Schooler and Engstler-Schooler (1990) in which participants watched a video of a simulated robbery and either verbally described the perpetrator or engaged in a control task. Participants who gave verbal descriptions were significantly less likely to correctly identify the perpetrator from a lineup test than those in the control condition. This somewhat counterintuitive finding, termed the “verbal overshadowing effect,” has potential implications for the criminal justice system.

Because follow-up research yielded mixed results and a meta-analysis yielded effect sizes much smaller than the original experiments (Meissner & Brigham, 2001), a large direct replication study was recently conducted on two of the original experiments (Experiments 1 and 4 of Schooler & Engstler-Schooler, 1990; Alogna et al. 2014). In both experiments, the main experimental

manipulation was the same: participants either verbally described the perpetrator or took part in a control task. The only difference between the experiments was the order of procedural events. As shown in Figure 3.1, the experimental manipulation took place immediately after presentation of the video in Experiment 4 of Schooler and Engstler-Schooler (Figure 3.1A) or 20 minutes after the presentation of the video in Experiment 1 of Schooler and Engstler-Schooler (Figure 3.1B). The effect replicated: compared to the control condition, the perpetrator was less likely to be identified from the lineup in both experiments, but the effect was much larger when the verbal description was provided 20 minutes after the video (and immediately before the lineup test).

In a typical eyewitness identification study, some participants are presented with a target-present lineup (i.e., a lineup that contains a photo of the guilty suspect) and other participants are presented with a target-absent lineup (i.e., a lineup in which the photo of the guilty suspect has been replaced by a photo of the innocent suspect). The measures of interest are the correct ID rate (the proportion of participants presented with a target-present lineup who correctly identify the guilty suspect) and the false ID rate (the proportion of participants presented with a target-absent lineup who incorrectly identify the innocent suspect). However, the original verbal-overshadowing experiments and the studies that recently replicated them included target-present lineups (i.e., lineups that contained a photo of the guilty suspect), but did not include target-absent lineups. What these studies therefore showed is that the correct ID rate

was lower in the verbal description condition compared to the control condition. The effect of that manipulation on the false ID rate is unknown. Thus, the only safe conclusion is that there was a reduction in the probability of correctly identifying the perpetrator when a verbal description was provided, but whether that reduction in the correct ID rate occurred because of reduced discriminability or because of a more conservative response bias is unknown. Distinguishing between those alternative interpretations requires that the probability of identifying the innocent suspect be measured as well (e.g., Clare & Lewandowsky, 2004, Mickes, 2016; Mickes & Wixted, 2015; Rotello, Heit, & Dube, 2015; Smith & Flowe, 2015). Moreover, the applied implications of the verbal overshadowing effect are fully dependent on whether it arises because of reduced discriminability or because of a conservative response bias (Mickes & Wixted, 2015).

Does providing a verbal description reduce discriminability?

The "verbal overshadowing effect" refers to impaired recognition memory performance. Interpreted in terms of signal detection theory, impaired recognition performance refers to reduced discriminability. Thus, from that perspective, a true verbal overshadowing effect is properly defined as a reduction in discriminability – that is, a reduction in the ability to discriminate innocent from guilty suspects – as a consequence of describing the perpetrator (Mickes, 2016). To measure discriminability, both the correct ID rate and false ID rate must be taken into account.

To measure discriminability, the most accurate approach is to measure not just one correct and false ID rate per condition (e.g., verbal description vs. control) but to measure the full range of correct and false ID rates that can be achieved in each condition across different levels of response bias. The entire family of achievable correct and false ID rates for a given condition is known as the receiver operating characteristic (ROC). ROC analysis is most easily performed by plotting correct vs. false identification rates across different levels of confidence. The ensuing ROC curves are constructed for both conditions and the area under the curve (AUC) for each condition is measured and statistically compared (for descriptions of how to conduct ROC analysis of lineup data, see Gronlund, Wixted & Mickes, 2014; Mickes, Flowe & Wixted, 2012). The larger the AUC, the better the discriminability. Evidence of a true verbal overshadowing effect would consist of a smaller AUC when a verbal description is provided compared to when a verbal description was not provided.

Does providing a verbal description affect reliability?

If verbal overshadowing does in fact reduce discriminability, it seems natural to suppose that it reduces the reliability of a suspect ID. However, whether or not providing a verbal description affects discriminability is a different question than whether or not providing a verbal description affects the *reliability* of a suspect identification from a lineup. Whether discriminability is low or high, an experimental manipulation that induces conservative responding will yield relatively high reliability (i.e., identifications will tend to be accurate), whereas a

manipulation that induces liberal responding will yield relatively low reliability (i.e., identifications will tend to be less accurate). Reliability can be measured in several different ways, including calibration analysis and confidence-accuracy characteristic (CAC) analysis (Mickes, 2015). Calibration analysis has the potential to underestimate reliability because the relevant equation includes filler identifications (for a comprehensive explanation of the differences between calibration and CAC analysis, see Wixted, Read, & Lindsay, 2016). CAC analysis, on the other hand, involves only suspect (guilty and innocent) ID accuracy as a function of confidence. This is the measure that is of most relevance to the legal system, which is interested in knowing the probability that a suspect who has been identified is actually guilty. When the base rates of target-present and target-absent lineups are equal (as is typically true of lab studies), this measure is given by:

$$PPV = \frac{S_g}{S_g + S_i}$$

where PPV is positive predictive value, S_g is the number of correct identifications, and S_i refers to the number of estimated innocent suspect identifications[1]. PPV is the probability that a suspect who was identified by a witness is in fact guilty, and it is computed separately for every level of confidence. For example, for participants who identify a suspect with high confidence, the PPV_{high} is given by:

[1] If there is no designated innocent suspect, the false suspect identification rate is estimated by dividing filler IDs from target-absent lineups by the number of lineup members.

$$PPV_{(high)} = \frac{S_{g(high)}}{S_{g(high)} + S_{i(high)}}$$

where $S_{g(high)}$ is the number of correct suspect IDs made with high confidence and $S_{i(high)}$ is the number of innocent suspect identifications made with high confidence. If the PPV for identifications made with high confidence were higher when no verbal description was provided, then reliability would be higher in that condition. It is possible, however, that even if discriminability is lower when a verbal description is provided, reliability could be higher (Mickes, 2016). This could happen if, for example, verbal descriptions reduced discriminability while at the same time induced very conservative responding.

Investigating the Effect of Verbal Descriptions on Discriminability and Reliability

To test the effect of verbal descriptions on discriminability and reliability, we directly replicated Schooler and Engstler-Schooler (1990) in four experiments. Following suit of the replication studies (Alogna et al., 2014), we replicated Experiments 1 and 4 of the original paper, with one critical difference – the inclusion of target-absent lineups. The original and replication studies tested memory on 8-person simultaneous lineups. To be able to use the same stimuli and include target-absent lineups, the lineup size was reduced to 6-person simultaneous lineups so that the perpetrator could be replaced with a filler for target-absent lineups. In our Experiment 1, the experimental manipulation (verbal description vs. control task) took place immediately after the study phase, and in

our Experiment 2, the experimental manipulation took place 20 minutes after the study phase (Figure 3.1).

In two additional experiments, we tested the effect of verbal descriptions on showups. Showups involve the presentation of only one person (the suspect) on the recognition test. Though showups are believed to be highly suggestive in nature (Goodsell, Wetmore, Neuschatz, & Gronlund, 2013; Steblay, Dysart, Fulero, & Lindsay, 2003) and have been found to yield lower discriminability than lineups (Wetmore et al., 2015; Mickes, 2015), showups will continue to be widely used by the police because they can be administered soon after a crime has been committed. As in Experiments 1 and 2, Experiments 3 and 4 retained the same procedural order as the original and replication experiments (see Figure 3.1). For these experiments, discriminability and reliability were again measured with ROC and CAC analysis, respectively. Finally, in a fifth experiment, we conducted a content analysis in an effort to determine why verbal descriptions have the effect they do on recognition memory performance.

Experiment 1

Method

Participants

Undergraduate students (N = 780) at the University of California, San Diego (UCSD) participated online for course credit. Sample size (for Experiments 1 and 2) was based on a power calculation that aimed to achieve 80% power (using results from an earlier lineup study, Mickes, Flowe, & Wixted, 2012, to

estimate the effect size). Participants (n = 63) reported that they previously viewed the video and were therefore not included in the analyses. Of the remaining (n = 717; 472 female, 239 male, and 6 did not specify), the average age = 20.5 years (sd = 2.55). Participants were randomly assigned to the control condition or the verbal condition and were tested on a target-absent lineup ($n_{control}$ = 188; n_{verbal} = 168) or a target-present lineup ($n_{control}$ = 171; n_{verbal} = 190) based on random assignment. The UCSD Institutional Review Board approved all of the experiments.

Materials

The stimuli included the 44 second video of the mock bank robbery and the eight photos (one of the perpetrator and seven fillers) used in original experiments (Schooler & Engstler-Schooler, 1990). The test phase included 6-person lineups with the images arranged in a 2x3 array. Images of the target and fillers from the original experiments were used for target-present and target-absent lineups. Target-present lineups were constructed using five of the seven fillers' images (that were randomly selected for each participant), and the photo of the perpetrator (and all of the images were randomly arranged for each participant). Target-absent lineups were constructed using six of the seven fillers' images (that were randomly selected and randomly arranged for each participant). The distractor task was an online crossword puzzle similar to the puzzle used in the original experiments (Schooler & Engstler-Schooler, 1990).

Procedure

The experiment was conducted online. Participants watched the video, typed as many countries and capitals as possible within 5-minutes (control condition) or provided a description of the perpetrator from the video for 5-minutes (verbal condition), and engaged in the 20-minute distractor task (see Figure 3.1A). The same instructions listed in the approved final protocol for the Alogna et al. (2014) study were used for the 5-minute writing task for participants in both conditions. Those in verbal conditions were given the following instructions from Alogna et al.: “Please describe the appearance of the bank robber in as much detail as possible. It is important that you attempt to describe all of his different facial features. Please write down everything that you can think of regarding the bank robber’s appearance. It is important that you try to describe him for the full 5 minutes” (pp. 559-560). After a 20-minute distractor task, memory for the perpetrator was tested on a lineup where participants were asked to try to identify the perpetrator or choose the “not present” option and rate their confidence on a 7-point scale (1 = guessing; 7 = certain).

Results and Discussion

Table 3.1 shows the frequency counts for each response type for target-present and target-absent lineups by levels of confidence. The average correct ID rate was higher for the verbal condition (0.52) than the control condition (0.49). False ID rates were estimated by dividing the number of filler identifications by the number of lineup members. The estimated false ID rate was lower for the verbal condition (0.09) than the control condition (0.11). Thus,

discriminability was, if anything, higher in the verbal group. We conducted ROC analysis to compare the locus of correct and false ID rates of both groups for a more complete assessment. The ROC curves in Figure 3.2A show higher discriminability for the verbal group. However, using a false ID rate cutoff of .458,[2] partial area under the curve (*pAUC*) analyses revealed that the difference between the verbal group (0.183) and the control group (0.153) was not significant, $D = 1.20$, $p = .232$. In other words, neither a verbal overshadowing effect nor its opposite was observed.

The curves in Figure 3.2A were generated from fitting a basic equal variance signal detection-based model to the ROC data from the control and the verbal conditions. In the model, memory strengths are distributed according to two Gaussian distributions, one representing fillers (which includes an innocent suspect) and one representing targets. This model assumes that to create target-absent lineups, six random draws are made from the filler distribution, and to create target-present lineups, five random draws are made from the filler distribution and one random draw is made from the target distribution. In the simplest version of this model, an identification is made if the memory strength of the most familiar face in the lineup exceeds a decision criterion.

---

[2] This value was selected because it is the false ID rate of the rightmost point on the ROC curve of the verbal condition, the more conservative of the two conditions (Gronlund, Wixted, & Mickes, 2014). Using the false ID value of the rightmost point on the ROC curve of the control condition as the cutoff does not change the conclusion ($p = .142$).

The filler distribution was set to $\mu_{lure} = 0$, $\sigma_{lure} = 1$, and the corresponding mean for the target distribution was estimated by fitting the model to the data. Correct and false identifications were binned into low (ratings of 1-3), medium (ratings of 4-5) and high (ratings of 6-7) levels of confidence and used for the different decision criteria. The model estimates d' and the three decision criteria. Fits were improved (but conclusions were not changed) by including another parameter, δ, which scales the estimated placements of the confidence criteria for target-present lineups relative to target-absent lineups (Seale-Carlisle & Mickes, 2016). Thus, there were a total of 10 parameters for both conditions, and each condition had 18 degrees of freedom: 3 degrees of freedom (filler identifications made with low, medium, or high confidence) for target-absent lineups (both conditions) and 6 degrees of freedom (filler identifications or suspect identifications made with low, medium, or high confidence) for target-present lineups (both conditions). The fits were performed simultaneously and had 8 degrees of freedom (18 degrees of freedom - 10 free parameters).

The parameters were adjusted until the difference between observed and predicted frequency counts was minimized using a chi-square goodness-of-fit statistic. The fit was good, $\chi^2(8) = 8.41$, $p = .394$. Constraining d' to be equal for verbal and control conditions also resulted in a good fit, $\chi^2(6) = 10.13$, $p = .119$. The full model fit and the constrained model fit did not differ significantly, $p = .190$, indicating that d' did not differ for the two conditions. Thus, as is typically

(but not necessarily) true, the results from the atheoretical *pAUC* analysis and the theoretical signal detection analysis agree.

The analyses presented above were concerned with discriminability. As noted earlier, reliability is a different issue. To measure the reliability of suspect IDs as a function of confidence in each condition, we conducted CAC analysis. Confidence ratings were binned in the same manner as for the model fits, and CAC was computed for identifications made with low, medium and high levels of confidence. The error bars represent standard error bars estimated using a bootstrap procedure (see Seale-Carlisle & Mickes, 2016). Figure 3.2B shows that the verbal group had higher reliability at each level of confidence, but none of the differences were significant.

Overall, neither discriminability nor reliability differed significantly between groups. In the original experiment in which the experimental manipulation occurred immediately after the study phase and 20 minutes prior to the identification procedure, the correct ID rate for the control group (0.71) was much higher than the verbal group (0.49) (Schooler & Engstler-Schooler, 1990). In the analogous replication experiment (Alogna et al., 2014), the average correct ID rate for the control group (0.55) was also higher than the verbal group (0.51), but the difference between the correct ID rates reported was considerably smaller. Three of the 31 participating laboratories found no difference between conditions, and 10 found a higher correct ID rate in the verbal condition than in the control

condition. Our results also revealed slightly higher correct ID rates for the verbal group (0.52) than the control group (0.49).

Confidence and accuracy were related for both groups. Identifications made with medium confidence were higher in accuracy than identifications made with low confidence, and lower in accuracy than identifications made with high confidence. Furthermore, CAC analysis revealed that identifications made with high confidence were comparably reliable for both groups.

Experiment 2

Experiment 2 was the same as Experiment 1 with the exception of swapping procedural order. In Experiment 2, the experimental manipulation took place 20 minutes after the study phase and immediately before the identification test (see Figure 3.1B). This was the order in which the greatest difference in correct identification rates resulted between groups in the replication experiments (Alogna et al., 2014). Also, as in Experiment 1, target-absent lineups were included to assess discriminability and reliability.

Method

Participants

Participants ($N = 780$) were recruited from Royal Holloway, University of London ($n = 138$), Amazon Mechanical Turk ($n = 245$), and SampleSize ($n = 397$). The participants ($n = 10$) who reported previously viewing the video were excluded from the analyses. Of the remaining ($n = 770$, 442 female; 318

male; 10 did not state), the average age = 27.9 years (*sd* = 11.1) Participants were randomly assigned to the control condition or the verbal condition and a target-absent lineup ($n_{control}$ = 179; n_{verbal} = 185) or a target-present lineup ($n_{control}$ = 196; n_{verbal} = 210). Royal Holloway, University of London Research Ethics Committee approved this study.

Materials and Procedure

The materials were the same used in Experiment 1. The procedure was the same with one exception: the experimental manipulation took place after the 20-minute distractor task and immediately before the test phase (see Figure 3.1B).

Results and Discussion

Table 3.1 shows the frequency counts for target-present and target-absent lineups by levels of confidence. The correct ID rate was lower in the verbal group (0.38) compared to the control group (0.62). The false ID rate was also lower in the verbal group (0.07) than the control group (0.09), which could mean that there is a difference in response bias, not discriminability, per se. We therefore conducted ROC analysis to measure discriminability independent of response bias. Figure 3.3A shows the ROC curves for both groups, and discriminability was lower in the verbal group than the control group. Using a false ID rate cutoff

of .584,[3] *pAUC* analysis revealed that the difference between the verbal (.096) and control (.155) groups was significant, $D = 3.06$, $p = .002$.

The ROC curves were generated from the same equal variance signal detection model as in Experiment 1. The ROC data were also fit, using the same parameters as in Experiment 1, and again, the fit was good, $\chi^2(8) = 6.12$, $p = .634$. However, when d' was constrained to be equal, the fit was worse, $\chi^2(6) = 17.53$, $p = .008$, and the fit was significantly different than when d' values were free to vary, $p < .001$. Thus, once again (and as expected), atheoretical *pAUC* analysis and theoretical signal detection analysis agree that discriminability was reduced in the verbal condition.

We next turned to the issue of reliability. The CAC curves, shown in Figure 3.3B, show slightly lower reliability across all three levels of confidence for the verbal group, but the differences were not significant at any of the levels of confidence. Moreover, confidence and accuracy are related (i.e., high confidence identifications are more accurate than low confidence identifications). High-confidence accuracy in the control condition was .95, whereas high-confidence accuracy in the verbal condition was .91. Thus, in both conditions, accuracy was high, and the small difference between them was not significant.

[3] Consistent with Experiment 1, this value was selected because it is the rightmost point on the ROC curve of the verbal condition (the more conservative condition). Using the false ID value of the rightmost point on the ROC curve of the control condition as the cutoff does not change the conclusion ($p = .002$).

In the current experiment, the correct ID rate was lower in the verbal condition compared to the control condition (0.38 vs. 0.62, respectively). This pattern was consistent with the original experiments (Schooler & Engstler-Schooler, 1990) and replication experiments (Alogna et al., 2014) when the verbal description task was delayed for 20 minutes after encoding. In the former, the correct ID rate was lower for the verbal condition (0.39) vs. the control condition (0.64). Similarly, in the latter, the average correct ID rate for the verbal condition (0.38) was lower than that of the control condition (0.54). Although those results are ambiguous as to whether they reflect either reduced discriminability or more conservative responding (or both), the ROC results reported here revealed significantly lower discriminability in the verbal condition (Figure 3.3A).

Despite the fact that discriminability was lower in the verbal condition, reliability was not significantly different (similar to the findings in Experiment 1) between the two conditions. Furthermore, high-confidence identifications were much more accurate than low-confidence identifications in both conditions. Thus, with regard to assessing the probative value of an ID, knowing confidence is far more informative than knowing whether or not the suspect's face was verbally described (despite the large verbal overshadowing effect). Next, we extended the replication further by testing the effect of verbal descriptions on discriminability and reliability when memory is tested using showups.

Experiment 3

In Experiment 3 we sought to replicate the pattern of results from Experiment 1. Thus, the procedure was held constant except that participants were tested on either a target-present or target-absent showup (i.e., the guilty suspect or innocent suspect, respectively). Again, we measured discriminability with ROC analysis and reliability with CAC analysis.

Method

Participants

UCSD undergraduate students participated online for course credit (N = 1,197; 410 male, 773 female, 14 unspecified; average age = 20.2 years, sd = 2.7). There are no earlier showup studies (i.e., showup vs. showup studies) to inform a power analysis, so sample size (for Experiments 3 and 4) was increased to 1,100 and we stopped data collection when the term ended. Participants were randomly assigned to the control condition or the verbal condition. Participants were also randomly assigned to a target-absent showup (control n = 300; verbal n = 293) or a target-present showup (control n = 328; verbal n = 276).

Materials

The materials were the same as those in Experiment 1 and 2, except an online game of Tetris was played instead of a crossword puzzle as the distractor task. Target-present showups were constructed by using the target photo, and target-absent showups were constructed by randomly selecting one of the seven filler photos.

Procedure

Procedural order was the same as in Experiment 1 (see Figure 3.1A). The only differences were that showups replaced lineups and participants rated their confidence on a 0-100% scale (0 = guessing; 100% = certain).

Results and Discussion

Table 3.1 shows the frequency counts for each response type for target-present and target-absent lineups by levels of confidence. The correct ID rate was higher for the control condition (0.65) than for the verbal condition (0.57). Likewise, the false ID rate was higher for the control condition (0.29) than for the verbal condition (0.18). Thus, on the surface, these results indicate that, at a minimum, verbal descriptions induced more conservative responding. Figure 3.4A shows that the two conditions yielded ROC curves that are not noticeably different, suggesting that verbal descriptions did not affect discriminability. The statistical comparison of the *AUC* values between the verbal (0.756) and control (0.735) conditions confirm this impression, $D = 0.71$, $p = .481$.

Next, correct and false identifications were binned into low (0-60%), medium (70-80%), and high (90-100%) confidence ratings to perform signal detection model fits to the ROC data and to conduct CAC analysis. The ROC curves in Figure 3.4A were generated by fitting the ROC data using the same equal variance signal detection model described previously with one less parameter (because there are no filler identifications with showups). The fit was good, $\chi^2(6) = 6.76$, $p = .344$. Constraining d' to be equal did not significantly

worsen the fit, $p = .663$.

The CAC curves (using the same confidence binning as for the model fits) in Figure 3.4B show no significant reliability differences between condition across the levels of confidence. Once again, confidence is predictive of accuracy, but the relationship for the verbal condition does not continue to increase from medium to high confidence. Also, high confidence identifications are noticeably lower in accuracy compared with what we observed for lineups (averaged across conditions, high-confidence showup accuracy = 0.80).

Experiment 4

In Experiment 4 we sought to replicate the pattern of results from Experiment 2, and like in Experiment 3, memory was tested on a target-absent or target-present showup. Again, we measured discriminability with ROC analysis and reliability with CAC analysis.

Method

Participants

UCSD undergraduate students participated online for course credit (N = 1,196; 364 male, 822 female, 10 unspecified; average age = 20.3 years, sd = 2.3). Participants were randomly assigned to the control condition or the verbal condition. Memory was tested on a target-absent showup ($n_{control} = 302$; $n_{verbal} = 322$) or a target-present showup ($n_{control} = 311$; $n_{verbal} = 261$).

Materials

All materials were the same as in Experiment 3.

Procedure

The procedure was the same as Experiment 3 with exception of the order of the distractor task and the experimental manipulation (the same order as Experiment 2; see Figure 3.1B). The writing task took place after the 20-minute distractor task and immediately before the test phase.

Results and Discussion

Table 3.1 shows the frequency counts for each response type for target-present and target-absent lineups by levels of confidence. Similar to the results in Experiment 2 (and the analogous replication study), the correct ID rate was lower in the verbal condition (0.43) than in the control condition (0.68). Also as in Experiment 2, the false ID rate was lower in the verbal condition (0.17) than the control condition (0.25). Again, these results are consistent with the idea that, at a minimum, providing verbal descriptions induced more conservative responding. To measure discriminability, ROC analysis was conducted. The ROC curves, as shown in Figure 3.5A, and AUC analysis reveal that discriminability is lower for the verbal condition (0.70) than the control condition (0.77), and that difference is significant, $D = 2.36$, p = .018. The curves in Figure 3.5A were generated by fitting the ROC data using the same equal variance signal detection model described in Experiment 3. Again, the fit was good, $\chi^2(6) = 3.28$, $p = .778$, and constraining d' to be equal worsened the fit to a marginally significant degree, $p = .052$. Thus, a verbal overshadowing effect is evident whether an atheoretical measure (AUC) or a theoretical measure (d') is used to interpret the results.

Despite the difference in discriminability, the CAC curves in Figure 3.5B again reveal no significant differences between conditions in reliability at all levels of confidence. However, there is a trend towards lower accuracy in the verbal conditions for IDs made with low or medium confidence. As in the other experiments, identifications made with high confidence are higher in accuracy than identifications made with medium and low confidence. High-confidence accuracy was 0.87 in both conditions. The results shown in Figure 3.5A and 3.5B illustrate a key point: a reduction in discriminability does not automatically translate into reduced reliability of IDs made with high confidence. As noted earlier, knowing the effect of a variable on discriminability does not automatically reveal the effect of that same variable on the reliability of an ID.

Experiment 5

In Experiments 1 and 3, participants in the verbal condition provided descriptions immediately after encoding, and discriminability did not differ from the control condition, regardless of whether memory was tested using a lineup or a showup. However, when verbal descriptions were provided after a delay (as in Experiments 2 and 4), discriminability was impaired for both procedures. What accounts for the difference in discriminability depending on whether or not the description is delayed?

The diagnostic feature-detection hypothesis may provide insight into this difference (Wixted & Mickes, 2014). The hypothesis was initially proposed to account for the discriminability advantage that simultaneous lineup presentations

have over procedures that involve showing an individual in isolation (as with sequential lineups or showups). By seeing the lineup members together, it is readily apparent to the witness that there are facial features shared across lineup members that should be discounted because they are not diagnostic of guilt. For example, if the perpetrator were a young, White male, then attaching weight to those features would not be helpful and would instead serve to impair discriminability because all of the lineup members would be young, White males. Having the faces presented simultaneously allows eyewitnesses to immediately detect and discount non-diagnostic features and to instead attach more weight on features that are not shared and are thus more diagnostic. This discrimination-enhancing strategy is less likely to be used when lineup members are presented individually because, under those conditions, it is harder to detect (and then discount) the common, non-diagnostic facial features.

The same concept may help to explain why verbal descriptions only impair discriminability when they are made after a delay. More specifically, participants may use more diagnostic feature descriptions immediately after encoding the perpetrator's face than they do after a delay. After a delay, by contrast, some forgetting will undoubtedly occur, and the description may become more general, perhaps becoming more likely to correspond to the common features that match everyone in the subsequently presented lineup. In that case, the participants may have a tendency to rely on the description they just gave when trying to identify the face of the perpetrator. To the extent that they rely on the general (common)

facial features mentioned in the verbal description, discriminability would be impaired.

To assess whether or not the diagnostic feature-detection hypothesis can help to account for the differences in discriminability when verbal descriptions are delayed, we first conducted a content analysis of the verbal descriptions provided in Experiments 1 through 4. We then conducted an experiment to test our theory.

Content Analysis

To conduct content analysis, 20 words were identified based on the appearance of the eight images of the perpetrator and fillers. Ten words were selected that were judged by the experimenters to be useful in differentiating the perpetrator from fillers (diagnostic-feature words), and 10 words were selected that were also judged by the experimenters to be less useful in differentiating the perpetrator from fillers (non-diagnostic-feature words). The latter words could have been used when selecting the fillers (e.g., White, male, attributes that related to hair color and stature). The diagnostic-feature words were descriptors that were not shared by all of the lineup members (see Appendix). The non-diagnostic-feature words were descriptors that were shared by all of the lineup members (see Appendix). The diagnostic-feature and non-diagnostic feature words were counted from descriptions provided by participants in the verbal condition in Experiments 1 and 3 (immediate descriptions) and compared with those descriptions in Experiments 2 and 4 (delayed descriptions).

Significantly more diagnostic-feature words were used when verbal descriptions were provided immediately after encoding (Experiments 1 and 3) compared to when verbal descriptions were provided 20 minutes after encoding (Experiments 2 and 4), $t(1945) = 4.75$, $p < .001$, Cohen's $d = 0.22$. However, there was no significant difference between the number of non-diagnostic feature words, $t(1945) = 1.28$, $p = .201$. A 2 × 2 analysis of variance revealed a significant interaction between type of feature (diagnostic vs. non-diagnostic) and time of verbal description (immediate vs. delayed), $F(1, 3890) = 15.96$, $p < .001$, Cohen's $d = 0.06$. These results provide evidence for the diagnostic feature-detection hypothesis.

In light of these findings, we conducted an experiment to test whether more diagnostic words were used when the verbal descriptions were provided immediately after encoding compared to after a delay. The participants in this experiment did not watch a video of the perpetrator but instead read either the descriptions that were written immediately after encoding or after a delay. They were then tasked with trying to identify the perpetrator from a lineup based on description only (i.e., they did not view the video). If the immediate descriptions contain more diagnostically useful information, then participants provided with those descriptions should be better able to identify the perpetrator from the lineup compared to the participants provided with the delayed descriptions.

Method

Participants

UCSD undergraduate participants took part in exchange for course credit (N = 128; 44 male, 81 female, 3 unspecified; mean age = 20.3 years, sd = 2.3). There are no previous studies to inform a power analysis, so we selected a sample size of 100 and stopped collecting data at the end of the term. Participants were randomly selected to read descriptions from Experiment 3 (n = 63) or Experiment 4 (n = 65). None had participated in the previous experiments.

Materials

The materials were the descriptions written by the participants in the verbal condition in Experiment 3 (written immediately after encoding; Figure 3/1A) and Experiment 4 (written after a 20-minute delay; Figure 3.1B), which were 569 and 583 descriptions, respectively.

Procedure

For each participant, one description was randomly selected from the pool of descriptions. Participants read the description and were immediately presented with an 8-person simultaneous target-present lineup (using the seven fillers and the perpetrator described in the previous experiments). The images were arranged in random order for each participant. Based on the description they read, participants attempted to identify the person they thought had committed the crime with no option to reject the lineup (i.e., no “not present” option).

Results and Discussion

Participants who read the descriptions that were written immediately after encoding were significantly more likely to correctly identify the perpetrator (M = 0.14, 95% CI = [.08, .25]) than participants who had read descriptions written after the delay (M = 0.03, 95% CI = [.01, .11]), z = 2.26, p = .024. Note that selecting a lineup member randomly from a perfectly fair 8-person lineup would result in a correct ID rate of 0.13. However, no lineup is perfectly fair. The low perpetrator selection rate in the control condition could either mean that the lineup was inherently biased towards one or more of the fillers (away from the perpetrator) or that the descriptions written after a delay had the effect of biasing selections towards one or more of the fillers (perhaps because they matched a more generic description than the perpetrator did).

In agreement with the diagnostic feature-detection hypothesis, more diagnostic-feature words were used in the descriptions when those descriptions were provided straightaway. Also consistent with the diagnostic feature-detection hypothesis, participants were able to identify the perpetrator more often if they read the description that was written by participants who provided the description immediately after encoding versus after a delay. Those descriptions therefore must have been more informative (i.e., more diagnostic). If participants in a verbal overshadowing experiment rely to some extent on their own descriptions when attempting to identify the perpetrator from the lineup, the prediction would be that discriminability should be impaired when descriptions are delayed (an effect that was observed in Experiments 2 and 4).

General Discussion

In a series of experiments we investigated the effects of verbal descriptions on discriminability and reliability on lineups and showups. The correct identification findings replicated the original verbal overshadowing Experiments 1 and 4 of Schooler and Engstler-Schooler (1990) and the replication efforts (Alogna et al., 2014). However, conclusions about memory performance based only on correct ID rates are tenuous. We therefore extended those findings by including target-absent lineups to be able to assess discriminability and reliability.

Effects of Verbal Reports on Discriminability

In Experiments 1-4, responding was more conservative in the verbal description condition. The relative conservatism is seen in the ROC curves in Figures 3.5 where the rightmost point on the verbal ROC is shifted leftward relative the rightmost point on the control ROC. Thus, one effect of providing verbal descriptions is to induce more conservative responding, and this phenomenon could account for the lower correct ID rates found in the original (Schooler & Engstler-Schooler, 1990) and the replication studies (Alogna et al., 2014). Why might this be? Clare and Lewandowsky (2004) proposed the idea that the task of describing the perpetrator makes participants realize that the task is challenging and as a result induces more cautious responding when faced with

making a lineup decision[4]. Our findings are consistent with this idea. However, above and beyond the conservative shift in responding, the results of Experiments 2 and 4 (involving delayed verbal descriptions) showed that discriminability in the verbal condition was also impaired.

Why is discriminability impaired by providing a verbal description after a delay but not by providing a verbal description immediately? This puzzling difference in discriminability could be explained by the diagnostic feature-detection hypothesis (Wixted & Mickes, 2014). The hypothesis holds that discriminability will be better when eyewitnesses rely more on diagnostic features than less diagnostic features. We tested this account in two ways, by conducting a content analysis and an experiment. In both analyses, the diagnostic feature-detection hypothesis provided a coherent interpretation of the data. Participants provided less diagnostic descriptions after a delay (presumably due to forgetting of more specific diagnostic details) compared to when descriptions were made immediately after encoding the face of the perpetrator, and other participants provided with those descriptions (but who did not see the mock-crime video) were better able to identify the perpetrator using the more diagnostic descriptions that had been written immediately after encoding.

[4] While Clare and Lewandowsky found lower discriminability (as measured by *d'*) for participants in one of their verbal conditions (the Holistic condition) compared to a control condition in Experiment 1 (the experiment most analogous to Experiments 1 and 2 here), they focused on the differences in criterion shifts.

Why people use less diagnostic information after time passes is a question that remains to be answered. Two potential theories may provide insight: fuzzy-trace theory and dual-process theories of recognition memory. Fuzzy-trace theory (Brainerd & Reyna, 1990) predicts that descriptions given after a delay would be based on gist representations versus descriptions given immediately, which would be based more on verbatim representations. This shift occurs because verbatim representations are thought to fade more rapidly than gist-based representations (e.g., Reyna, 2012). Indeed, Schooler (1998) once broadly linked verbal overshadowing with fuzzy trace theory, and it may be time to revisit this connection with more focus on the differential time course of gist and verbatim traces. Similarly, dual process theories might predict that descriptions provided soon after encoding are based on recollection, whereas descriptions provided later are based more on familiarity (e.g., Wais, Wixted, Hopkins, & Squire, 2006, but see e.g., Fortin, Wright, & Eichenbaum, 2004). Determining the usefulness of these theories could be a target for future research efforts.

One possibility that cannot be ruled out by our findings is that participants who provided a description immediately after encoding may rely on their description less than participants who provided a description after a delay. Adding an additional 20-minute delay would be one way to assess this possibility. Another possibility that cannot be ruled out by our findings is that when the description is provided after a delay, participants rely less on diagnostic

information despite the fact that the memory is intact and more diagnostic information can be culled in ways that were not tested here. Relatedly, future research efforts could involve investigations of ways to induce eyewitnesses to generate diagnostic descriptions even after time passes.

Effects of Verbal Reports on Reliability

In Experiments 1 through 4, the reliability of suspect IDs was comparable between conditions. Thus, even when discriminability was lower in the verbal condition, as was the case in Experiments 2 and 4, reliability was not appreciably different. Furthermore, adding to the body of literature that confidence and accuracy are related (e.g., Juslin, Olsson, & Winman, 1996; Brewer & Wells, 2006; Palmer, Brewer, Weber, & Nagesh, 2013; Sauer, Brewer, Zweck, & Weber, 2010; Dodson & Dobolyi, 2016; Mickes, 2015; Wixted, Read, & Lindsay, 2016), the relationship was strong for both of the conditions in these experiments. That is, PPV for high confidence identifications was higher than PPV for medium confidence identifications, which was higher for low confidence identifications. This was true even when the effect of verbal overshadowing on discriminability was strong (Experiments 2 and 4).

A similar pattern (reduced discriminability without a concomitant reduction in reliability) has now been reported for manipulations such as retention interval (Palmer et al., 2013; Sauer et al., 2010; Wixted et al., 2016), same-vs.-cross race (Dodson & Dobolyi, 2016; Nguyen, Pezdek & Wixted, in press), and both exposure duration and divided attention (Palmer et al., 2013). In each case, the

manipulation in question had a strong effect on discriminability while having little to no effect on the reliability of an ID made with high confidence. Although fewer high-confidence IDs occur in the low-discriminability condition, when they do occur in that condition, they are typically as accurate (or nearly so) as high-confidence IDs in the high-discriminability condition.

Practical Implications

The implications of these results for the criminal justice system seem straightforward. The results from ROC and CAC analyses are of interest to different decision-makers with ROC analysis being important for policymakers, who decide whether and when to ask for a verbal description, and CAC analysis being important for judges and jurors, who have no control over police policy but ought to know how reliable an ID is likely to be (Mickes, 2015; Mickes, 2016).

Our ROC results suggest that, as they presumably already do, police should encourage reporting crimes immediately and then take down the description of the perpetrator as soon as possible. By doing so, the adverse effects of verbal descriptions on discriminability would be mitigated. Future research efforts should manipulate different timings of the verbal descriptions, including a more protracted time course (Mickes, 2016) so that evidence for the optimal time points could be determined.

The CAC results are a matter of importance for judges and jurors who make decisions about culpability (Mickes, 2015; Mickes, 2016). On this issue, the message of our research likely differs from what has thought to be true of the

effect of verbal overshadowing. More specifically, our results suggest that, regardless of whether a verbal description was provided, the reliability of an ID made from a lineup or a showup was comparable. Moreover, high-confidence IDs from a lineup were quite accurate in both conditions (greater than 90% correct), whereas high-confidence IDs made from a showup were less accurate in both conditions. The fact that identifications made with high confidence are associated with lower PPV when memory is tested on showups than lineups should signal to judges and jurors that those identifications may be less trustworthy and thus should be taken with caution.

Table 3.1. Frequencies of suspect IDs, filler IDs, and no IDs for target-absent and target-present lineups for all levels of confidence in the control and verbal conditions in Experiments 1-4.

	Control					Verbal				
	Target-present			Target-absent		Target-present			Target-absent	
Confidence	Suspect IDs	Filler IDs	No IDs	Filler IDs	No IDs	Suspect IDs	Filler IDs	No IDs	Filler IDs	No IDs
	Experiment 1									
1	3	2		5		1	0		2	
2	1	4		11		3	7		5	
3	10	6		22		5	10		13	
4	17	15	34	35	60	18	12	41	20	77
5	19	13		35		24	14		32	
6	21	9		14		31	5		14	
7	13	4		6		16	3		5	
	Experiment 2									
1	0	2		1		0	1		1	
2	3	3		4		2	5		5	
3	7	5		13		5	8		8	
4	21	10	41	29	86	12	10	87	16	108
5	47	10		33		29	13		29	
6	25	3		10		30	4		14	
7	19	0		3		2	2		4	

	Target-present		Target-absent		Target-present		Target-absent	
	Experiment 3							
	Suspect IDs	No IDs	Suspect IDs	No IDs	Suspect IDs	No IDs	Suspect IDs	No IDs
0%	0		1		0		1	
10%	0		0		1		0	
20%	1		0		0		1	
30%	3		0		2		1	
40%	6		3		7		2	
50%	15	114	7	214	9	120	5	240
60%	25		22		18		9	
70%	51		23		42		15	
80%	47		15		39		8	
90%	30		9		20		4	
100%	36		6		18		7	
	Experiment4							
0%	1		0		0		0	
10%	1		0		0		0	
20%	1		0		1		1	
30%	5		1		0		0	
40%	8		4		4		4	
50%	7	100	10	226	5	149	9	266
60%	23		11		22		11	
70%	52		19		25		11	
80%	54		21		19		14	
90%	36		4		22		5	
100%	23		6		14		1	

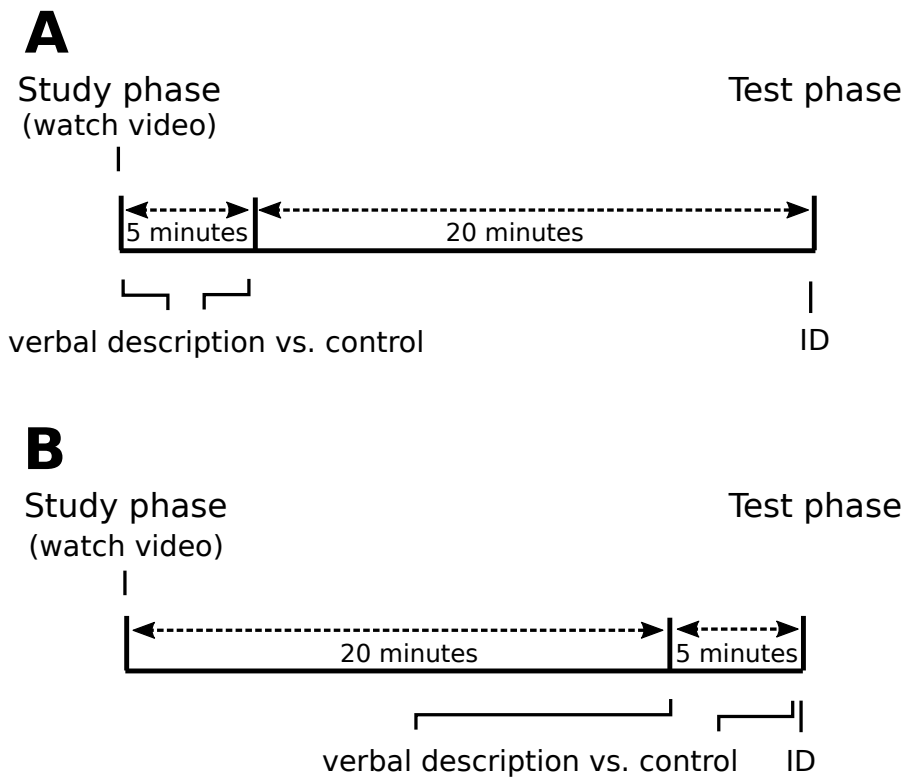


Figure 3.1. Procedural order of the original Experiments 4 (**A**) and 1 (**B**) (Schooler & Engstler-Schooler, 1990); Alogna et al. (2014) RRR1 (**A**) and RRR2 (**B**); and the current Experiments 1 and 3 (**A**) and Experiments 2 and 4 (**B**). (Diagram adapted from Mickes, 2016)

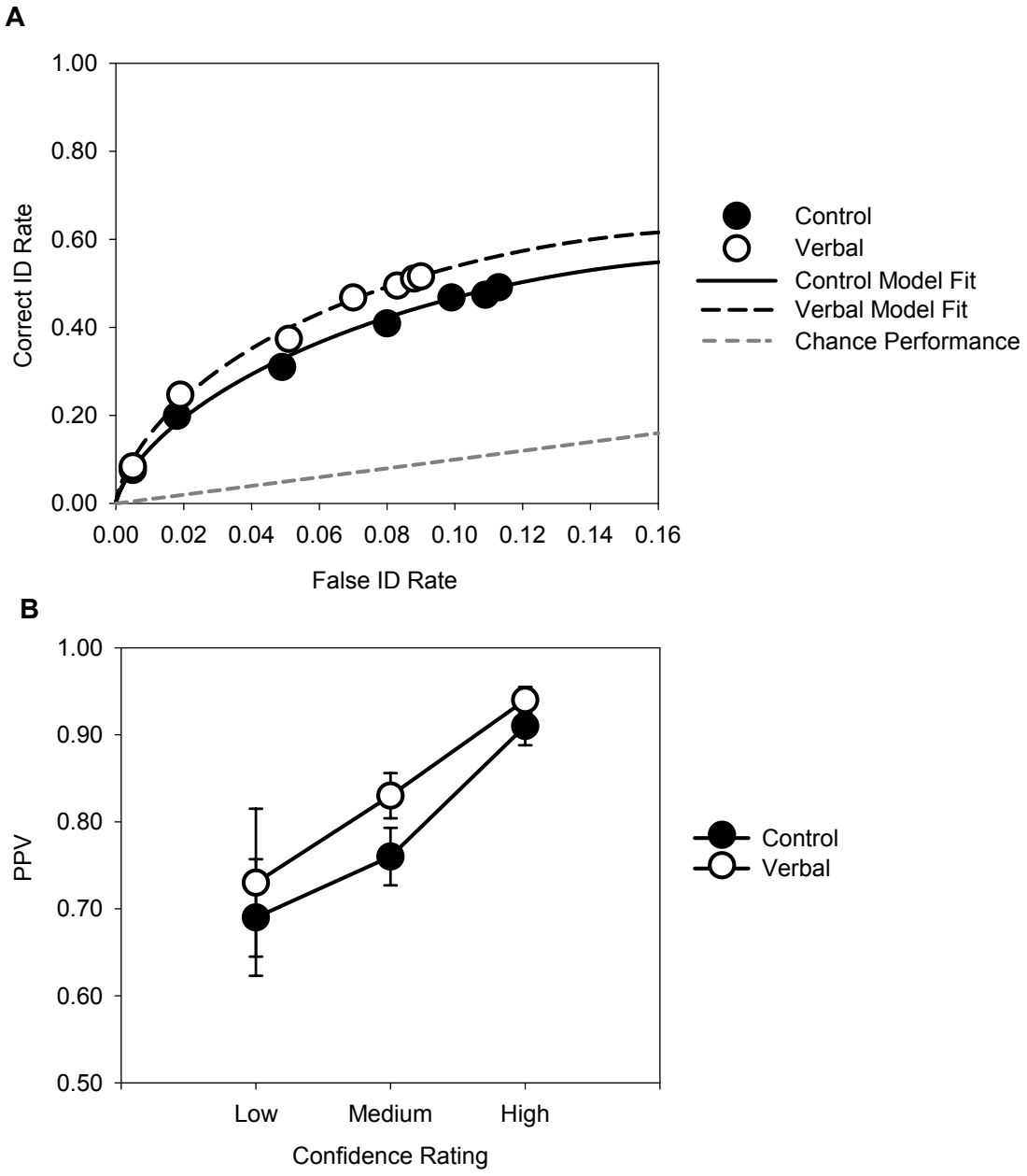


Figure 3.2. Receiver operating characteristic (ROC) and confidence-accuracy characteristic (CAC) plots for the verbal and control conditions in Experiment 1. **A)** ROC data and curves that represent the fit of the signal detection model. The grey dashed line represents the line of chance performance. **B)** CAC plot of positive predictive value (PPV) as a function of confidence. Bars represent standard error bars estimated using a bootstrap procedure.

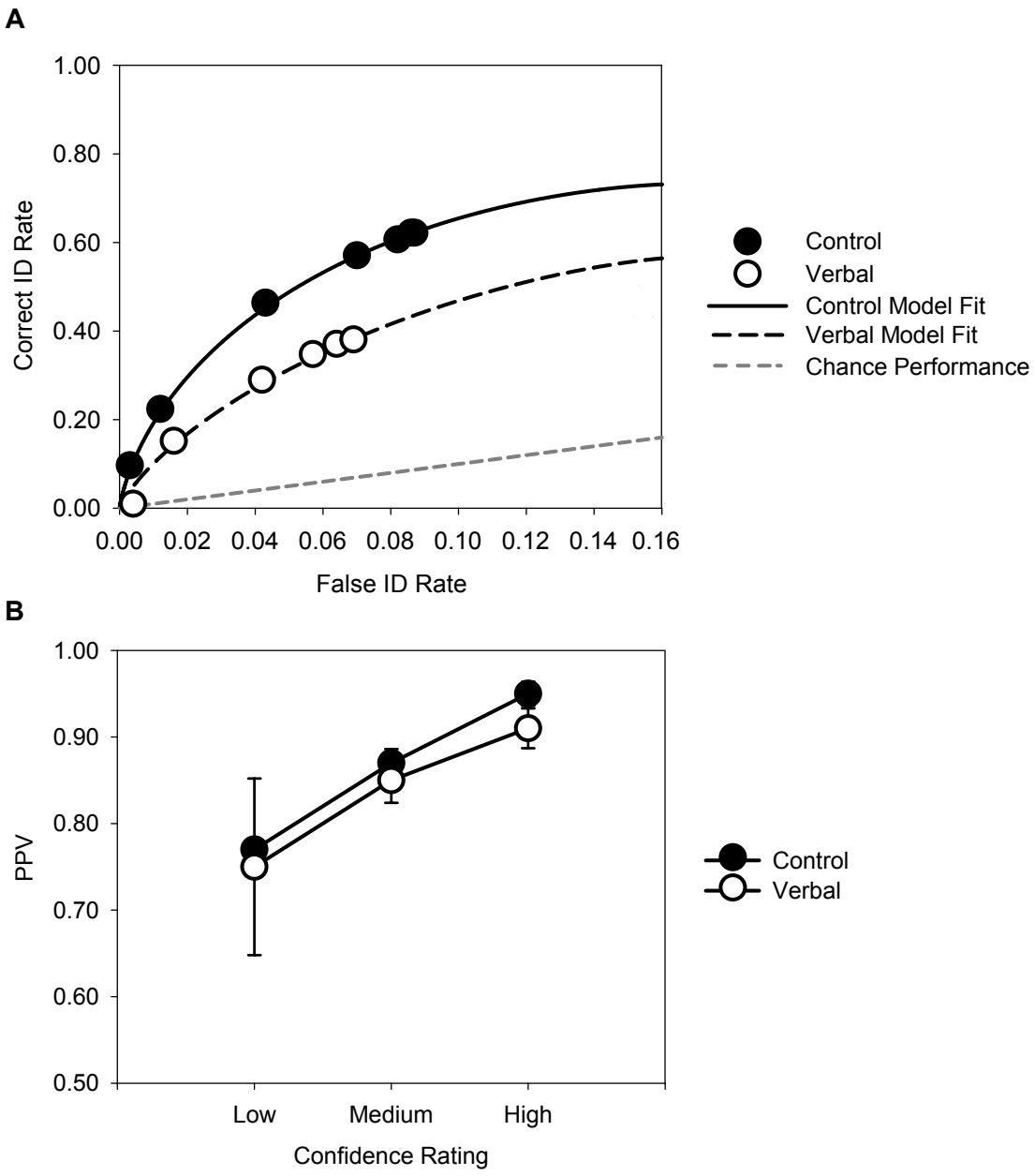


Figure 3.3. Receiver operating characteristic (ROC) and confidence-accuracy characteristic (CAC) plots for the verbal and control conditions in Experiment 2. **A)** ROC data and curves that represent the fit of the signal detection model. The grey dashed line represents the line of chance performance. **B)** CAC plot of positive predictive value (PPV) as a function of confidence. Bars represent standard error bars estimated using a bootstrap procedure.

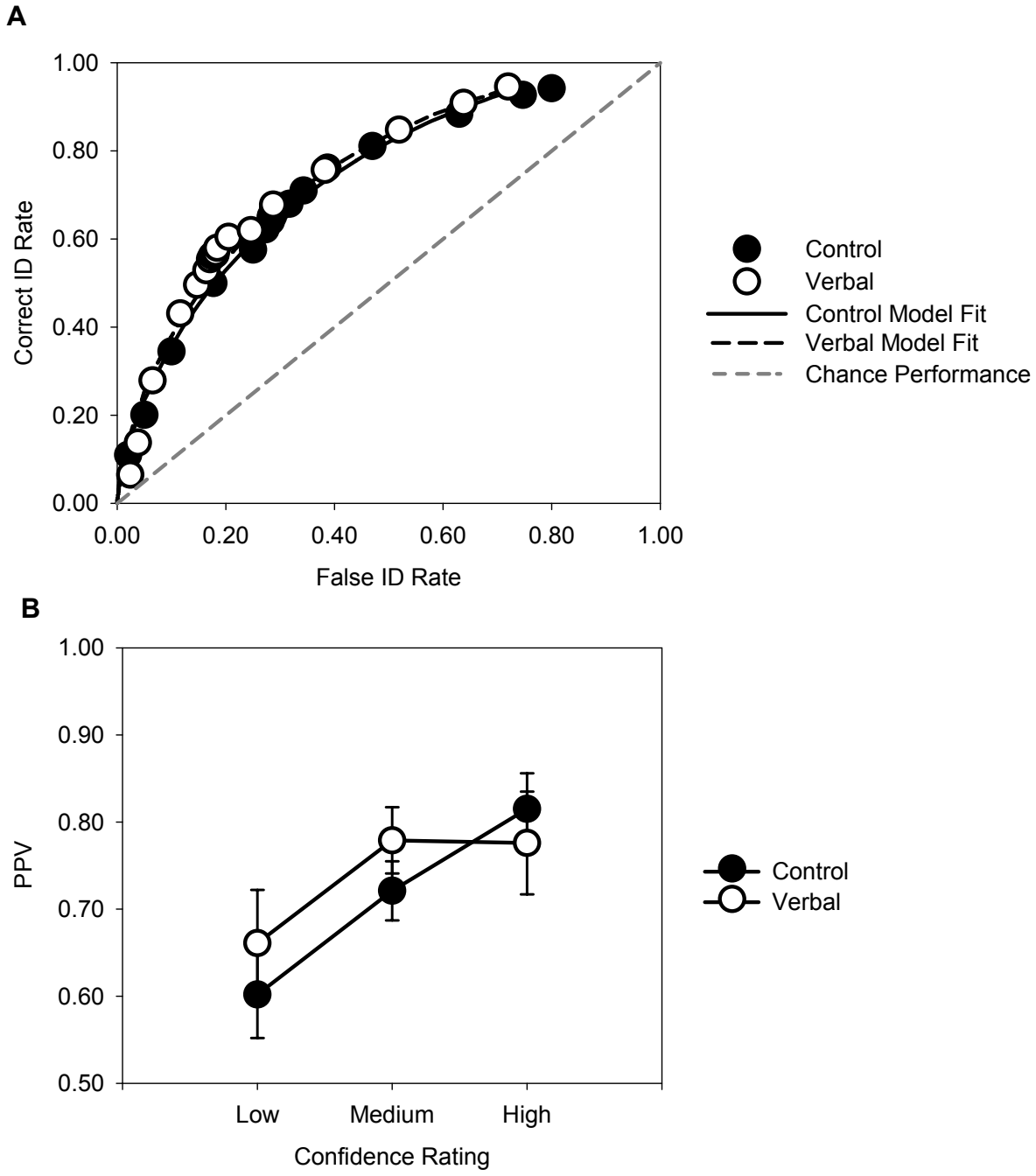


Figure 3.4. Receiver operating characteristic (ROC) and confidence-accuracy characteristic (CAC) plots for the verbal and control conditions in Experiment 3. **A)** ROC data and curves that represent the fit of the signal detection model. The grey dashed line represents the line of chance performance. **B)** CAC plot of positive predictive value (PPV) as a function of confidence. Bars represent standard error bars estimated using a bootstrap procedure.

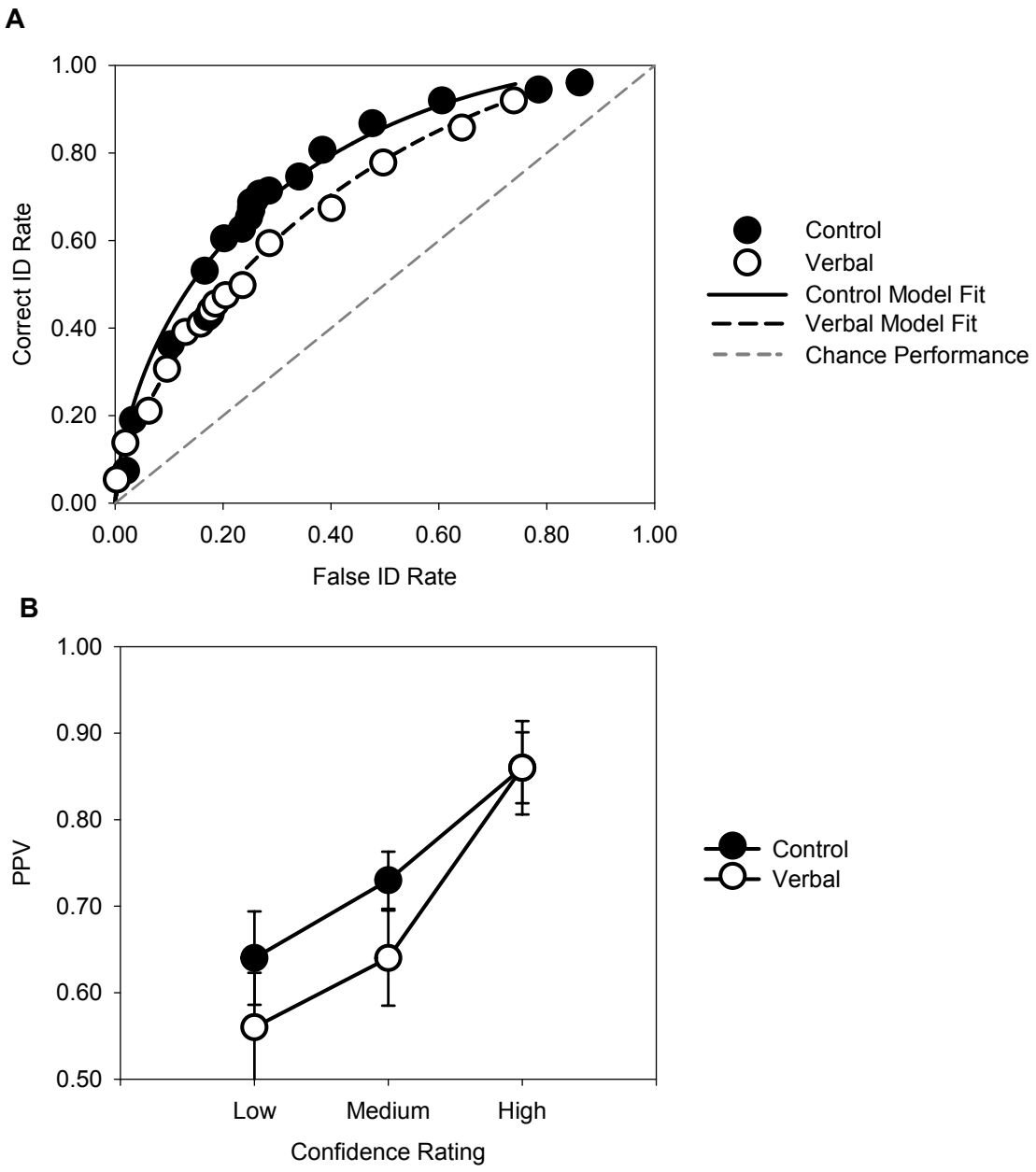


Figure 3.5. Receiver operating characteristic (ROC) and confidence-accuracy characteristic (CAC) plots for the verbal and control conditions in Experiment 4. **A)** ROC data and curves that represent the fit of the signal detection model. The grey dashed line represents the line of chance performance. **B)** CAC plot of positive predictive value (PPV) as a function of confidence. Bars represent standard error bars estimated using a bootstrap procedure.

Appendix

Diagnostic			Non-Diagnostic		
Words	Immediate	Delayed	Words	Immediate	Delayed
chin	76	65	white	491	517
jaw	76	67	male	330	323
cheek	127	89	age	266	295
brow	560	531	brown	449	422
forehead	51	38	black	586	526
eye	673	623	moustache	59	140
oval	33	15	dark	467	555
round	201	186	weight	30	39
wavy	116	86	build	71	94
point	79	53	height	159	161

Note. Different participants used different adjectives. For example, because “chin” (e.g., “pointy chin”) and “jaw” (e.g., “chiseled jaw”) were mentioned meant that there was something notable about them that was more diagnostic than ethnicity (“White”) and gender (“male”).

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., & Birt, A. R., . . .Zwaan, R.A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556-579.

Brainerd, C. J. & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review, 10,* 3-47.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relation in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12,* 11-30.

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 739-755.

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology, 30,* 113-125.

Fortin, N. J., Wright, S. P., & Eichenbaum, H. (2004). Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature, 9,* 188-191.

Goodsell, C. A., Wetmore, S. A., Neuschatz, J. S., & Gronlund, S. D. (2013). Showups vs. lineups: A review of two identification techniques. In B. Cutler (Ed.), *Reform of eyewitness identification procedures* (pp.45-64). Washington, DC: American Psychological Association.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science, 23,* 3-10.

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,* 1304-1316.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology, 15*, 603-616.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*(2), 93-102.

Mickes, L. (2016). The effects of verbal descriptions on eyewitness memory: Implications for the real-world. *Journal of Applied Research in Memory and Cognition, 5,* 270-276.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous vs. sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361-376.

Mickes, L. & Wixted, J. T. (2015). On the applied implications of the "Verbal Overshadowing Effect". *Perspectives on Psychological Science, 10,* 400-403.

Nguyen, T. B., Pezdek, K. & Wixted, J. T. (in press). Evidence for a Confidence-Accuracy Relationship in Memory for Same- and Cross-Race Faces. *Quarterly Journal of Experimental Psychology*.

Palmer, M., Brewer, N., Weber, N. & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55-71.

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision Making, 7,* 332-359.

Rotello, C. M., Heit, E., & Dube, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin and Review, 22,* 944-954.

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337–347.

Schooler, J. W. (1998). The distinctions of false and fuzzy memories. *Journal of Experimental Child Psychology*, 71(2), 130-143.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: some things are better left unsaid. *Cognitive Psychology, 22,* 36-71.

Seale-Carlisle, T. M. & Mickes, L. (2016). US lineups outperform UK lineups. *Royal Society Open Science.* DOI: 10.1098/rsos.160300

Smith, H. M. J., & Flowe, H. D. (2015). ROC analysis of the verbal overshadowing effect: Testing the effect of verbalisation on memory sensitivity. *Applied Cognitive Psychology, 29,* 159-168.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 27,* 523-540.

Wais, P. E., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2006). The hippocampus dupports both the recollection and the familiarity components of recognition memory. *Neuron, 49,* 459-466.

Wetmore, S., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition, 4,* 4-18.

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121,* 262-276.

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence-accuracy relationship. *Journal of Applied Research in Memory and Cognition, 5,* 192-203.

Chapter 3, in full, has been submitted for publication as Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. “The Effects of Verbal Descriptions on Performance in Lineups and Showups.” The dissertation author was the primary investigator and author of this manuscript.

CHAPTER 4

The Cross-Race Effect in Eyewitness Identification:

Reduced Discriminability Does Not Necessarily Imply Reduced Reliability

Abstract

The cross-race effect or own-race bias is a well-known finding in psychology wherein memory is worse for faces of a different race. This research has considerable implications for the legal system, and expert witnesses regularly testify that cross-race identifications are less trustworthy than same race-identifications. Lower overall memory (d′), however, is not the question of interest to the legal system. The legal system wants to know how likely a high-confidence identification is to be accurate. We replicate the standard cross-race effect and show that despite lower discriminability, high-confidence same-race and cross-race IDs were highly (and almost equally) reliable. In both cases, high-confidence accuracy exceeded 95% correct.

Since Munsterberg (1908) first wrote about psychology and crime, researchers have studied factors that affect the accuracy of eyewitness memory. An extensive body of research shows that eyewitness accounts can be influenced by both system and estimator variables (Lampinen, Neuschatz, & Cling, 2012; Wells, 1978; Wells et al., 2000). System variables are procedures that can be controlled by the criminal justice system, such as questioning of witnesses (Loftus, 1975; Loftus 1996), law enforcement-witness interactions (Clark, Marshall, & Rosenthal, 2009), and identification protocols (Clark, Brower, Rosenthal, Hicks, & Moreland, 2013; Douglass, Smith, & Fraser-Thill, 2005; Haw & Fisher, 2004; Wells et al., 1998). Estimator variables, by contrast, are not under the control of the criminal justice system because they are characteristics of the crime itself. Estimator variables include, for example, the presence of a weapon, witness stress, and the relationship between the race of the perpetrator and the race of the witness.

The same-vs.-cross-race estimator variable has been studied for nearly 50 years (Malpass & Kravitz, 1969; Meissner & Brigham, 2001; Sporer, 2001) and is the focus of this article. Previous research has consistently found a cross-race effect (or "own-race bias") in that people are better at recognizing members of their own race than members of other races (Brigham & Malpass, 1985; Chiroro, Tredoux, Radaelli, & Meissner, 2008; Malpass, 1974; Pezdek, O'Brien, & Wasson, 2012; Wright, Boyd, & Tredoux, 2003). More specifically, numerous studies have documented the fact that overall discriminability (e.g., as measured

by *d'*) is higher for same-race than cross-race faces, and various theoretical explanations have been advanced to explain that effect. For example, some research has shown that the cross-race effect is reduced (Brigham, Maass, Snyder, & Spaulding, 1982; Carroo, 1986; Chiroro & Valentine, 1995) or even reversed (Sangrigoli, Pallier, Argenti, Ventureyra, & de Schonen, 2005) with more exposure to faces from different races.

Research on the cross-race effect has been cited by expert witnesses in courts of law for the past several decades as a reason to question the trustworthiness of cross-race eyewitness identifications. Such testimony would appear to be supported not only by the research discussed above but also by expert consensus opinion. For example, in one survey of experts in eyewitness memory, the cross-race effect was described as follows: “eyewitnesses are more accurate when identifying members of their own race than members of other races” (Kassin, Tubb, Hosch, & Memon, 2001, p. 408). Ninety percent of those surveyed indicated that the evidence bearing on this statement was reliable, and over 70% indicated that they would be willing to testify about it in a court of law. However, this statement does not clearly differentiate between overall memory being worse for someone of a different race (e.g., as measured by *d'*) and the reliability of a cross-race identification (e.g., how accurate an ID made with high confidence is). Although the word "accuracy" is used to characterize both aspects of memory performance, they are not the same, and they are not equally important to the legal system. Here, we will use the word "accuracy" to refer to

the ability to distinguish between innocent and guilty suspects (typically measured by *d'*) and "reliability" to refer to the probability that an identification made with a particular level of confidence is correct (Mickes, 2015). For an estimator variable, the legal system is primarily concerned with the reliability of an identification that has occurred, not with the overall ability to distinguish between innocent and guilty suspects.

The eyewitness experts surveyed by Kassin, Tubb, Hosch, and Memon (2001) probably would not have considered the reliability of same- vs. cross-race identifications for different levels of confidence because 90% of them agreed that there was reliable evidence that an “eyewitness's confidence is not a good predictor of his or her identification accuracy” (p. 408). Recent evidence indicates that on an initial memory test using a properly administered lineup, confidence is an extremely good predictor of accuracy (Wixted et al., 2015; Wixted & Wells, 2017). Indeed, it is theoretically possible that an estimator variable that impairs overall memory accuracy has that effect by shifting high-accuracy high-confidence IDs to low-accuracy low-confidence IDs. Overall accuracy would be reduced (because of the increased proportion of low-accuracy low-confidence IDs) without necessarily changing the reliability of an ID made with a particular level of confidence. Measures that do not compute identification accuracy for each level of confidence and instead combine all identifications into a single number (e.g., *d'*) cannot differentiate between factors that affect overall memory performance and factors that affect confidence-specific reliability. Does the

cross-race factor affect only affect overall accuracy (as much prior research has shown to be true), or does it also affect the reliability of IDs made with a particular level of confidence? That question is the focus of this article.

Recent work suggests that the cross-race effect may be largely limited to overall accuracy, not to confidence-specific reliability. For example, in a reanalysis of several studies that investigated the cross-race effect using a list-memory procedure, Nguyen, Pezdek and Wixted (in press) found that in three prior studies from the Pezdek lab for which overall performance was greater than chance, *d'* was greater for same- than cross-race faces (i.e., the typical cross-race effect was found), but high-confidence same-race and cross-race identifications were nevertheless nearly equally reliable. Similarly, using a photo lineup paradigm, Dodson and Dobolyi (2016) also found the typical cross-race effect (i.e., significantly higher *d'* for same- vs. cross-race IDs), but reliability as we have defined it here – that is, proportion correct as a function of confidence – was very similar for both conditions. In both conditions, high-confidence IDs were far more accurate than low-confidence IDs. By contrast, within each level of confidence, the difference in accuracy for same-vs.-cross-race IDs was small. For example, for IDs made with the lowest level of confidence, same-race accuracy was approximately 15% correct, whereas cross-race accuracy was approximately 11% correct. For IDs made with the highest level of confidence, same-race accuracy was approximately 80% correct, whereas cross-race accuracy was approximately 77% correct. Thus, in this study, as in Nguyen et al.

in press (in press), confidence provided a great deal of information about the reliability of an ID, where the same-vs.-cross-race variable provided a much smaller amount of information about the reliability of an ID. It seems fair to say that this is the opposite of what has long been thought to be true about the reliability of eyewitness identification.

In their lineup experiment, Dodson and Dobolyi (2016) measured the effect of the same-vs.-cross-race variable on calibration accuracy for “choosers.” Choosers are participants who identify someone from a lineup, whether the suspect (innocent or guilty) or a filler. The accuracy of choosers is of most importance to the legal system because eyewitnesses who do not choose someone from a lineup do not imperil any of the lineup members. Choosers, by contrast, sometimes choose an innocent suspect, which can result in a wrongful conviction. Thus, the performance of choosers has long been analyzed separately from non-choosers. However, the argument has recently been made that choosers who pick fillers also do not imperil a member of the lineup, so the question of most interest is suspect ID accuracy (Mickes, 2015). That is, the measure of most interest is the accuracy of participants who choose the suspect (innocent or guilty).

The effect of the cross-race variable on the confidence-specific accuracy of suspect IDs has not been investigated, but we do so here. We report a new lineup experiment that investigates the effect of the cross-race variable on suspect ID accuracy for different levels of confidence, and we reanalyze the data

reported by Dodson and Dobolyi (2016) to determine what their lineup study indicates about the effect of same-vs.-cross-race on suspect ID accuracy.

Method

Participants

UCSD undergraduate students participated online for course credit (N = 1646; 464 male, 1172 female, 10 unspecified; average age = 20.5 years, sd = 2.3). At the end of the experiment, participants indicated their ethnicity. There were 912 Asians (268 male, 640 female, 4 unspecified; average age = 20.3 years, sd = 2.0) and 306 Caucasians (85 male, 219 female, 2 unspecified; average age = 21.0 years, sd = 2.4).

Materials

Faces of 36 Asians with neutral expressions and 36 Caucasians with neutral expressions were used. All faces had a white background. A web-based version of Tetris was used for the distractor task.

Procedure

The 36 Asian faces were randomly sorted into three groups of 12 faces. This random sorting was repeated three times, giving a total of nine different sets of 12 Asian faces. This same procedure was also used for the Caucasian faces.

For each participant, one set of Asian faces and one set of Caucasian faces were randomly selected. During the study phase one Asian face was randomly selected from the set and presented for 3 s, and one Caucasian face was randomly selected from the set and presented for 3 s. The presentation

order of the two faces was also randomized. Participants then had a 10-min distractor task, which was immediately followed by the test phase.

During the test phase, six of the Asian faces from the set were randomly presented in one simultaneous lineup along with a "Not Present" option. Six of the Caucasian faces from the set were randomly presented in another lineup along with a "Not Present" option. Randomly selecting six of the 12 faces gave a 50% probability of having a target-present lineup and a 50% probability of having a target-absent lineup. The order of the two lineups was also randomized. Participants could either select one of the faces or "Not Present" for each lineup. They then rated their confidence on a 0-100% scale (0 = guessing; 100% = certain).

Results

For the general question of whether or not memory is worse for cross-race than same-race identifications, receiver operating characteristic (ROC) curves were constructed. An ROC curve measures discriminability, just as the more typical *d′* measure does, but it does so without relying on theoretical assumptions. The cross-race condition included all identifications made to a cross-race face (i.e., Caucasian participants viewing an Asian face and Asian participants viewing a Caucasian face). The same-race condition included all identifications made to a same-race face (i.e., Caucasian participants viewing a Caucasian face and Asian participants viewing an Asian face). As can be observed in the ROC curves shown in Figure 1, discriminability was higher for

same-race identifications than for cross-race identifications. Using a false ID rate cutoff of .035,[5] *pAUC* analysis revealed that the difference between the same-race (0.022) and cross-race (0.018) groups was significant, $D = 3.02$, $p = .003$. Not surprisingly, the same result is obtained when we compute d' for the same-race (3.27) and cross-race (2.88) conditions. Thus, the oft-replicated cross-race effect was replicated again here. As noted earlier, however, that result does not, in and of itself, imply that the reliability of a cross-race ID differs from the reliability of a same-race ID. Figure 2 shows the cross-race and same-race effect for the Caucasian and Asian faces separately, which shows the same pattern as the combined plots. We plot the data by faces because this does not require any assumption that the face stimuli are perfectly equated in terms of memorability, whereas plotting the data by participant instead requires this assumption in order to be interpretable.

For the applied question of whether or not the same-vs.-cross-race variable affected suspect ID accuracy for different levels of confidence, we constructed CAC plots (see Mickes, 2015 for details). The results are presented in Figure 3. The first point to make about these results is that accuracy is fairly high even for low-confidence IDs. The usual pattern where accuracy becomes higher as confidence increases is also observed. At the highest level of

[5] This value was selected because it is the rightmost point on the ROC curve of the cross-race condition. Using the rightmost point on the ROC curve of the same-race condition does not change the conclusion ($p < .001$).

confidence, same- and cross-race identifications are virtually identical. Figure 4 shows the CAC plots for Caucasian and Asian faces separately, which shows the same general pattern as the combined plots.

The CAC plots in Figure 3 assume equal base rates of target-present and target-absent lineups. However, this number may be lower, which would lower the accuracy of positive identifications. Wixted, Mickes, Dunn, Clark, and Wells (2016) estimated the actual base rate of target-present lineups to be 35%. Figure 5 shows the CAC plots for same- and cross-race identifications with a 35% base rate (see Wixted and Wells, 2017 for details about calculations).

Reanalysis of Dodson and Dobolyi's (2016) Calibration Data. We contacted the authors of Dodson and Dobolyi (2016) about our interest in re-plotting their data, and they kindly supplied their results plotted in terms of suspect ID accuracy. Figure 6 shows their reanalyzed results. Although accuracy covers a greater range in their study than in ours, the conclusion is similar: high-confidence suspect IDs are highly – and almost identically – accurate for same- and cross-race suspect IDs, but a small difference emerges for IDs made with lower levels of confidence.

Discussion

Extensive research has examined the cross-race effect and has consistently shown that memory is worse for cross-race than same-race identifications. However, the question of most interest to the legal system – namely, are high-confidence cross-race identifications less reliable than high-

confidence same-race identifications? – has been almost entirely ignored. Here, we replicated the standard cross-race effect according to which overall memory is worse for cross-race identifications than same-race identifications. Nevertheless, our results show that high-confidence cross-race identifications are highly accurate and nearly as accurate as high-confidence same-race identifications. This finding is true not only in our new data set comparing Caucasian and Asian faces but also in our reanalysis of the Dodson and Dobolyi (2016) data set comparing Caucasian and Black faces. In both studies, high-confidence same-race and cross-race suspect IDs exceeded 97% correct.

For decades, eyewitness memory researchers concluded there was little relationship between eyewitness confidence and accuracy (see Wixted, Mickes, Clark, Gronlund, & Roediger, 2015). Wells and Murray (1984) concluded, "the eyewitness accuracy–confidence relationship is weak under good laboratory conditions and functionally useless in forensically representative settings" (p. 165). This belief likely lead eyewitness memory researchers to ignore the most forensically relevant question about cross-race identifications. If researchers thought level of confidence did not provide any useful information, considering differences in high-confidence accuracy between conditions would be futile. However, eyewitness memory researchers now know that eyewitness confidence provides important information about the likely accuracy of an identification (Wixted & Wells, 2017). This is clearly seen with cross-race identifications as well. That is, on an initial and properly administered lineup, knowing the

eyewitness's confidence level provides a lot of information about how likely the identification will be accurate, but knowing if the eyewitness is making a same-race or cross-race identification provides very little information about identification accuracy (particularly for IDs made with high confidence, which are of most importance for cases that result in prosecution).

The measure of overall recognition accuracy provides information about the theoretical question of which types of faces are better remembered but reveals little about the reliability of a high-confidence identification. To date, most research on the cross-race effect has examined the question of whether memory is worse for cross-race than same-race faces. This has resulted in numerous theories that explain the mechanism for reduced memory performance. These theories explain why discriminability (*d'*) is lower for cross-race than same-race identifications. This, however, is not the most critical concern for the judicial system. The judicial system is interested in knowing the accuracy of high-confidence cross-race identifications. In our research we replicate the standard cross-race effect where memory is worse for cross-race identification than for same-race identification. However, our study also shows that high-confidence cross-race identifications are nearly as accurate as same-race identifications.

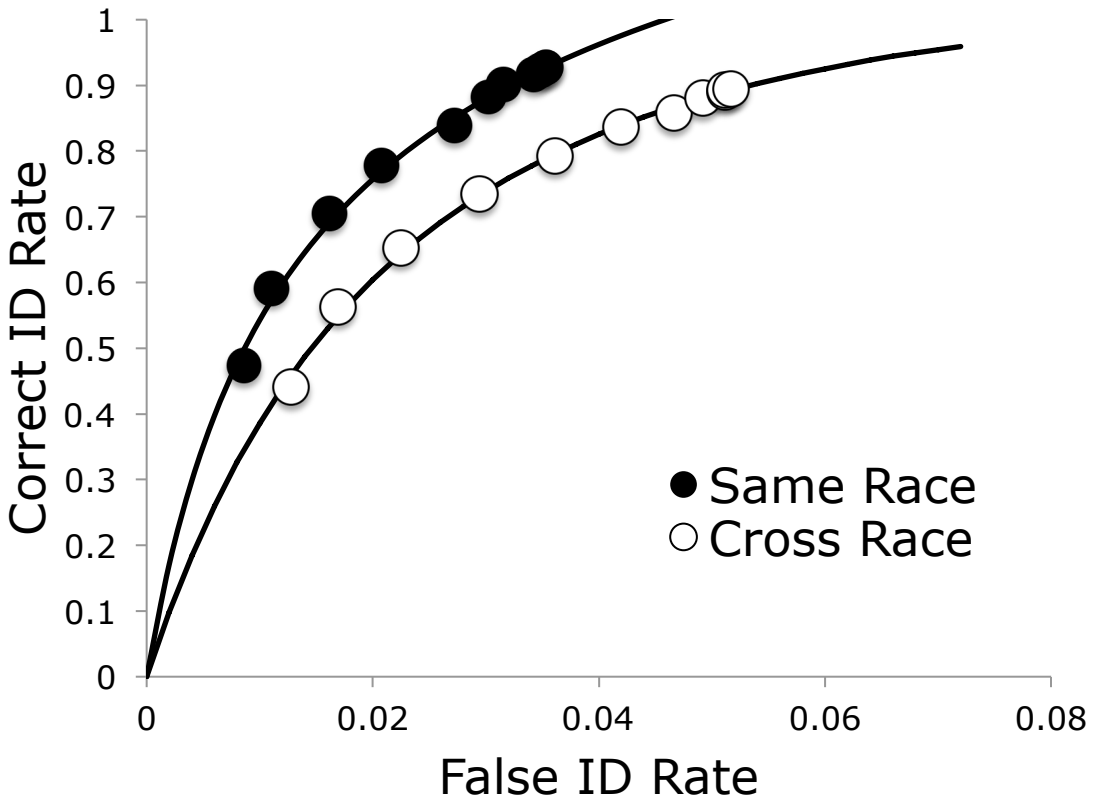
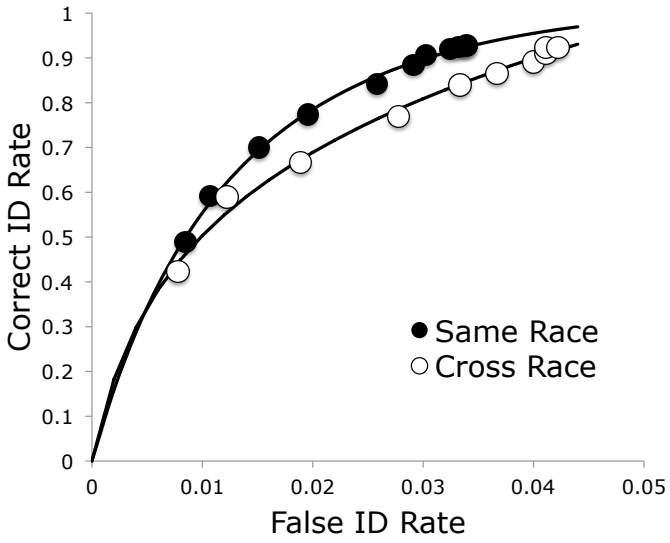


Figure 4.1. Receiver operating characteristic (ROC) plots for same-race and cross-race identifications.

A



B

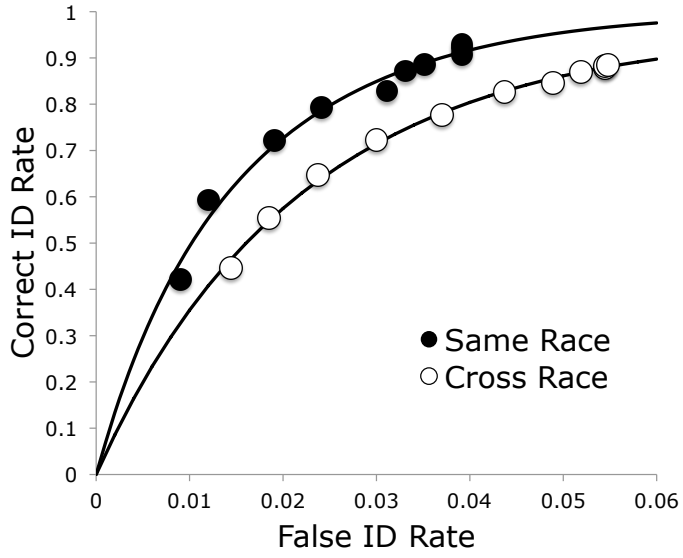


Figure 4.2. Receiver operating characteristic (ROC) plots for same-race and cross-race identifications for Asian faces (Panel A) and Caucasian faces (Panel B).

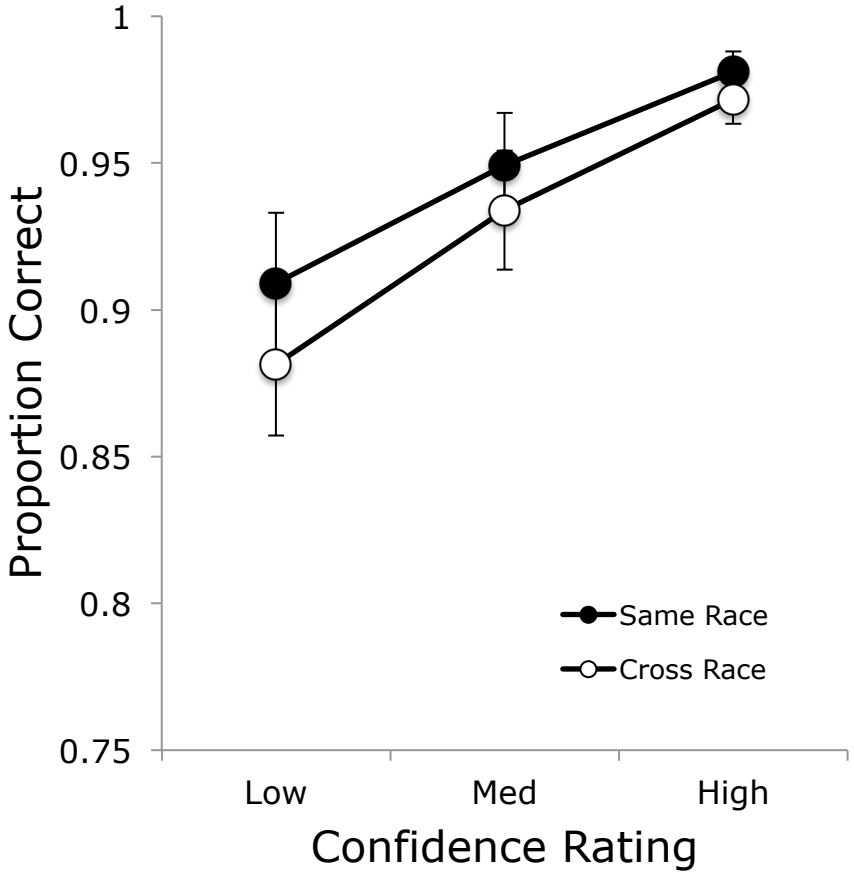
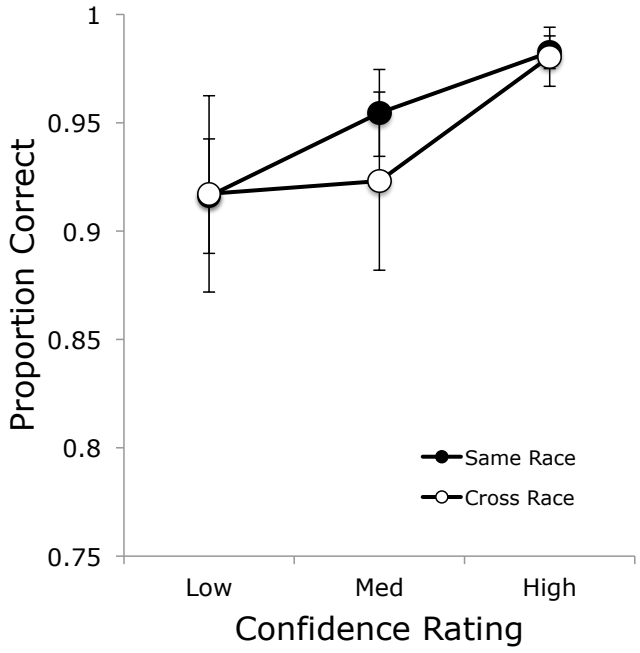


Figure 4.3. Confidence-accuracy characteristic (CAC) plots for same-race and cross-race identifications.

A



B

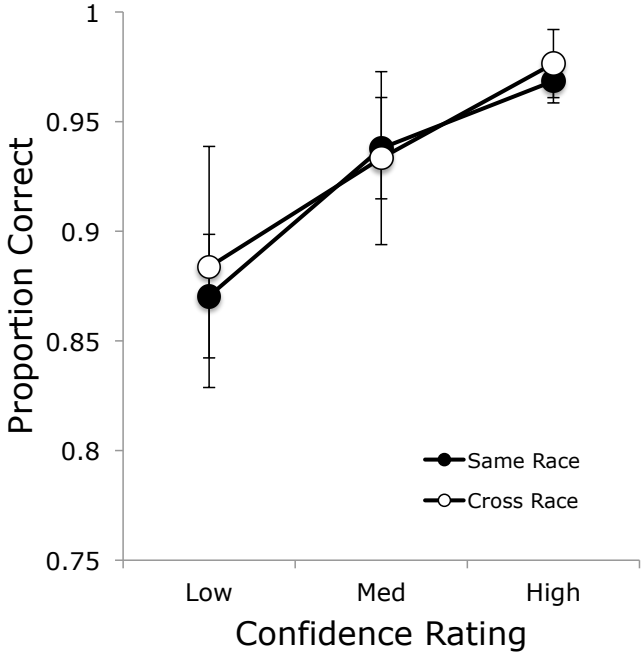


Figure 4.4. Confidence-accuracy characteristic (CAC) plots of same-race and cross-race identifications for Asian faces (Panel A) and Caucasian faces (Panel B).

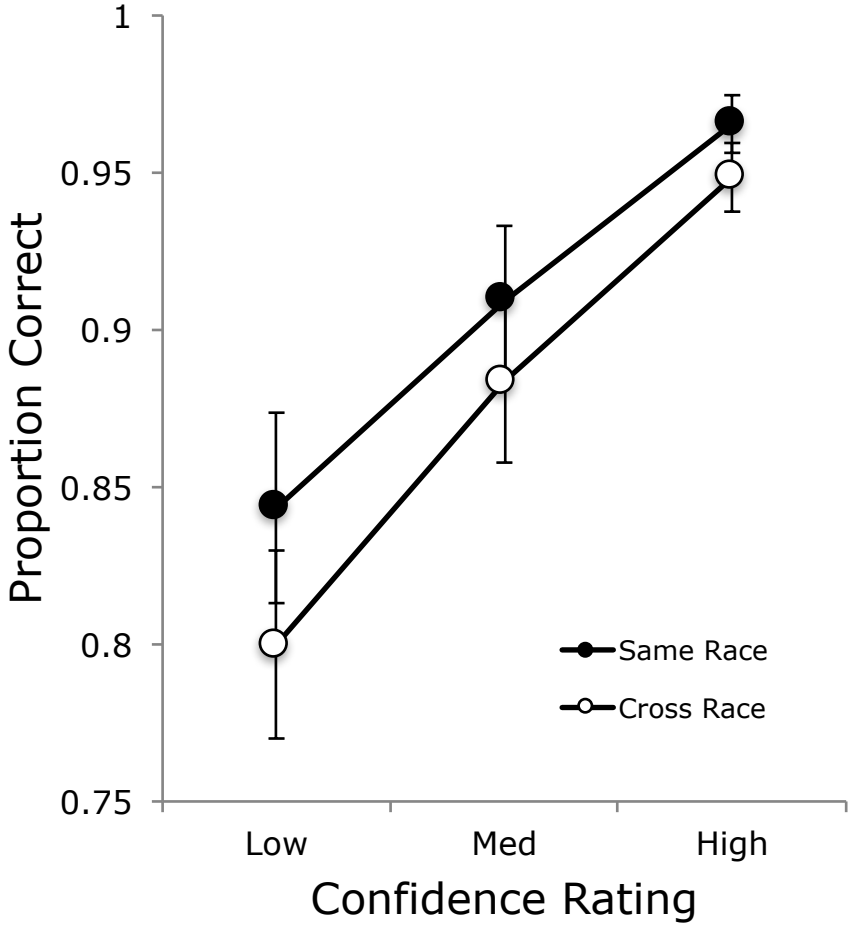


Figure 4.5. Confidence-accuracy characteristic (CAC) plots for same-race and cross-race identifications if only 35% of lineups actually contain the perpetrator.

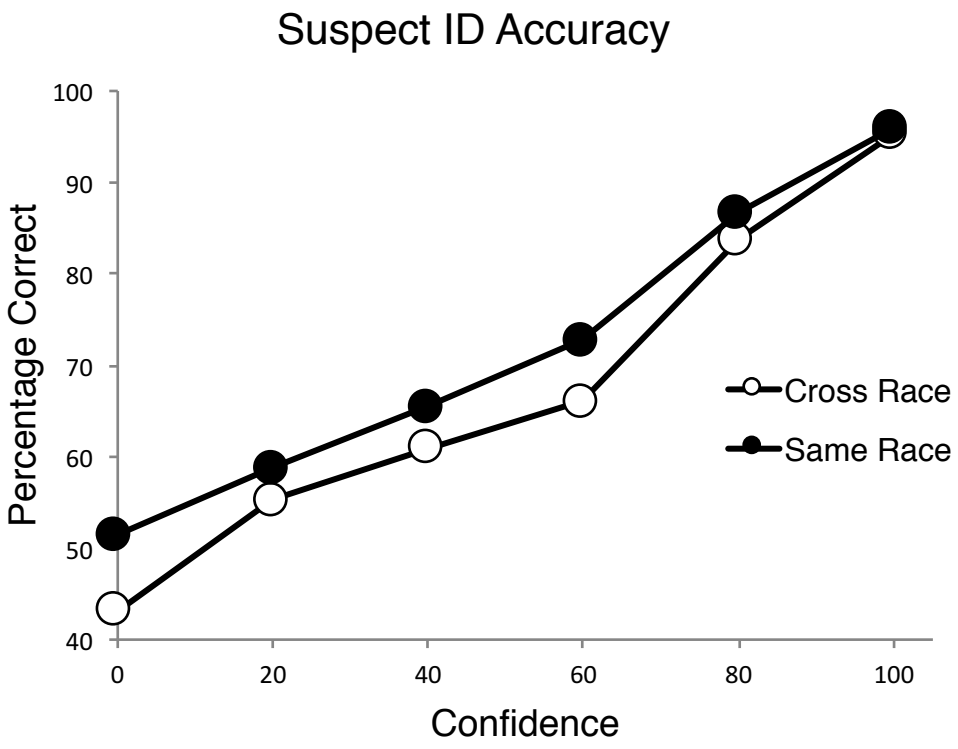


Figure 4.6. Confidence-accuracy characteristic (CAC) plots for same-race and cross-race identifications from the reanalyzed data from Dodson and Dobolyi (2016).

References

Brigham, J. C., Maass, A., Snyder, L. D., & Spaulding, K. (1982). Accuracy of eyewitness identification in a field setting. *Journal of Personality and Social Psychology*, *42*(4), 673-681.

Brigham, J. C., & Malpass, R. S. (1985). The role of experience and contact in the recognition of faces of own-and other-race persons. *Journal of Social Issues*, *41*(3), 139-155.

Carroo, A. W. (1986). Other race recognition: A comparison of Black American and African subjects. *Perceptual and Motor Skills*, *62*(1), 135-138.

Chiroro, P. M., Tredoux, C. G., Radaelli, S., & Meissner, C. A. (2008). Recognizing faces across continents: The effect of within-race variations on the own-race bias in face recognition. *Psychonomic Bulletin & Review*, *15*(6), 1089-1092.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology*, *48*(4), 879-894.

Clark, S. E., Brower, G. L., Rosenthal, R., Hicks, J. M., & Moreland, M. B. (2013). Lineup administrator influences on eyewitness identification and eyewitness confidence. *Journal of Applied Research in Memory and Cognition*, *2*(3), 158-165.

Clark, S. E., Marshall, T. E., & Rosenthal, R. (2009). Lineup administrator influences on eyewitness identification decisions. *Journal of Experimental Psychology: Applied*, *15*(1), 63-75.

Dodson, C. S., & Dobolyi, D. G. (2015). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology*, *30*(1), 113-125.

Douglass, A. B., Smith, C., & Fraser-Thill, R. (2005). A problem with double-blind photospread procedures: Photospread administrators use one eyewitness's confidence to influence the identification of another eyewitness. *Law and Human Behavior*, *29*(5), 543-562.

Haw, R. M., & Fisher, R. P. (2004). Effects of administrator-witness contact on eyewitness identification accuracy. *Journal of Applied Psychology*, *89*(6), 1106-1112.

Kassin, S. M., Ellsworth, P. C., & Smith, V. L. (1989). The "general acceptance" of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist*, *44*(8), 1089-1098.

Kassin, S. M., Tubb, V. A., Hosch, H. M.., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research: A new survey of the experts. *American Psychologist*, *56*(5), 405-416.

Lampinen, J. M., Neuschatz, J. S., Cling, A. D. (2012). *The psychology of eyewitness identification*. New York, NY: Psychology Press.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*(4), 560-572..

Loftus, E. F (1996). *Eyewitness testimony.* Cambridge, MA: Harvard University Press.

Malpass, R. S. (1974). Racial Bias in Eyewitness Identification? *Proceedings of the Division of Personality and Society Psychology*, *1*(1), 42-44.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, *13*(4), 330-334.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3-35.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*(2), 93-102.

Munsterberg, H. (1908). *On the witness stand*. New York: Doubleday.

Sangrigoli, S., Pallier, C., Argenti, A. M., Ventureyra, V. A. G., & De Schonen, S. (2005). Reversibility of the other-race effect in face recognition during childhood. *Psychological Science*, *16*(6), 440-444.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law*, *7*(1), 36-97.

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology*, *36*(12), 1546-1557.

Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J. W., & Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist*, *55*(6), 581-598.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human behavior*, *22*(6), 603-647.

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger III, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*(6), 515.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*.

Young, S. G., Hugenberg, K., Bernstein, M. J., & Sacco, D. F. (2012). Perception and motivation in face recognition: A critical review of theories of the cross-race effect. *Personality and Social Psychology Review*, *16*(2), 116-142.

Chapter 4, in full, has been submitted for publication as Wilson, B. M. & Wixted, J. T. “The Cross-Race Effect in Eyewitness Identification: Reduced Discriminability Does Not Necessarily Imply Reduced Reliability.” The dissertation author was the primary investigator and author of this manuscript.

CHAPTER 5

Conclusion

In this dissertation, I have examined how people differentiate between internal and external memory sources and discussed implications for real-world memory. The work described in Chapter 2 provides new evidence supporting the contention of reality-monitoring theory that cognitive operations are important for knowing whether or not a memory comes from an internal source. The work described in Chapter 3 shows how internally-generated information (verbal descriptions of a face) can be confused with an external memory source when the internally-generated information is not sufficiently detailed. Finally, the work described in Chapter 4 shows that people can be good at judging the strength of their memory signal when deciding whether or not an identification should be made with high confidence even when discriminability is impaired (specifically, when recognizing cross-race faces). In all of the chapters I have used signal detection theory to examine higher-level cognitive components that influence memory.

As a research tool, signal detection theory can improve understanding of real-world memory by clarifying the components that are being examined. Different types of evidence values can help differentiate between internal and external memories and between external and new memories. A key insight that signal detection theory provides is that the decision criterion can move in addition to the distributions of evidence values. It is especially interesting to also understand that just because a signal detection statistic indicates that the decision criterion moves, this does not mean that the decision criterion has

actually moved. As described next, this is one of the complications with calculating only signal detection parameters without fully understanding the underlying structure of the problem.

Signal detection theory provides useful tools for measuring memory, but even more importantly than these tools is the way it allows a problem to be better understood at a conceptual level. One frequent confusion with signal detection theory is that if a statistic measuring the criterion placement (e.g., *c*) changes, this automatically indicates that the criterion has been moved. This confusion is noted in Chapter 2. When only one distribution moves but all else remains constant (including the placement of the decision criterion), the statistic measuring criterion placement will change even if the criterion itself does not move at all.

In Chapter 2, this very phenomenon was observed in a study of reality monitoring and mindfulness meditation. Reality monitoring is the process by which people determine if a memory is internally generated or actually encountered in the real world. Mindfulness meditation provides a way of reducing cognitive operations so that researchers can observe its impact on internal and external memory sources. A signal detection analysis indicated that mindfulness meditation affects only internal memories and not external memories. The implication is that mindfulness meditation should not impair the ability to differentiate between two external memory sources but should be selective to situations where both internal and external memory signals can be

confused with each other. This is the case in the Deese-Roediger-McDermott (DRM) paradigm and also the case in the reality-monitoring paradigm used in Experiment 3 of Chapter 2.

Future research should examine the impact of mindfulness in other situations where confusions can occur between internal and external memory representations. Some research suggests that mindfulness can improve memory (Brown, Goodman, Ryan, & Anālayo, 2016), and my research would not necessarily disagree with these conclusions. The model proposed in Chapter 2 indicates that discriminability should be lower after mindfulness meditation only in situations when internal information generated during encoding can be confused with external information. In most situations, this is not a confusion that would likely be an issue for reducing discriminability. For example, discriminating between a face seen during encoding and a new face presented during test (i.e., discriminating between two external sources of a memory) would unlikely be affected by mindfulness meditation.

According to the account proposed in Chapter 2, discriminability would be lower as a result of mindfulness meditation only if a person internally imagined a face and then this internally-generated face appears on the test as a foil. It is unlikely in most scenarios that an internally-generated face would happen to match a foil used in a lineup. However, in the case of the verbal overshadowing effect examined in Chapter 3, internal and external confusions are possible because providing a verbal description of a face requires a person to internally

examine a mental image. Examining this mental image can then result in internal-external confusion because the verbal description creates a new internal representation. Future research should examine the effect of mindfulness meditation on verbal overshadowing because this is a situation where my account would predict that internal-external confusions are likely to occur.

People may always have internal-external confusions after providing a verbal description of a face. However, internal-external confusions are only problematic to the extent they contain differing information. When the verbal description is given immediately after seeing the face, the internal description is sufficiently detailed for people to be able to later identify the perpetrator on a memory test. However, when the verbal description is given after a delay, the internal description is no longer sufficiently detailed for people to be able to later identify the perpetrator on a memory test.

Reality-monitoring theory proposes that "externally generated representations are more semantically detailed—that is, contain more information or more specific information—than internally generated representations" (Johnson & Raye, 1981, p. 71). This greater detail helps people to later know that something was actually encountered in the real world rather than being internally generated. As noted in Chapter 3, this difference in the level of detail between internal and external information can also explain why a verbal overshadowing effect is found when a verbal description is provided after a 20-minute delay but not when a verbal description is provided immediately. The

external information from the actual video has a high level of detail. However, the internally-generated verbal information deteriorates as time elapses. When this internally-generated verbal information no longer has the level of detail necessary to discriminate the guilty suspect from foils, performance on the memory test is lower. This effect can be observed in the ROC data in Chapter 3.

Chapters 2 and 3 both focus on the importance of being able to discriminate between internal and external sources of information. However, it is usually just as important to be able to differentiate between different sources of external information. The ability to make this discrimination involves being able to properly evaluate internal cues as to the strength of an externally created mental representation. My research indicates that people can be quite good at being able to know if an internal strength cue is high or low. One's ability to be in tune with internal memory-strength information is revealed in the CAC plots for same-race and cross-race identifications in Chapter 4. For both same-race and cross-race identifications, accuracy is higher for high-confidence identifications than for low-confidence identifications, and accuracy is essentially equally high for identifications made with high confidence. If people did not have the ability to judge their internal strength cues, high-confidence accuracy would be reduced when discriminability is reduced (as it is in the cross-race condition). Indeed, for decades, the field of eyewitness identification has mistakenly interpreted the lower discriminability associated with cross-race memory to mean that high-

confidence cross-race IDs are less trustworthy than same-race IDs. Once properly examined, the data call that longstanding conclusion into question.

In total my dissertation examines people's abilities to attend to memory cues to be able to determine if something has actually been encountered in the real world. This is often challenging, however, because both internally-generated and other externally-encountered information can closely resemble true memories for specific events. The CAC curves in Chapters 3 and 4 suggest that people normally have a good ability to evaluate the strength of their memory cues. People know when memory strength is high and when memory strength is low. Future research should examine the exact mechanism by which people develop the ability to evaluate whether a particular memory is weak or strong. This process likely involves confirming and disconfirming feedback as to the accuracy of a particular memory decision (Mickes, Hwe, Wais, & Wixted, 2011).

Adults seem to be naturally good at judging the strength of various memory cues. One of the reasons people are naturally good at this is likely because people automatically categorize their experiences at the time they are being encountered or generated. This categorization can then later help them to know more about the source of a memory above and beyond the information itself. A primary goal of mindfulness meditation is to get people to avoid spontaneously categorizing and judging experiences. This can improve well-being as has been demonstrated in other research (Brown & Ryan, 2003). However, mindfulness meditation can also make it more difficult for people to

later know the source of memory information. Specifically, mindfulness meditation can make it difficult for people to know that something that was internally generated was actually internally generated.

My dissertation endeavored to enhance our understanding of how people judge internally-generated and externally-generated memories. My work in Chapter 2 adds support for reality monitoring theory about the importance of cognitive operations for helping to identify internally-generated information as having been internally generated. In Chapter 3, my research provides a new explanation for why a verbal overshadowing effect is observed after a delay but not when a verbal description is provided immediately after study. This can be understood by considering differences in the level of detail between internal and external memory sources, which is a critical distinction in reality monitoring theory. Chapter 4 corrects a misconception in the eyewitness memory literature by showing that even though discriminability is lower for cross-race identifications than for same-race identifications, high-confidence identifications are highly accurate for both. My dissertation connects basic and applied memory research by using signal detection theory to examine memory in the real world.

References

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*, 822–848. doi:10.1037/0022-3514.84.4.822

Brown, K. W., Goodman, R. J., Ryan, R. M., & Anālayo, B. (2016). Mindfulness enhances episodic memory performance: evidence from a multimethod investigation. *PLoS ONE*, *11*(4), e0153309.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*, 67–85.

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140*, 239-257.