Title

The confidence-accuracy relationship: Deepening our understanding of confidence and uncertainty

Author

Killeen, Isabella Mackenzie

Publication Date

2018

UNIVERSITY OF CALIFORNIA SAN DIEGO

The confidence-accuracy relationship:
Deepening our understanding of confidence and uncertainty

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

in

Psychology

by

Isabella M. Killeen

Committee in charge:

Professor Caren M. Walker, Co-Chair
Professor John T. Wixted, Co-Chair
Professor Gedeon Deák
Professor Timothy Rickard
Professor Federico Rossano

2018

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

Co-Chair

Co-Chair

University of California San Diego

2018

DEDICATION

For Pamela Squier-Solman, William Castle, and Alice Squier:

Gone but never forgotten.

TABLE OF CONTENTS

Signature Page…………………………………………………………………………………iii

Dedication…………………………………………………………………………………...iv

Table of Contents……………………………………………………………………………v

List of Figures…………………………………………………………………………………vi

List of Tables……………………………………………………………………………….viii

Acknowledgements……………………………………………………………………………ix

Vita……………………………………………………………………………………………...x

Abstract………………………………………………………………………………………...xi

Chapter 1 Introduction to the Dissertation……………………………………………………...1

Chapter 2 Eyewitness Identification: A Reconsideration of the Confidence-Accuracy Relationship in Children…………………………………………………………………12

Chapter 3 The Development of a Recognition-Memory Confidence-Accuracy Relationship for Objects and Faces in Preschoolers………………………………………………………….46

Chapter 4 A Signal-Detection-Based Comparison of Simultaneous and Elimination Lineups using Receiver Operating Characteristic Analysis………………………………………………68

Chapter 5 Confidence Scale Use in Preschool-Aged Children: Effects of Disconfirming Evidence……………………………………………………………………………………...101

Chapter 6 Conclusion………………………………………………………………………...124

LIST OF FIGURES

Figure 2.1 Confidence-Accuracy Characteristic for kindergartener's and third-grader's old and new recognition judgments……………………………………………………………………35

Figure 2.2 Confidence-Accuracy Characteristic for children with Autism Spectrum Disorder and typically developing children………………………………………………………….……36

Figure 2.3 Confidence-Accuracy Characteristic for adults' and children's old and new recognition judgments………………………………………………………………….…..37

Figure 2.4 Confidence-Accuracy Characteristic for 2-alternative forced choice object recognition in 3, 4, and 5-year-olds…………………………………………………………………..38

Figure 2.5 The average Confidence-Accuracy Characteristic for children, young adults, and older adults (Shing et al., 2009)…………………………………………………………………..…39

Figure 2.6 The average Confidence-Accuracy Characteristic for children, young adults, and older adults (Fandakova et al., 2013)……………………………………………………………..40

Figure 2.7 Confidence-Accuracy Characteristic for Suspect IDs in children and adolescents….41

Figure 2.8 Aggregate Confidence-Accuracy Characteristic for Suspect IDs in children………..42

Figure 3.1. Average confidence for correct and incorrect answers for line drawings of objects (Experiment 1)…………………………………………………………………………………...62

Figure 3.2. Confidence-Accuracy Characteristic curves for line drawings of objects (Experiment 1)……………………………………………………………………………………………63

Figure 3.3. Average confidence for correct and incorrect answers for faces (Experiment 2)…..64

Figure 3.4. Confidence-Accuracy Characteristic curves for faces (Experiment 2)……………...65

Figure 4.1. ROC data for the simultaneous and traditional elimination lineup procedures with fitted curves……………………………………………………………………………………93

Figure 4.2. Confidence-accuracy data for simultaneous and traditional elimination lineups……94

Figure 4.3. ROC data for the simultaneous and novel elimination lineup procedures with fitted curves……………………………………………………………………………………………95

Figure 4.4. Confidence-accuracy data for simultaneous and novel elimination lineups………...96

Figure 5.1.a. Novel 3-point confidence scale……………………………………………..114

Figure 5.1.b. Standard confidence scale……………………………………………..………114

Figure 5.2. Window stimuli with a target shape…………………………………………..…115

Figure 5.3. Target shapes paired by column with their corresponding “disconfirming shapes”.116

Figure 5.4. Average confidence across clear, partially, and fully occluded windows for the novel and standard confidence scales…………………………………………………..………..117

Figure 5.5.a. Average confidence as a function of window occlusion for participants who saw disconfirming evidence after the first trial……………………………………………...……118

Figure 5.5.b. Average confidence as a function of window occlusion for participants who saw confirming evidence after the first trial…………………………………………………...…118

Figure 5.6. Average confidence as a function of window occlusion for 3-year-olds, separated by trial…………………………………………………………………………………...…119

Figure 5.7. Average confidence as a function of window occlusion for 4-year-olds, separated by trial…………………………………………………………………………………...……120

Figure 5.8. Average confidence as a function of window occlusion for 5-year-olds, separated by trial…………………………………………………………………………..……….121

LIST OF TABLES

Table 4.1. Number of participants in each condition for Experiment 1……………………………91

Table 4.2. Number of participants in each condition for Experiment 2……………………………92

ACKNOWLEDGEMENTS

I would like to thank John Wixted and Caren Walker for their continued support and guidance, for being my coauthors, and for co-chairing this committee.

I would like to thank my coauthor and fellow Wixted Lab member Brent Wilson for his support and collaboration.

I would like to thank the research assistants of the Wixted Lab and the Early Learning & Cognition Lab who helped make this work possible: Sierra Ampudia, May Jaber, Nadia Keddo, Sally La, Joelle Robinett, Abigail Sumi, and Jordan Viernes. I would also like to thank the lab coordinators of the Early Learning & Cognition Lab: Andie Nishimi and Nicky Sullivan.

Chapter 2, in full, is currently being prepared for submission for of the material. Killeen, Isabella M. The dissertation author was the primary investigator and author of this material.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Killeen, Isabella M.; Walker, Caren M.; Wixted, John T. The dissertation author was the primary investigator and author of this material.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Killeen, Isabella M.; Wilson, Brent M.; Wixted, John T. The dissertation author was the primary investigator and author of this material.

Chapter 5, in full, is currently being prepared for submission for publication of the material. Killeen, Isabella M.; Walker, Caren M. The dissertation author was the primary investigator and author of this material.

VITA

Education

2018 Doctor of Philosophy in Psychology, University of California San Diego

2016 Master of Arts in Psychology, University of California San Diego

2014 Bachelor of Science in Psychology, *Magna Cum Laude*, Colorado State University

Conference Presentations

Killeen, I.M. & Walker, C.M. (2017). Confidence scale use in preschool-aged children: Effects of disconfirming evidence. Poster at the Cognitive Development Society annual meeting. Portland, OR.

Killeen, I.M., Vieane, A.Z., & Clegg, B.A. (2015). The effects of one's expressed confidence on the memory and metacognition of others. Poster at the Western Psychological Society annual meeting. Las Vegas, NV.

ABSTRACT OF THE DISSERTATION

The confidence-accuracy relationship: Deepening our understanding of confidence and uncertainty

by

Isabella M. Killeen

Doctor of Philosophy in Psychology

University of California San Diego, 2018

Professor Caren M. Walker, Co-Chair
Professor John T. Wixted, Co-Chair

This dissertation builds upon our existing understanding of confidence, uncertainty, and their relationship to recognition memory accuracy. Its basic approach is to measure overt expressions of confidence in children and adults and to examine both the development of confidence and its relationship to memory accuracy. The work consists primarily of one review paper (Chapter 2) and three research papers (Chapters 3, 4 and 5) that each seek to deepen our understanding of confidence in a unique way. Chapter 2 reviews the existing literature on the confidence-accuracy relationship in children, both in developmental and eyewitness contexts. Chapter 3 provides new evidence for the emergence of a relationship between confidence and

recognition memory accuracy during the preschool years. Chapter 4 tests for a confidence-accuracy relationship in adults using an eyewitness memory paradigm developed for children. Chapter 5 introduces a novel paradigm that uses disconfirming evidence (e.g. evidence that disconfirms a previously held hypothesis about the state of the world) to improve confidence scale use in preschool-aged children. This research expands our knowledge of how confidence develops, as well as how the relationship between confidence and memory accuracy changes between early childhood and adulthood.

CHAPTER 1

Introduction to the Dissertation

Adults use uncertainty monitoring and overt expressions of confidence as a way to (among other things) indicate the likely accuracy of their memories. Basic memory research has shown that when an adult expresses a high level of confidence in a memory (e.g. "I am 100% sure I have encountered this before) that memory is very likely to be accurate. Conversely, when an adult expresses a low level of confidence in a memory (e.g. "I'm only 30% sure that I have encountered this before) that memory is likely to be inaccurate or error-prone. This relationship between confidence and memory accuracy is referred to as the "confidence-accuracy relationship." In adults, it has most commonly been shown using "old/new" paradigms, where a participant is asked to recognize whether a picture or word on a test list was presented on a previously encoded list (e.g. it was recently encountered and is "old") or not (e.g. it was not recently encountered and is "new") (e.g. Hiller & Weber, 2013; Mickes, Hwe, Wais, & Wixted, 2011; Tekin & Roediger, 2017). However, it has also been shown in eyewitness memory paradigms, where an adult is shown a video of a mock crime, then asked to identify the "culprit" (e.g. the person in the video) out of a six- to eight-person photo lineup where the photos are presented either simultaneously (e.g. in an array) or sequentially (e.g. one at a time consecutively) (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted & Wells, 2017). This is important to the legal system because if an eyewitness's confidence for their initial identification of a suspect from a lineup is indicative of the accuracy of that decision (e.g. the likelihood that the suspect is guilty), eyewitnesses can be considered "reliable" and their identification used to further police investigations and legal proceedings.

While there has been a significant amount of research on the confidence-accuracy relationship in adults both in basic memory and eyewitness contexts, less research has been conducted in children. Some recognition memory studies have collected confidence ratings from children, but instead of reporting the confidence-accuracy characteristic they instead report children's mean confidence for their correct responses versus their mean confidence for their incorrect responses. This analysis is performed when confidence is being used as an explicit measure of uncertainty monitoring. If children express significantly higher confidence for their correct answers when compared to their incorrect answers, they are thought to have developed uncertainty monitoring. Uncertainty monitoring has consistently been seen in children older than five (Destan, Hembacher, Ghetti, & Roebers, 2014; Roderer & Roebers, 2010) for recognition memory tasks.

When do children develop the ability to monitor their own uncertainty? Very little research has been done on these subjects. What research has been done suggests that for perceptual tasks (e.g. identifying an object out of intact vs. degraded images, identifying colors) and lexical tasks (e.g. identifying the names of common vs. uncommon items), uncertainty monitoring develops at around three years old (Lyons & Ghetti, 2011; Beran, Decker, Schwartz & Smith, 2012). For recognition memory tasks, uncertainty monitoring develops around four years old (Hembacher & Ghetti, 2014). However, these are the only three studies conducted in this age group. The results of Hembacher & Ghetti (2014) suggest that for recognition memory decisions, 3-year-olds lack the ability to monitor their own uncertainty, and therefore give similar confidence ratings for both correct and incorrect answers. 5-year-olds, on the other hand, give significantly higher confidence ratings for correct vs. incorrect answers, and thus are thought to have developed uncertainty monitoring. However, even if a child is able to differentiate when they are correct from when they are incorrect using confidence, it does not speak to the overall accuracy of their memory. In the eyewitness literature, the measure of interest is how well confidence is able to predict the overall accuracy of identifications (usually measured as percent correct). This is not to say that these measures do not have some level of redundancy. More likely than not, children who show uncertainty monitoring will also show a confidence-accuracy relationship. But how accurate, exactly, is a decision made with high confidence? A different analysis is needed to answer that question.

Without uncertainty monitoring, the confidence-accuracy relationship cannot exist, because children would not be able to use confidence as an outward expression of their metacognitive state. As mentioned above, children express their uncertainty though their explicit confidence judgments for incorrect answers being lower than for correct answers. Thus, the underlying mechanism by which children learn to use confidence is rooted in their ability to understand when they are incorrect. Initially, children that have developed the ability to monitor their own uncertainty may still show a weak confidence-accuracy relationship. This is because the overall accuracy of their memories still may be rather low. For example, prior research has shown that children do not show adult-like recognition memory accuracy for faces until early adolescence (Bruce et al., 2000; Carey, Diamond & Woods; 1980; Mondloch, Geldart, Maurer & Le Grand, 2003). Thus, even for high-confidence decisions they may be only 60 to 70 percent accurate. But through uncertainty monitoring children may be able to distinguish which of their answers are more likely to be correct, even if their overall accuracy is similarly low across all levels of confidence. Over time they will become more calibrated and thus more selective about when they express high confidence, eventually only expressing it for their strongest memory signals. It is at this point that a strong confidence-accuracy relationship would emerge.

Children are often said to start out as “eternal optimists,” (Mickes et al., 2011) where they are always certain their memories are correct and always express high confidence. Then, as the children get older, they begin to learn that sometimes their decisions based on their memories are incorrect, and from there they begin to appreciate that errors tend to occur when memory strength is weak and that uncertainty is warranted under those conditions. As a result of this understanding of uncertainty, children start expressing confidence based on their interpretation of their memory strength, rather than simply always expressing high confidence.

No studies have tested for the development of the confidence-accuracy relationship in children. However, there are studies that report a strong confidence-accuracy relationship in children. Both 5- to 9-year-old (Berch & Evans, 1973) and 8- to 12-year-old (Hiller & Weber, 2013) children show a strong relationship between expressed confidence and memory accuracy for old/new recognition decisions. The reanalysis in Chapter 2 also showed a confidence-accuracy relationship for 7- to 12-year-old children in old/new, 2-alternative forced-choice, and eyewitness paradigms. Additionally, children have shown a confidence-accuracy relationship and uncertainty monitoring in recall studies. Many studies have shown uncertainty monitoring in recall for children as young as seven years old in both basic and eyewitness tasks (Roebers, 2002; Roderer & Roebers, 2003; Roebers, von der Linden, Schneider, & Howie, 2007). One study has also shown a confidence-accuracy relationship for 8- and 10-year-old children in an eyewitness recall task (Howie & Roebers, 2007). Howie & Roebers (2007) found that children showed a strong, adult-like confidence-accuracy relationship for unbiased questions, but showed almost no confidence-accuracy relationship for misleading questions. Unbiased questions provide no evidence for a particular answer, while misleading questions do. An example on an unbiased question would be, “What color was the girl’s hair?”, while an example of a misleading question would be, “The girl’s hair was red, right?” This is consistent with other research that has shown children are more likely to trust and remember misinformation than adults (Ceci & Bruck, 1993; Ackil & Zaragoza, 1995).

If children develop uncertainty monitoring for memory tasks around three years old, and show a strong confidence-accuracy relationship for memory tasks at five years old, then it is plausible that the confidence-accuracy relationship develops sometime during the preschool years. Chapter 3 demonstrates that for the same set of stimuli used by Hembacher & Ghetti (2014)- line drawings of objects- 4- and 5-year-olds show a relationship between their expressed confidence and the accuracy of their memories. However, when the stimuli were switched from objects to faces, 4- and 5-year-olds did not show a confidence-accuracy relationship. For faces, the confidence-accuracy relationship may not develop until children are older than five.

The stimuli in Chapter 3 were switched from line drawings to faces because the use of face stimuli, even in a recognition memory paradigm, can be informative about the reliability of eyewitnesses. In the case of Chapter 3, preschool-age children may not be reliable eyewitnesses if they do not show a confidence-accuracy relationship for face stimuli. However, we know that adults (Wixted et al, 2015; Wixted & Wells, 2017) and 14- to 17-year-old adolescents (Brewer & Day, 2005) show a strong confidence-accuracy relationship. Based on the reanalysis in Chapter 1, it appears that 8- to 12-year-old children also show a strong confidence-accuracy relationship. In both cases, this strong-confidence accuracy relationship was evoked using a "simultaneous lineup", where six to eight faces are presented simultaneously in an array. This is an eyewitness procedure commonly used for adults. In Chapter 4, adults are tested using an eyewitness paradigm originally designed for children called the "elimination lineup." The elimination lineup separates the identification process into two separate judgments: first, a relative judgment where the child selects the most familiar face in the lineup, then an absolute judgment where the child decides whether the face they have chosen is the "culprit." Previous research has concluded that this two-step process is an improved procedure for both child and adult eyewitnesses (Pozzulo &

Lindsay, 1999; Pozzulo et al., 2008). However, neither of these studies measured the confidence-accuracy relationship. In Chapter 4, we found that for both the simultaneous and the elimination lineup, identifications made with high confidence were over 95% correct.

Based on the results of Chapter 3 and previous literature, there is evidence that 3- and 4-year-olds show a weaker relationship between confidence and accuracy than 5-year-olds. One possibility for why this is the case is that 3- and 4-year-olds are less capable of using a confidence scale than 5-year-olds. Even if 3- and 4-year-olds are able to monitor their own uncertainty, in order to show a confidence-accuracy relationship they need to be able to overtly express confidence using a scale. In Chapter 5, we test a novel paradigm with the goal of improving the confidence scale use of children in this age group. We find that, through the use of *disconfirming evidence,* we can improve confidence scale use in 3- and 4-year-olds. *Disconfirming evidence* is evidence presented to the child that disconfirms a preexisting hypothesis that the child holds. In this case, children hold a preexisting hypothesis that a partially occluded shape matches a target shape shown next to it. When we reveal to the children that the partially occluded shape does not match the target shape, this evidence disconfirms the hypothesis that they held. A variation of the novel paradigm developed for Chapter 5 may function as a training task used to familiarize young children with a confidence scale prior to its use as a part of a different experiment.

This dissertation aims to expand our understanding of the relationship between confidence and memory accuracy, particularly in children. Gaining a deeper understanding of how the confidence-accuracy relationship develops may lead to a stronger theory about the cognitive mechanisms that underlie confidence and uncertainty. This dissertation also provides evidence for the emergence of the confidence-accuracy relationship during the preschool years.

Knowing when the confidence-accuracy relationship first emerges, and how we can facilitate is emergence by improving confidence scale use in young children, helps us map a more detailed developmental trajectory of this relationship. As we continue to refine this developmental trajectory, we will come closer to understanding how humans go from “eternal optimists” (Mickes et al., 2011) as young children to experts at expressing confidence as adults.

References

Ackil, J.K. & Zaragoza, M.S. (1995). Developmental differences in eyewitness suggestibility and memory for source. *Journal of Experimental Child Psychology, 60*(1), 57-83.

Beran, M.J., Decker, S.L., Schwartz, A. & Smith, J.D. (2012). Uncertainty monitoring by young children in a computerized task. *Scientifica, 2012.*

Berch, D.B. & Evans, R.C. (1973). Decision processes in children's recognition memory. *Journal of Experimental Psychology, 16*, 148-164.

Brewer, N. & Day, K. (2005). The confidence-accuracy and decision latency-accuracy relationships in children's eyewitness identification. *Psychiatry, Psychology, and Law, 12*(1), 119-128.

Bruce, V., Campbell, R. N., Doherty-Sneddon, G., Import, A., Langton, S., McAuley, S., & Wright, R. (2000). Testing face processing skills in children. *British Journal of Developmental Psychology, 18*, 319– 333.

Carey, S., Diamond, R., & Woods, B. (1980). Development of face recognition: A maturational component? *Developmental Psychology, 16*, 257-269.

Ceci, S.J. & Bruck, M. Suggestibility of the child eyewitness: A historical review and synthesis. *Psychological Bulletin, 113*(3), 403-439.

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C.M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology, 126*, 213-228.

Hembacher, E. & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*(9), 1768-1776.

Hiller, R.M. & Weber, N. (2013). A comparison of adults' and children's metacognition for yes/no recognition decisions. *Journal of Applied Research in Memory and Cognition, 2*, 185-191.

Howie, P. & Roebers, C.M. (2007) Developmental progression in the confidence-accuracy relationship in event recall: Insights provided by a calibration perspective. *Applied Cognitive Psychology, 21*, 871-893.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

Koriat, A., Goldsmith, M., Schneider, W., & Nakash-Dura, M. (2001). The credibility of children's testimony: Can children control the accuracy of their memory reports? *Journal of Experimental Child Psychology, 79*, 405–437.

Lyons, K.E. & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development, 82*(6). 1778-1787.

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.

Mondloch, C.J., Geldart, S., Maurer, D., & Le Grand, R. (2003) Developmental changes in face processing skills. *Journal of Experimental Child Psychology, 86*, 67-84.

Plude, D. J., Nelson, T. O., & Scholnick, E. K. (1998). Analytical research on developmental aspects of metamemory. *European Journal of Psychology of Education, 13*, 29–42.

Pozzulo, J.D., Dempsey, J., Corey, S., Girardi, A., Lawandi, A., & Aston, C. (2008). Can a lineup procedure designed for child witnesses work for adults? Comparing simultaneous, sequential, and elimination lineup procedures. *Journal of Applied Social Psychology, 38*(9), 2195-2209.

Pozzulo, J.D. & Lindsay, R.C.L. (1999). Elimination lineups: An improved identification procedure for child eyewitnesses. *Journal of Applied Psychology, 84*(2), 167-176.

Roderer, T. & Roebers, C.M. (2003). Confidence judgments in event recall: Developmental progression in the impact of question format. *Journal of Experimental Child Psychology, 85*, 352-371.

Roderer, T. & Roebers, C.M. (2010). Explicit and implicit confidence judgments and developmental differences in metamemory: An eye-tracking approach. *Metacognition & Learning, 5*, 229-250.

Roebers, C.M. (2002). Confidence judgments in children's and adults' event recall and suggestibility. *Developmental Psychology, 38*(6), 1052-1067.

Roebers, C.M., von der Linden, N., Schneider, W., & Howie, P. (2007). Children's metamemorial judgments in an event recall task. *Journal of Experimental Child Psychology, 97*, 117-137.

Sanders, J.I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron, 90*, 499-506.

Tekin, E. & Roediger, H.L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*(1), 49.

Wixted, J.T., Mickes, L., Clark, S.E., Gronlund, S.D., & Roediger H.L. (2015). Initial eyewitness confidence reliably predicts identification accuracy. *American Psychologist, 70*, 515-526.

Wixted, J.T. & Wells, G.L. (2017) The relationship between eyewitness confidence and Identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10-65.

Chapter 2

Eyewitness Identification: A Reconsideration of the Confidence-Accuracy Relationship in Children

Abstract

In the eyewitness identification literature, children are believed to be unreliable eyewitnesses. The basis for this belief is that eyewitness identification studies in children have generally reported a weak confidence-accuracy relationship. However, basic developmental research (e.g. memory for word pairs) has shown a strong confidence-accuracy relationship in children. This paper reviews the apparently conflicting evidence concerning the confidence-accuracy relationship in children. Previously reported results from studies that tested children using fair lineups are reanalyzed, and all of the relevant findings are plotted in terms of accuracy as a function of confidence. This straightforward plot is known as the confidence-accuracy characteristic (CAC), and it is now widely used in the adult eyewitness identification literature. The re-analysis reported here reveals a strong confidence-accuracy relationship for children, though not as strong as adults.

Introduction

Notable exceptions notwithstanding, until recently, the general consensus in the field of eyewitness identification was that there is, at best, a modest relationship between an adult eyewitness's initial confidence in their lineup decision and the accuracy of that decision (Wixted, Mickes & Fisher, 2018). This consensus is directly opposed to the consensus of basic memory researchers, who almost invariably find a strong relationship between confidence and accuracy for adults (e.g., Mickes, Hwe, Wais, & Wixted, 2011). It was originally thought that the reason for this apparent discrepancy was some fundamental difference between lineup recognition decisions in eyewitness identification (ID) experiments and old/new recognition decisions in list memory experiments. However, it has recently become clear that there is more to the story than that. Using confidence-accuracy characteristic analysis (which simply plots confidence as a function of accuracy instead of computing a correlation coefficient), researchers found that adults show a strong relationship between their confidence in a lineup decision and the accuracy of that decision (Wixted et al., 2015; Wixted & Wells, 2017).

Here, I inquire into the issue of whether or not there is a similarly strong relationship between confidence and accuracy in children, particularly after they have reached 8-10 years old, by reviewing the relevant basic and applied literatures. As in the adult cognitive psychology literature, the basic developmental psychology literature consistently finds that around 7 years of age (if not earlier), children develop the ability to monitor the strength of their memory signals and express their confidence accordingly (Roebers, 2014). Their performance may not be quite as good as adults, and they may be somewhat overconfident, but there is still a strong relationship such that expressed confidence holds considerable information value. More specifically, this research shows that children express significantly higher confidence in their accurate memories as compared to their inaccurate memories.

The confidence-accuracy relationship in children is an important issue because the legal system sometimes has to rely upon the testimony of child eyewitnesses. Based on psychological research, evidence suggests that children are regarded in the legal system as being unreliable and overconfident (e.g., Keast, Brewer & Wells, 2007). For example, *Expert Evidence: Law, Practice, Procedure, and Advocacy* is a book that has "has been cited by superior courts in every jurisdiction in Australia and New Zealand." In its chapter on eyewitness testimony, it states that confidence is not a useful guide to accuracy for children's identification responses (Powell, Garry, & Brewer, 2013). However, research in this domain as it applies to children has largely come to a standstill despite more informative measures having been developed for adults in recent years. Would the evidence still show a weak confidence-accuracy relationship for children if newer methods of analysis were brought to bear on the problem? It seems like a timely question to ask given that newer methods of analysis have recently changed how the field understands the confidence-accuracy relationship for adults (Wixted & Wells, 2017).

The goal of this paper is to bring attention back to an issue seemingly left unsolved, by exploring in some detail why there is such a disagreement between the basic and eyewitness literature when it comes to the reliability of children's confidence in their memories. The confidence that a child expresses when presented with a lineup may hold considerable information value, but, if so, the legal system would not be aware of that. The fact that the basic developmental literature has concluded that children are able to distinguish their correct from incorrect answers using confidence raises the possibility that the confidence-accuracy measures used in the eyewitness identification literature may incorrectly imply that child eyewitnesses are generally unreliable (as was true of the adult literature).

It should be noted that this review only considers the confidence-accuracy relationship for an initial identification made using a scientifically tested lineup procedure, under pristine (e.g. fair and unbiased) conditions. The initial test involves administering a fair lineup with an immediate confidence judgment, and not allowing for any influence from the lineup administrator (e.g. not allowing instructions that bias the eyewitness towards making or withholding their identification). A fair lineup is one where all of the filler faces look similar to the suspect, and share the same level of similarity (e.g. a Caucasian suspect should not be surrounded by some Hispanic and some African-American faces). Repeated testing (e.g. having an eyewitness view more than one lineup) is known to inflate the confidence-accuracy relationship (Deffenbacher, Bornstein, & Penrod, 2006; Wixted & Wells, 2017), such that only confidence in response to the first lineup is a reliable indicator of accuracy (Wixted et al., 2015; Wixted & Wells, 2017). The confidence-accuracy relationship is known to be significantly weaker when certain non-pristine testing conditions are used, in particular, when unfair lineups are used (Wixted & Wells, 2017). In this paper, all of the eyewitness identification studies considered were run under pristine conditions. This paper also only considers the confidence-accuracy relationship for lineup procedures that have faced rigorous scientific testing, such as simultaneous and sequential lineup techniques. Other lineups, like the elimination lineup, fall outside the scope of this review.

Research on the Confidence-Accuracy Relationship in Adults

When the eyewitness memory field began more than 30 years ago, a statistic called the point biserial correlation coefficient was typically used to measure the relationship between confidence and accuracy. This approach reduced the confidence-accuracy relationship down to one number – a Pearson correlation between whether a response is correct or incorrect (coded, for example, as 0 or 1) and its corresponding confidence rating (e.g., on a 1 to 5 Likert scale). Initially, this method did not involve separating those individuals who were "choosers" (e.g. those who actually identify a person from a lineup) from "non-choosers" (e.g. those who reject the lineup). When these two groups were combined, an early review of the literature by Wells and Murray (1984) reported a point biserial correlation coefficient of .07. Based on that result, they concluded that "the eyewitness accuracy-confidence relationship is weak under good laboratory conditions and functionally useless in forensically representative settings" (p. 165). Many other studies reported similar findings. Then, in a later meta-analysis of this literature, Sporer, Penrod, Read and Cutler (1995) separated the "choosers" from the "non-choosers". This separation improved the information value of the correlation coefficient. Only choosers have the potential to imperil someone in a court of law, because they actually identified someone from the lineup. If a witness does not choose an individual from a lineup, their testimony will most likely not be used against the suspect. Thus, it makes sense to remove non-choosers from the calculation of the correlation coefficient, because their inclusion may mask a stronger relationship between confidence and accuracy for choosers. Sporer et al. (1995) showed this exact masking effect, because when they excluded non-choosers from their calculations, the coefficient improved to 0.41 on average. Thus, these authors noted that confidence accounts for about 17% of the variance in accuracy. Despite this improvement, Penrod and Culter (1995) concluded that even just for choosers, eyewitness confidence "… is a weak indicator of eyewitness accuracy even when measured at the time an ID is made and under relatively 'pristine' laboratory conditions" (p. 830).

A short time later, Juslin, Olsson, & Winman (1996) showed that the point biserial correlation coefficient can be low even when there is an undeniably strong relationship between confidence and accuracy. In fact, it can be low even when the confidence and accuracy are perfectly calibrated. Perfect calibration means that when an eyewitness indicates that they are 60% confident, they will be correct 60% of the time. That is, if eyewitnesses indicate 70% confidence, they will be correct 70% of the time, and so on. Even when this perfect calibration is achieved, the point biserial correlation coefficient can be low or high, depending on how the responses are distributed across the confidence levels.

With this discovery, many in the eyewitness field shifted to using calibration curves as an alternative measurement. After Sporer et al. (1995), calibration curves were calculated separately for choosers and non-choosers. The calibration between confidence and accuracy is computed by plotting the proportion of correct identifications as a function of confidence from 0 to 100. As mentioned above, perfect calibration occurs when an eyewitness's confidence is directly proportional to their accuracy (e.g. 60% confidence indicates 60% accuracy). Under fair, unbiased conditions, calibration studies revealed a stronger-than-expected relationship between initial eyewitness confidence and the accuracy of a lineup decision (e.g., Brewer & Wells, 2006). For example, identification decisions made with low-confidence (e.g., 0 to 20%) were correct about 30% of the time, whereas decisions made with high confidence (e.g., 90 to 100%) were correct about 80% of the time. Still, eyewitnesses were considered to be overconfident in their memories at the high end of the scale (Brewer & Wells, 2006).

The calibration measure is not ideally suited to the lineup task. In a lineup task, or any sort of recognition memory task, 0% confidence is not actually expected to be indicative of 0% accuracy because random chance performance on a recognition memory task is above 0%

accuracy. In a two-alternative forced-choice task, for example, chance performance is 50%. In a standard six-photo simultaneous lineup, chance performance is approximately 17%. Non-zero chance performance distorts the calibration curves and makes it appear that participants are expressing confidence that does correspond appropriately to memory strength.

It has also been argued that there are even more important reasons why calibration curves are not the most informative way to calculate the confidence-accuracy relationship (Mickes, 2015). In a lineup decision, a person who makes a positive identification can have identified either the suspect or a filler. Calibration curves combine people who make either of these two types of identifications (suspect or filler) into the choosers group. This is problematic because suspect IDs and filler IDs indicate completely different things about the likely guilt of the suspect. If a person makes a suspect ID, it is evidence of guilt, while if a person makes a filler ID, it is evidence of innocence. High-confidence suspect IDs rarely occur when the suspect is innocent (Wixted, Mickes, Clark, Gronlund & Roediger, 2015), so putting these IDs into the same category as filler IDs, which more often happen in target-absent compared to target-present lineups (Wells & Olson, 2002), obscures the information value of suspect IDs. Thus, the measure that is arguably of most relevance to a case involving eyewitness testimony against a defendant is suspect ID accuracy (Mickes, 2015). It follows from these considerations that a more informative way to quantify the confidence-accuracy relationship is to calculate just the suspect ID accuracy rate as a function of confidence. Using this method, termed "confidence-accuracy characteristic" (CAC) analysis (Mickes, 2015), confidence is typically found to be highly indicative of accuracy for fair, unbiased lineups. Additionally, accuracy for high-confidence identifications is very high (Wixted et al., 2015). A key unanswered question is whether or not the same is true for children.

Theoretical Basis for a Confidence-Accuracy Relationship in Children

“Confidence” has many definitions across the psychological literature. One definition considers confidence to be the expression of a metacognitive ability called “uncertainty monitoring” which refers to one’s capacity to introspect on the likely (in)accuracy of one’s decisions (e.g. to monitor one’s own uncertainty). In other words, uncertainty monitoring is the term for an internal metacognitive process, and confidence is the outward expression that occurs as a result of that process (Koriat & Goldsmith, 1996). Confidence has also been characterized from a purely statistical standpoint as the Bayesian posterior probability that a decision maker is correct, particularly when the decision is made under conditions of uncertainty (Sanders, Hangya, & Kepecs, 2016). These two definitions are not mutually exclusive. They can be combined to say that confidence is an expression of the Bayesian posterior probability that one’s decision is accurate, based on one’s ability to introspect about the evidence presented.

Once children achieve some level of autonomy through the development of movement and coordination, as well as language, they are confronted with constant dilemmas that they must negotiate. Theoretically, if they develop the ability to monitor and introspect about their internal cognitive signals, they would be able to make more adaptive choices in situations like that (Koriat & Goldsmith, 1996). In this regard, age-related improvements in metamemory have been found to yield improvements in memory accuracy across many tasks (Flavell, 1989; Koriat, Goldsmith, Schneider, & Nakash-Dura, 2001; Plude, Nelson, & Scholnick, 1998). One relevant aspect of metamemory is children’s confidence in their correct and incorrect memories.

Uncertainty Monitoring. An improvement in metamemory in general would presumably extend to expressions of confidence as well. The relevant studies of metamemory in children usually report mean confidence for correct responses versus mean confidence for incorrect responses to measure their proficiency in “uncertainty monitoring.” If children express significantly higher confidence for their correct answers compared to their incorrect answers, they are said to have developed uncertainty monitoring. Uncertainty monitoring has consistently been found in children older than five (Destan, Hembacher, Ghetti, & Roebers, 2014; Roderer & Roebers, 2010), and other studies have found that uncertainty monitoring for recognition memory in particular develops around the preschool ages, between three and five years old (Hembacher & Ghetti, 2014). 3-year-olds lack the ability to monitor their own uncertainty and therefore give similar confidence ratings for both correct and incorrect answers. 5-year-olds, on the other hand, give significantly higher confidence ratings for correct vs. incorrect answers, and thus are said to have developed uncertainty monitoring.

Children have also shown a clear evidence of uncertainty monitoring in studies of recall. In the recognition memory studies discussed above, participants were presented with the actual items that were encoded along with some new items, and they were asked which items are old and which are new. In a recall study, participants are instead simply asked to state everything they remember about what was encoded (free recall) or they are cued using a certain word or phrase to help them remember what they encoded (cued recall). Many studies have shown uncertainty monitoring in recall for children as young as seven years old in both basic and eyewitness tasks (Roebers, 2002; Roderer & Roebers, 2003; Roebers, von der Linden, Schneider, & Howie, 2007).

The existence of uncertainty monitoring in children suggests that confidence – even in children – can be informative with respect to accuracy. Nevertheless, studies of reality monitoring generally do not directly characterize the confidence-accuracy relationship in children. For example, even if a child is able to differentiate when they are correct from when they are incorrect using confidence, it does not speak to the overall accuracy of their memory (e.g., it does not mean that high-confidence accuracy is very high). This is not to say that these measures do not have some level of redundancy. More likely than not, children who show uncertainty monitoring will also show a clear confidence-accuracy relationship. But how accurate, exactly, is a decision made with high confidence? A question like that is relevant to the legal system, but a different analysis is needed to answer it.

The Confidence-Accuracy Relationship in Children. Though studies of uncertainty monitoring do not provide direct evidence for a strong confidence-accuracy relationship in children, it seems reasonable to suppose that as a child becomes better at monitoring their own uncertainty, they will also show improvements in their confidence-accuracy relationship. Indeed, without some capacity for uncertainty monitoring, a meaningful confidence-accuracy relationship could not exist because children would be unable to use confidence as an outward expression of their metacognitive state. That is, the underlying mechanism by which children learn to use confidence is presumably rooted in their ability to understand when they are incorrect.

Children are often said to start out as "eternal optimists," (Mickes et al., 2011) where they are always certain their memories are correct and always express high confidence. Then, as the children get older, they begin to learn that sometimes their decisions based on their memories are incorrect, and from there they begin to appreciate that errors tend to occur when their memory strength is weak, and that uncertainty is warranted under those conditions. As a result of this understanding of uncertainty, children begin expressing confidence based on their interpretation of the memory strength signal generated by an item (e.g., on a recognition test), rather than simply always expressing high confidence.

Despite evidence of uncertainty monitoring in children as young as 5 years of age, eyewitness identification studies with children that have investigated the confidence-accuracy relationship with children have found no meaningful relationship in children as old as 11 (Keast et al., 2007). Thus, researchers studying basic developmental processes in children report clear evidence of uncertainty monitoring (the presumed basis for a strong confidence-accuracy relationship), and they have also provided a theoretical basis for the existence of a relationship between confidence and accuracy in children. By contrast, researchers studying the confidence-accuracy relationship for lineup decisions in children report that children's confidence ratings show little to no relationship with their accuracy. What explains this apparent discrepancy?

Conceivably, this dichotomy between basic and applied research is due to a fundamental difference between lineup tasks and of the kinds of recognition and recall memory tasks used in basic research. However, this is the same explanation that once seemed to explain a similar discrepancy in the adult literature prior to the introduction of calibration curves and confidence-accuracy characteristic analysis, which showed a much stronger confidence-accuracy relationship than previously thought. Thus, an alternative possibility is that, in children, the discrepancy between basic and applied research is also more apparent than real. In the adult literature, it turned out that the confidence measures that were commonly used masked the presence of a confidence-accuracy relationship for eyewitness tasks, mainly because filler identifications and/or non-choosers were included in the analysis. This might also be the case for eyewitness studies in children.

Missing from the literature thus far is a comprehensive overview of the confidence-accuracy relationship in children. What do the data typically look like – in both the basic and applied literatures – when accuracy is plotted as a function of confidence using CAC analysis?

Next, I address that question using data from basic developmental research, reanalyzing the originally reported data where necessary (e.g., when the data were reported in terms of uncertainty monitoring). I then consider in more detail what has been learned about the confidence-accuracy relationship in children from eyewitness identification research, re-plotting those data in terms of CAC analysis as well.

Developmental Research on the Confidence-Accuracy Relationship

Although only a few basic recognition memory studies in children collected explicit confidence ratings, those that have done so found a strong relationship between confidence and accuracy in children. For example, Berch and Evans (1973) tested 5- to 9-year-old children on a continuous recognition task. Children were shown a set of 90 cards with 2-digit numbers on them. There were 45 unique numbers, so each number appeared twice in the set. The set was divided into 3 blocks of 30 cards, with 15 “new” items and 15 “old” items appearing in each block. New items were numbers that had not been seen before in that block and old items were numbers that had been seen once previously. The children were asked to state whether each number shown was new or old, and to rate their confidence in each of their decisions (Sure vs. Not Sure). Berch and Evans (1973) analyzed their results using a probability function for old/new judgments as a function of confidence and found that “both the kindergartners and 3rd graders produced monotonic decreasing functions. For example, the lower the [subject’s] level of confidence in judging an item as old, the lower the probability of that item actually being old.” In this case, an “old” decision is analogous to being a “chooser” in an eyewitness identification paradigm. For 3rd-graders making an “old” decision, the confidence-accuracy relationship was strong (less so for kindergarteners). Their results are shown plotted as CAC in Figure 1.

In 2010, Wilkinson and colleagues compared typically developing children between 9 and 17 years old to children with Autism Spectrum Disorder (ASD) of the same age on an old/new facial recognition paradigm. During the learning phase, participants viewed 24 female faces sequentially. After delay, a memory test was given that consisted of 48 faces, 24 old (e.g. shown in the learning phase) and 24 new. The faces were presented one at a time, and for each, participants were asked whether or not they had seen the face in the learning phase. After each decision they were asked to rate their confidence. The results were analyzed using CAC analysis (i.e., proportion correct was plotted as a function of accuracy). The typically developing children showed a strong confidence-accuracy relationship, while the children with ASD showed no relationship. These results are shown in Figure 2.

Hiller and Weber (2013) tested 8 to 12-year-old children and adults on an associative word pair recognition paradigm, aiming to test for the presence of a confidence-accuracy relationship in children on a basic recognition memory task. Participants were shown 28 word pairs sequentially in the encoding phase, and after a delay were given a memory test also consisting of 28 word pairs. They had to recognize each word pair in the test as old or new, and after each decision rate their confidence using a confidence scale that ranged from 50 (guessing) to 100 (certain). They analyzed their results using logistic mixed-effects modeling. This is similar to confidence-accuracy characteristic analysis, except that predicted log odds of a recognition decision being correct or incorrect is plotted as a function of confidence rather than percent correct being plotted as a function of confidence. They found that children showed a strong confidence-accuracy relationship for this paradigm, though not as strong as adults. Their results are shown plotted as CAC in Figure 3, with predicted log odds converted to percent correct. Because this information was only shown in a plot, nearly exact values were estimated using WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/). When children made "yes" decisions (i.e., on trials in which they were "choosers"), their accuracy was approximately 30% correct for low-confidence decisions, and their accuracy was 86% correct for high-confidence decisions. Thus, the confidence-accuracy relationship was strong.

Adults have been expressing confidence for significantly longer than children, so it is perhaps not surprising that they are able to give more precise confidence judgments based on their memory strength when compared to children. However, as is clearly apparent in Figure 3, this in no way means that children lack this ability or that their confidence does not hold considerable information value. It instead likely means that children still have not fully developed adult-level metacognitive abilities. Despite the strong confidence-accuracy relationship they found in children, Hiller and Weber (2013) concluded that their results do not contradict the seemingly well-established *lack* of a confidence-accuracy relationship on eyewitness identification tasks. The fact their results clearly indicate a strong confidence-accuracy relationship in children was attributed by them to the fact that they used a basic list-memory procedure, not an eyewitness identification procedure (under the belief that the types of tasks are fundamentally different).

Hembacher & Ghetti (2014) tested uncertainty monitoring in children. 3- to 5-year-old children completed an object recognition task using two-alternative forced-choice trials. During the learning phase the children viewed 30 line drawings of common objects. After a delay, they completed a recognition test in which they had to decide which of two line drawings was one they had seen in the learning phase. After making their decision, their confidence was measured on a 3-point scale. They concluded that 3-year-olds were unable to monitor their own uncertainty, because their mean confidence was the same regardless of if their answers were correct or incorrect. 5-year-olds, on the other hand, showed significantly higher mean confidence for correct answers compared to incorrect answers, and thus it was concluded that they were able to monitor their uncertainty. 4-year-olds fell somewhere in between, showing a developmental trajectory for uncertainty monitoring across this age group. I obtained the data for this study through the Open Science Framework, and was able to also determine if a similar developmental trajectory existed for the confidence-accuracy relationship. Perhaps not surprisingly, when calculated using CAC, the confidence-accuracy relationship matched the developmental trajectory of uncertainty monitoring. The CAC data are pictured in Figure 4. The CAC data reveal a strikingly similar relationship to the uncertainty monitoring data. 3-year-olds show virtually no confidence-accuracy relationship, with both low-confidence and high-confidence responses resulting in similar levels of overall accuracy (both around 85%). 4-year-olds show a moderate confidence-accuracy relationship, with low-confidence responses being approximately 70% correct and high-confidence responses being approximately 86% correct. 5-year-olds showed a strong confidence-accuracy relationship with low-confidence responses being approximately 63% correct and high-confidence responses being approximately 92% correct. This is evidence that at least for basic recognition memory tasks, children’s confidence is a strong indicator of their overall memory accuracy even as young as 5 years old. It is also evidence that the confidence-accuracy relationship and uncertainty monitoring may be different representations of the same metacognitive process, or at least that these two measures follow a similar developmental trajectory.

Shing et al. (2009) compared the confidence-accuracy relationship between children (ages 10 to 12), teenagers (13 to 15), adults (ages 20 to 25), and older adults (ages 70 to 75). They were interested in testing which group is the most likely to commit high-confidence errors.

A high-confidence error occurs when a person is certain that they remember something that in reality they have never seen before. To test this, they used a word-pair associative recognition task. In this task, a word in the participant's native language is paired either with a second native language word (for the control group) or with a foreign language word (for the experimental group. They were shown a list of 45 pairs that were either both native words or one native and one foreign, and after a delay given a test where they were shown 60 pairs and asked to recognize them as being old or new, rating their confidence in each decision after it is made. Of interest in this study was whether their child and teenage participants showed a confidence-accuracy relationship in addition to the adults and older adults. This paper reported the "percent 'sure' responses" as a function of the hit and false alarm rates for each condition. The "percent 'sure' responses" measure represents what proportion of the participants' responses were made with high confidence. Because this information was only shown in a plot, nearly exact values were estimated using WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/). To convert these data into CAC, the first step is to multiply the percent "sure" hits by the overall hit rate, thereby providing the high-confidence hit rate. This process was then repeated for "not sure" responses.

The average CAC plots for each age group (averaged across the pre-and post-strategy conditions of the experiment) are shown in Figure 5. For all age groups, the confidence-accuracy relationship improved for the post-strategy condition. On average, children showed a strong confidence-accuracy relationship, though not as strong as young adults. For high-confidence responses they were approximately 87% accurate, and for low-confidence responses they were approximately 61% accurate. It also could be evidence that although children may spontaneously show an adult-like confidence-accuracy relationship, they may be able to improve if they are given a strategy to employ. Additionally, children showed significantly better overall accuracy compared to older adults. Older adults still showed a confidence-accuracy relationship, but they were generally overconfident. On average, for high-confidence responses they were only 80% accurate. However, in the post-strategy condition they were 86% accurate for high-confidence responses. This is further evidence for elaborative imagery improving the confidence-accuracy relationship.

Fandakova and colleagues (2012) tested children (ages 10-12), adults (ages 20-27), and older adults (ages 68-76) on a repeated continuous recognition task of word pairs. They were testing how each age group's rate of false alarms, particularly high-confidence false alarms, changed from run to run. In this case, all age groups went through three runs. They also looked at how the false alarm rate differed for lures that were completely new word pairs versus lures that were rearranged target word pairs (e.g. Tree-Duck would become Duck-Tree). This paper reported the overall proportion of hits and false alarms for each run, as well as the proportion "sure" hits and proportion "sure" false alarms. Because this information was only shown in a plot, nearly exact values were again estimated using WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/). Before calculating CAC, we averaged the hit and false alarm rates across the three runs. To calculate the proportion of "not sure" hits and false alarms, we assumed all responses not considered "sure" were considered "not sure" or low-confidence, and thus subtracted the proportion "sure" hits and false alarms from the overall proportion of hits and false alarms. The average CAC plots for each group are shown in Figure 6.

As in Shing et al. (2009), young adults performed the best out of the three age groups, followed by children and then older adults. Young adults were 89% correct for "sure" responses, and 61% correct for "not sure" responses. Children were 79% correct for "sure" responses, and 60% correct for "not sure" responses. Older adults were 71% correct for "sure" responses, and

54% correct for “not sure” responses. The CAC for all groups in this study was not as impressive as in some others, and children did not perform as well compared to young adults as in other cases, but there was still a strong confidence-accuracy relationship for all three groups. There was apparently something about this task that caused performance to be less impressive than usual.

Eyewitness Research on the Confidence-Accuracy Relationship in Children

While many revelations about adults as eyewitnesses have been made in the past few years, research on children as eyewitnesses has fallen behind. For instance, to this day, the overwhelming majority of research on the confidence-accuracy relationship for lineup decisions in children has been conducted using the point biserial correlation coefficient, particularly in the eyewitness setting (Parker, Haverfield, & Baker-Thomas, 1986; Parker & Carranza, 1989; Parker & Ryan, 1993). There are also some studies that use calibration curves, but no studies that use CAC analysis, despite the general consensus now shown in the adult literature that it is an informative way to measure the confidence-accuracy relationship for both basic memory studies and eyewitness memory studies.

Studies using the point biserial correlation coefficient to measure the confidence-accuracy relationship for child eyewitnesses generally reported only a weak to moderate relationship. Parker and colleagues (1986) tested children (with a mean age of 8 years) and adults (with a mean age of 24 years) on a simultaneous lineup task. In a simultaneous lineup task, the suspect and five or more physically similar “filler” faces are presented simultaneously in a photo array. The eyewitness is told that the perpetrator may or may not be in the lineup and to identify the perpetrator if they are present. Once the identification has been made, they are asked to rate their confidence. Unusually, in this study, they reported a point-biserial correlation coefficient of *r* = 0.47 for children and *r* = 0.34 for adults, concluding that “confidence was strongly related to accuracy.” However, this was deemed to be a result of the experiment using only target-present and not target-absent lineups, and thus they concluded that this was not necessarily showing evidence for a confidence-accuracy relationship in child eyewitnesses more generally. Parker & Carranza (1989) also tested children (with a mean age of 9 years) and adults (with a mean age of 21 years) on a simultaneous lineup task. In this study, they reported a point-biserial correlation coefficient of *r* = 0.00 for the initial identification and confidence and concluded that “the eyewitness accuracy-confidence relationship once more appears to be of little use forensically.” This conclusion encompasses both adults and children. Parker & Ryan (1993) tested children and adults on simultaneous and sequential lineup tasks. A sequential lineup task, like a simultaneous lineup task, involves showing a participant six or more faces, but in the sequential task, the faces are presented one at a time sequentially rather than simultaneously. They measured the point biserial correlation coefficient across conditions, and found overall *r* = 0.22, concluding “confidence/accuracy correlations in the present study revealed no overall correlations.”

Studies with children using calibration curves have found generally the same result as studies using the point biserial correlation coefficient. There are two notable studies that have employed this analysis on lineup data from children. Both tested many children, and the results of these studies appeared to be so definitive that they seem to have been taken as the last word on the subject. Brewer and Day (2005) tested 240 8 to 10-year-old children and 159 14- to 17-year-old adolescents on a simultaneous lineup task. They concluded: “In sum, it appears that the confidence in the identification decision by itself will not provide a useful accuracy marker in children” (p. 126) However, they also concluded that the adolescents did show a strong confidence-accuracy relationship. Thus, based on their conclusions, at some point between the ages of 10 and 14, children seem to go from being “extremely overconfident” to showing adult-like calibration. Additionally, Keast et al. (2007) tested 1,415 10 to 14-year-old children on an identical simultaneous lineup task across two experiments. In this study, the same crime video was used as in Brewer and Day (2005), which depicted the theft of a credit card at a restaurant, but participants were asked to identify both the thief and the waiter in the video using lineups, unlike the previous study where they only identified the thief. They concluded, across both identifications and experiments that, "in contrast to adults, children’s identification confidence provides no useful guide for investigators about the likely guilt or innocence of a suspect" (p. 286).

It is possible to convert these calibration curves into CAC. Doing this may reveal that the calibration curves have been masking a stronger confidence-accuracy relationship due to their inclusion of filler IDs in the accuracy score. As noted earlier, although filler IDs are errors, they are not relevant to cases in which a witness has identified a suspect (i.e., they are not relevant to a trial involving eyewitness identification evidence). To convert their reported calibration curve into CAC, the exact calibration data was estimated using the figures provided in the original papers. Estimated accuracy scores (e.g. proportion correct for each level of confidence) were calculated by inputting the points from the calibration plots into WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/). Next, this aggregate accuracy score was converted to a Suspect ID accuracy score by taking the reported accuracy score for a given level of confidence, $a1$, converting to an odds score, o, where $o = a1 / (1 - a1)$, and then computing suspect ID accuracy, 100% * $a2$, using the formula $a2 = o / (o + 1/n)$, where n = lineup size (Wixted & Wells, 2017).

In both cases, once the data are shown in a CAC plot, a strong relationship between confidence and accuracy is clearly apparent. CAC plots for Brewer and Day (2005) are shown in Figure 7 and CAC plots for Keast et al. (2007) are shown in Figure 8. In Brewer and Day's (2005) study, adolescents showed an adult-like confidence-accuracy relationship for the simultaneous lineup task, with low-confidence IDs being approximately 83% accurate and high-confidence IDs being approximately 99% accurate. Thus, our reanalysis confirms Brewer & Day (2005)'s initial conclusion for adolescents showing adult-like confidence-accuracy relationship, but does not confirm their claim that children ages 8 to 10 are "extremely overconfident." For those children, low-confidence accuracy was 44% correct, whereas high-confidence accuracy was 88% correct. According to the CAC plot shown in Figure 7, children's confidence was highly indicative of accuracy, and high-confidence accuracy, while exhibiting some degree of overconfidence (i.e., 88% is clearly less than 100% correct), is reasonably accurate nonetheless. The same basic conclusion follows from the CAC plot from Keast et al. (2007) shown in Figure 8, where low-confidence accuracy was 47% correct and high-confidence accuracy was 86% correct.

Conclusion

Even though it is not as strong as the confidence-accuracy relationship in adults, children are still reasonably accurate (~85% correct) when expressing high confidence about a lineup identification. In addition, when children express low confidence, that value decreases to about 50% accurate. This means that when children express low confidence, they are aware of the fact that the memory signal associated with the identified individual is weak, and that they are most likely not making an accurate identification. This information is just as important in a court of law as a high confidence identification, because it means that the child may not actually know who the suspect is at the time of the initial identification (despite what might be recorded by the police as a “positive ID”). By the time they are brought to testify at a trial, however, their memory may have become contaminated from seeing the person they identified with low confidence being depicted as the person who committed the crime, and the suspect’s face may be associated with a higher familiarity value after having seen it multiple times. All of this could lead to them expressing high confidence in court, thereby providing what seems like evidence of the suspect’s guilt, even when they initially made a highly error-prone low-confidence ID.

The key take-home message is that the confidence-accuracy relationship in children tested using an eyewitness identification procedure is essentially the opposite of what researchers and the legal system believe to have been definitively established by scientific research. Instead of being unrelated to accuracy, a child’s expression of confidence provides considerable information about the likely accuracy of a suspect ID. Not only does the present reanalysis make that important point, it also reconciles what has previously seemed to be a contradiction between what has been learned about the confidence-accuracy relationship in children in the basic developmental literature and what has been learned about that relationship in the eyewitness identification literature. As it turns out, just as in the adult literature, the contradiction was more apparent than real. Whether tested using a basic list-memory paradigm or an eyewitness identification paradigm, the confidence-accuracy relation in children older than 8 years of age (and perhaps even younger) is clearly strong.

Chapter 2, in full, is currently being prepared for submission for of the material. Killeen, Isabella M. The dissertation author was the primary investigator and author of this material.

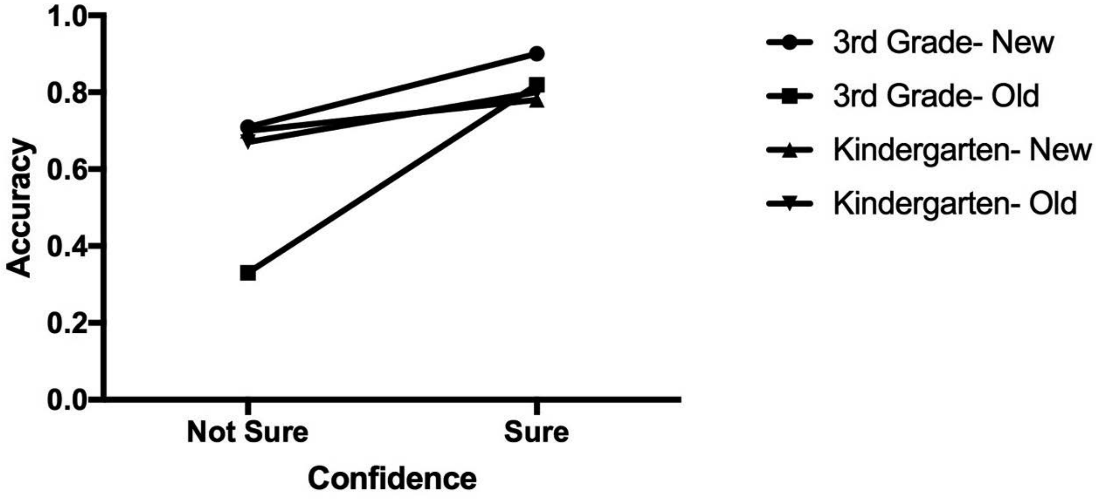

Figure 2.1. Confidence-Accuracy Characteristic for kindergartener's and third-grader's old and new recognition judgments (Berch & Evans, 1973)

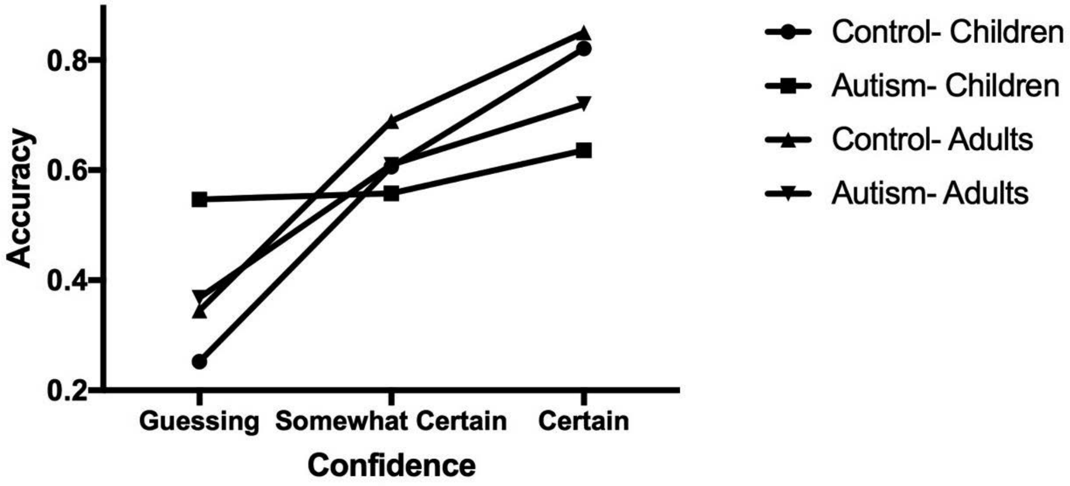

Figure 2.2. Confidence-Accuracy Characteristic for children with Autism Spectrum Disorder and typically developing children (Wilkinson et al., 2010).

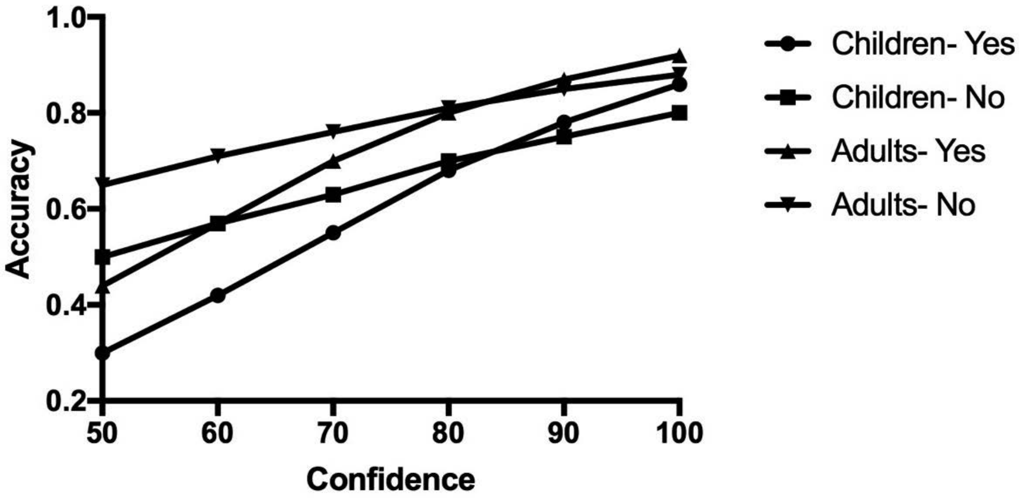

Figure 2.3. Confidence-Accuracy Characteristic for adults' and children's old and new recognition judgments (Hiller & Weber, 2013)

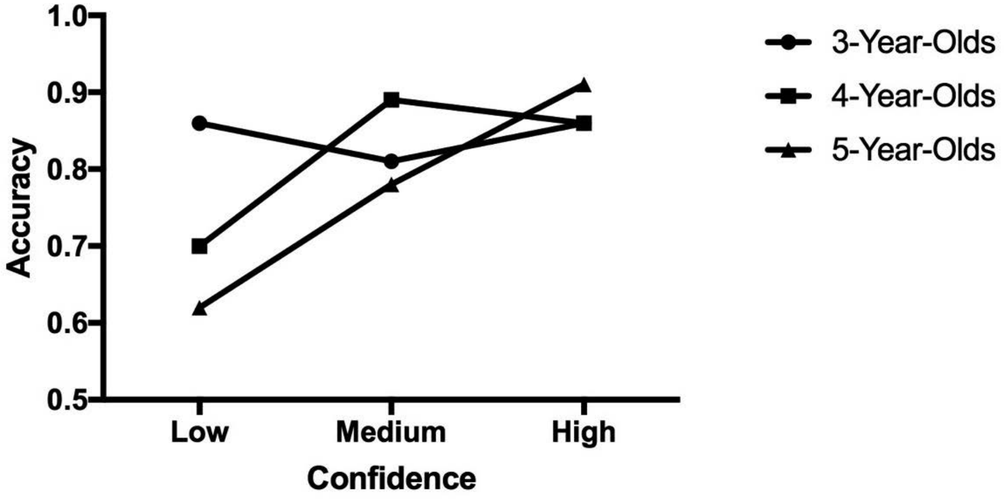

Figure 2.4. Confidence-Accuracy Characteristic for 2-alternative forced choice object recognition in 3, 4, and 5-year-olds (Hembacher & Ghetti, 2014)

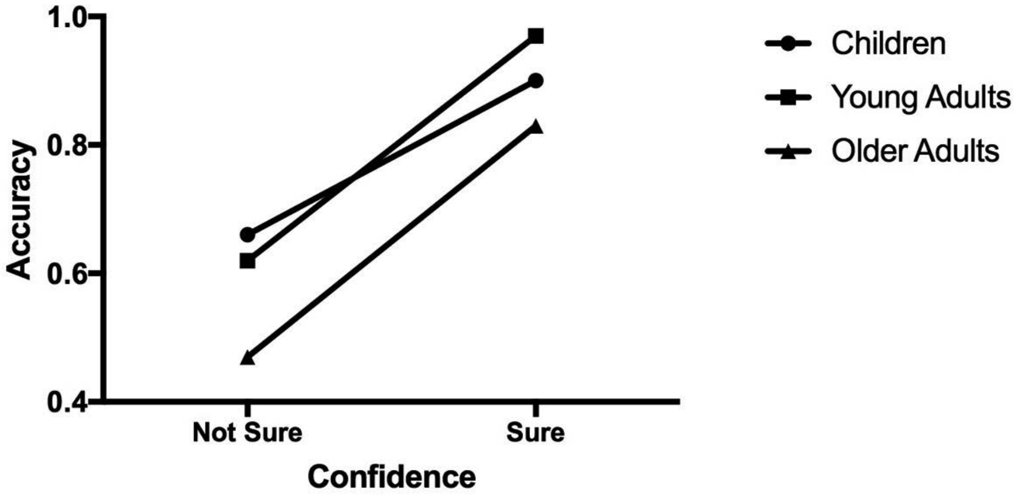

Figure 2.5. The average Confidence-Accuracy Characteristic for children, young adults, and older adults (Shing et al., 2009)

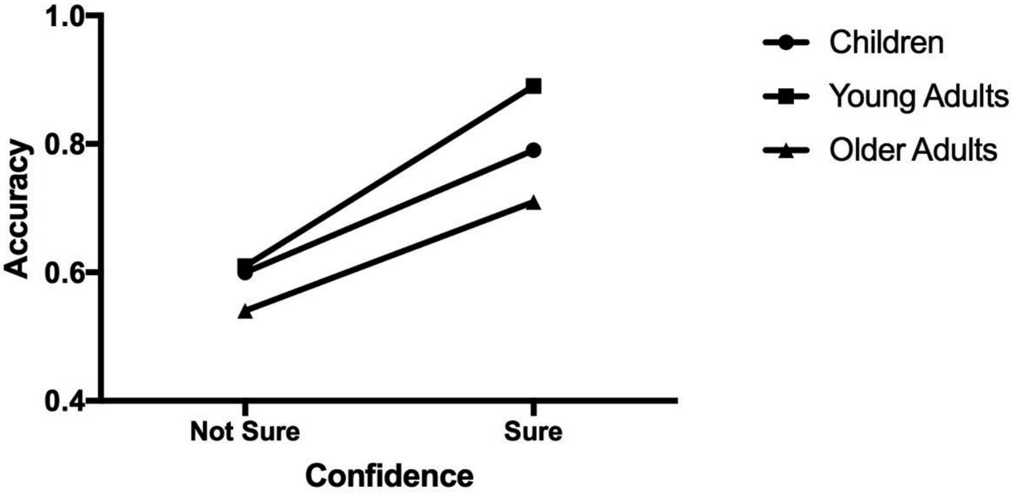

Figure 2.6. The average Confidence-Accuracy Characteristic for children, young adults, and older adults (Fandakova et al., 2013)

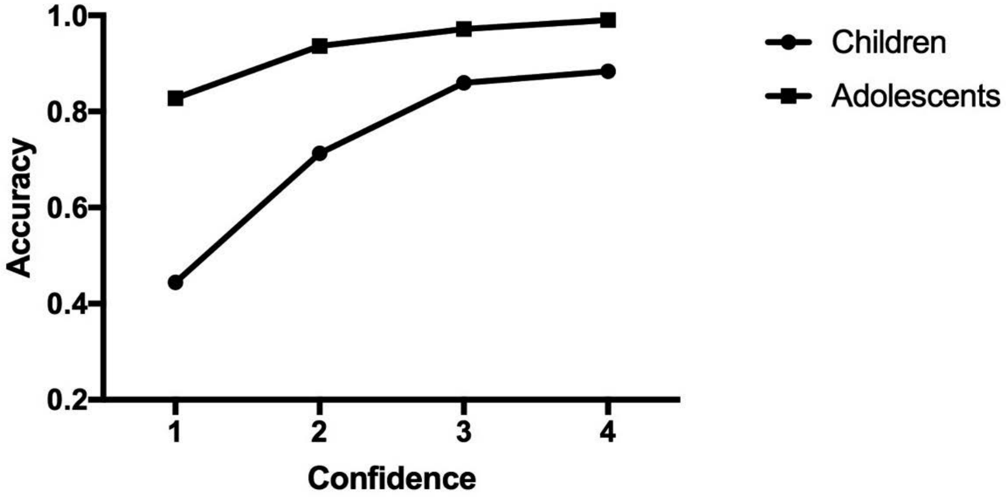

Figure 2.7. Confidence-Accuracy Characteristic for Suspect IDs in children and adolescents (Brewer & Day, 2005)

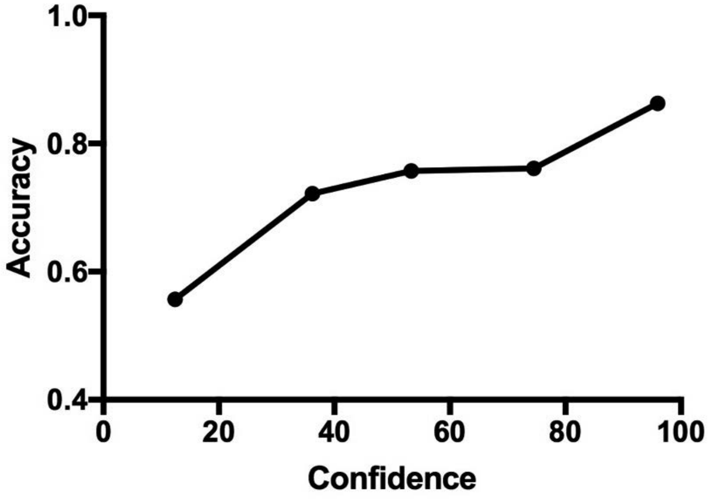

Figure 2.8. Aggregate Confidence-Accuracy Characteristic for Suspect IDs in children (Keast et al., 2007)

References

Ackil, J.K. & Zaragoza, M.S. (1995). Developmental differences in eyewitness suggestibility and memory for source. *Journal of Experimental Child Psychology, 60*(1), 57-83.

Berch, D.B. & Evans, R.C. (1973). Decision processes in children's recognition memory. *Journal of Experimental Psychology, 16*, 148-164.

Brewer, N. & Day, K. (2005). The confidence-accuracy and decision latency-accuracy relationships in children's eyewitness identification. *Psychiatry, Psychology, and Law, 12*(1), 119-128.

Brewer, N. & Wells, G.L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*(1), 11-30.

Ceci, S.J. & Bruck, M. Suggestibility of the child eyewitness: A historical review and synthesis. *Psychological Bulletin, 113*(3), 403-439.

Deffenbacher, K. A., Bornstein, B. H., & Penrod, S. D. (2006). Mugshot exposure effects: Retroactive interference, mugshot commitment, source confusion, and unconscious transference. *Law and Human Behavior, 30,* 287–307.

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C.M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology, 126*, 213-228.

Fandakova, Y., Shing, Y.L., & Lindenberger, U. (2013). Differences in binding and monitoring mechanisms contribute to lifetime differences in false memory. *Developmental Psychology, 49*(10), 1822-1832.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1995). Young children's knowledge about thinking. *Monographs of the Society for Research in Child Development, 60*, 1–96.

Hembacher, E. & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*(9), 1768-1776.

Hiller, R.M. & Weber, N. (2013). A comparison of adults' and children's metacognition for yes/no recognition decisions. *Journal of Applied Research in Memory and Cognition, 2*, 185-191.

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

Keast, A., Brewer, N. & Wells, G.L. (2007). Children's metacognitive judgments in an eyewitness identification task. *Journal of Experimental Child Psychology, 97*, 286-314.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

Koriat, A., Goldsmith, M., Schneider, W., & Nakash-Dura, M. (2001). The credibility of children's testimony: Can children control the accuracy of their memory reports? *Journal of Experimental Child Psychology, 79*, 405–437.

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.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness. *Journal of Experimental Psychology: Applied, 18*, 361-376.

Parker, J.F. & Carranza, L.E. (1989). Eyewitness testimony of children in target-present and target-absent lineups. *Law and Human Behavior, 13*, 133-149.

Parker, J.F., Haverfield, E., & Baker-Thomas, S. (1986) Eyewitness testimony of children. *Journal of Applied Social Psychology, 16*(4), 287-302.

Parker, J.F. & Ryan, V. (1993). An attempt to reduce guessing behavior in children's and adults' eyewitness identifications. *Law and Human Behavior, 17*, 11-26.

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law, 1*, 817–845.

Plude, D. J., Nelson, T. O., & Scholnick, E. K. (1998). Analytical research on developmental aspects of metamemory. *European Journal of Psychology of Education, 13*, 29–42.

Powell, M., Garry, M., & Brewer, N. (2013) Eyewitness testimony. In I. Freckleton & H. Selby (Eds.), *Expert Evidence: Law, Practice, Procedure, and Advocacy, Volume V*. Sydney, Australia: Thomson Reuters.

Roderer, T. & Roebers, C.M. (2003). Confidence judgments in event recall: Developmental progression in the impact of question format. *Journal of Experimental Child Psychology, 85*, 352-371.

Roderer, T. & Roebers, C.M. (2010). Explicit and implicit confidence judgments and developmental differences in metamemory: An eye-tracking approach. *Metacognition & Learning, 5*, 229-250.

Roebers, C.M. (2002). Confidence judgments in children's and adults' event recall and suggestibility. *Developmental Psychology, 38*(6), 1052-1067.

Roebers, C.M. (2014). Children's deliberate memory development: The contribution of strategies and metacognitive processes. In P.J. Bauer & R. Fivush (Eds.), *The Wiley Handbook on the Development of Children's Memory, Volume I/II* (pp. 865-894). Chichester, UK: John Wiley & Sons Ltd.

Roebers, C.M., von der Linden, N., Schneider, W., & Howie, P. (2007). Children's metamemorial judgments in an event recall task. *Journal of Experimental Child Psychology, 97*, 117-137.

Sanders, J.I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron, 90*, 499-506.

Shing, Y.L., Werkle-Bergner, M., Li, S., & Lindenberger, U. (2009). Committing memory errors with high confidence: Older adults do but children don't. *Memory, 17*(2). 169-179.

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

Wells, G. L., & Olson, E. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied, 8*, 155-167.

Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). New York, NY: Cambridge University Press.

Wilkinson, D.A., Best, C.A., Minshew, N.J., & Strauss, M.S. (2010) Memory awareness for faces in individuals with autism. *Journal of Autism and Developmental Disorders, 40*, 1371-1377.

Wixted, J.T., Mickes, L., Clark, S.E., Gronlund, S.D., & Roediger H.L. (2015). Initial eyewitness confidence reliably predicts identification accuracy. *American Psychologist, 70*, 515-526.

Wixted, J.T., Mickes, L. & Fisher, R. F. (2018). Rethinking the reliability of eyewitness memory. *Perspectives on Psychological Science, 13*, 324-335.

Wixted, J.T. & Wells, G.L. (2017). The Relationship between Eyewitness Confidence and Identification Accuracy: A New Synthesis. *Psychological Science in the Public Interest*, *18*, 10-65.

CHAPTER 3

The Development of a Recognition-Memory Confidence-Accuracy Relationship for Objects and Faces in Preschoolers

Abstract

Recent research has shown that preschoolers possess *uncertainty monitoring,* meaning they express higher confidence about their correct answers compared to their incorrect answers. Additionally, preschoolers have been shown to express higher confidence about their accurate memories (e.g. when they correctly identify an item they have encountered previously) compared to their incorrect memories (e.g. when they incorrectly state they have encountered a novel item before). A similar but unique measure that has never been tested in children of this age is the proportion of their memories that are accurate for a given level of confidence. This is referred to as the *confidence-accuracy relationship.* In this study, we tested for both uncertainty monitoring and a confidence-accuracy relationship in 3- 4- and 5-year-olds. Our results show that 4- and 5-year-olds show uncertainty monitoring and a confidence-accuracy relationship for their memory of objects. However, only 4-year-olds showed uncertainty monitoring and none of the ages showed a confidence-accuracy relationship for faces. We conclude that this is due to large differences in overall memory for objects compared to faces. All ages showed performance that was well above chance for their memory of objects, but all ages showed performance that was not above or barely above chance for their memory of faces. We discuss possible explanations for this disparity between stimulus sets.

Introduction

Adults show a strong relationship between their expressed confidence and the accuracy of their recognition memory decisions. This has typically been shown in old/new recognition paradigms where a participant is asked to recognize whether a picture or word on a test list was presented on a previously encoded list (e.g. it was recently encountered and is "old") or not (e.g. it was not recently encountered and is "new") (e.g. Hiller & Weber, 2013; Mickes, Hwe, Wais, & Wixted, 2011; Tekin & Roediger, 2017). Adults show a strong confidence-accuracy relationship not only in basic memory research but also in eyewitness memory research, where a participant watches a video of a mock crime and is then asked to identify the person from the video (e.g. the "culprit") out of a lineup of faces presented either simultaneously (e.g. 6-8 faces are shown at once in an array) or sequentially (e.g. 6-8 faces are shown one at a time consecutively) (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted & Wells, 2017). Typically, for adults, confidence is rated on a numeric scale with 5 to 20 points. Tekin & Roediger (2017) showed that the amount of points in the confidence scale does not appreciably affect the relationship between expressed confidence and memory accuracy for old/new recognition decisions. The relationship between confidence and accuracy is measured either as proportion correct for items from a list or proportion correct for suspect identifications from a lineup – as a function of confidence. This straightforward and easy-to-interpret plot is known as the confidence-accuracy characteristic (CAC).

Although only a few basic recognition memory studies in children have collected explicit confidence ratings, those that have found a strong relationship between confidence and accuracy. Similar to adults, this result has typically been observed in old/new recognition paradigms. Both 5- to 9-year-old (Berch & Evans, 1973) and 8- to 12-year-old (Hiller & Weber, 2013) children have shown a strong relationship between expressed confidence and memory accuracy for old/new recognition decisions. If children as young as five already show a strong relationship between confidence and accuracy, when does this relationship develop? Young children have been referred to as "eternal optimists" with little understanding of uncertainty (Mickes et al., 2011), so children must develop this understanding at some point in early childhood. Only one study has examined the relationship between confidence and recognition memory accuracy in children younger than five, though other studies have concluded that children show metacognitive monitoring as early as two to three years old (Lyons & Ghetti, 2011; Marazita & Merriman, 2004). Hembacher & Ghetti (2014) tested 3- to 5-year-old children using a two-alternative forced choice paradigm. In this paradigm, participants are asked which of two simultaneously presented items was seen on a previously encoded list. The stimuli used were line drawings of common objects. They found that 5-year-olds showed significantly higher confidence for their correct answers (e.g. they correctly identified which of two items had been shown to them previously) compared to their incorrect answers (e.g. the incorrectly selected an item they had not seen previously), and interpreted this as 5-year-olds showing *uncertainty monitoring.*

Uncertainty monitoring is defined as the process by which a learner considers whether a decision made under unreliable conditions is likely to be correct (Koriat & Goldsmith, 1996). It is described as introspective because uncertainty monitoring requires a metacognitive awareness of the relative strength of different cognitive signals, whether they be perceptual, memorial, or linguistic. An awareness of the likelihood of error in a decision is also what underlies the confidence-accuracy relationship, so it is likely that this measure and CAC are two ways of showing the same effect. Additionally, age-related improvements in metamemory, including uncertainty monitoring, are associated with improvements in memory accuracy across many tasks (Flavell, 1989; Koriat, Goldsmith, Schneider, & Nakash-Dura, 2001; Plude, Nelson, & Scholnick, 1998).

Hembacher & Ghetti (2014) also showed that 4-year-olds expressed significantly higher confidence for their correct answers compared to their incorrect answers, but only when the item they correctly identified had been shown to them multiple times. When an item is presented multiple times, the strength of the memory created for that item increases compared to when an item is only presented once. 3-year-olds showed no differences in average confidence between correct and incorrect answers. These results appear to show the development of a confidence-accuracy relationship during the preschool years. However, this is the only study of its kind. In our study, we first look to replicate the results of Hembacher & Ghetti (2014) in a very similar paradigm. We used the same stimulus set (line drawings of objects) and the same novel confidence scale that they developed. In addition to replicating their results, using their paradigm allows us to compare their analysis technique (average confidence for correct vs. incorrect answers) to CAC analysis. If these two analyses measure the same effect, then the developmental trajectory seen in Hembacher & Ghetti (2014)'s measure should also be evident in CAC analysis.

In addition to using the same object stimuli as Hembacher & Ghetti (2014), we expanded our stimulus set to include face stimuli. As mentioned previously, adults show a strong confidence-accuracy relationship for words and pictures in old/new recognition, as well as faces in eyewitness memory lineup procedures. Hembacher & Ghetti (2014) showed that preschool-aged children may show a confidence-accuracy relationship for object recognition, but no previous study has tested memory for faces in this age group. In 2010, Wilkinson and colleagues concluded that typically developing 9- to 17-year-old children show a significant relationship between their confidence and the accuracy of their memories for face stimuli presented in an old/new recognition paradigm, which provides some evidence that even for older children the confidence-accuracy relationship exists independent of the type of stimulus presented. If this is also true for preschool-aged children, then they will show similar performance for face stimuli compared to object stimuli. If Hembacher & Ghetti (2014)'s analysis and CAC are measuring the same effect, we expect to see a similar developmental trajectory for confidence-accuracy relationship across both analyses for object stimuli and face stimuli.

Experiment 1

Methods

Participants

Fourteen 3-year-olds (M= 39.1 months, SD = 2.39, fourteen 4-year-olds (M= 54.66 months, SD = 2.35), and fourteen 5-year-olds (M= 65.12 months, SD = 2.77) participated in this study for a total of 42 children. An additional 8 participants were excluded because they did not complete the entire study, 1 participant due to caretaker interference, and 3 participants because they either failed an attention check or showed a complete bias for answers presented on one side of the screen. Participants were recruited from local preschools and the local science museum in a primarily urban setting. While specific demographic information was not collected from individual participants, demographics of the recruitment locations (taken from US Census Bureau data) suggest the participants were predominately white (44.5%) and middle-class (median household income of $73,900).

Materials

The study was completed on an iPad using Qualtrics survey software. Elements of the materials and paradigm were modeled after the methods used by Hembacher & Ghetti (2014).

Stimuli were 60 line drawings selected from Cycowicz, Friedman, Rothstein, & Snodgrass (1997). Each child saw 30 of the line drawings at encoding, with the 30 not shown at encoding serving as lures during the retrieval phase. Each of the encoded images (targets) was matched with a perceptually similar lure at retrieval, with the encoded images and lures counterbalanced across two groups. The 3-point confidence scale used in this study was the same scale used by Hembacher & Ghetti (2014).

Procedure

Encoding Period. Children were tested one-on-one with the experimenter and were told to watch the iPad screen very carefully because a lot of pictures were going to appear on the screen. They were told to keep watching until the screen until no more pictures appeared. The experimenter would confirm that the child was ready to begin the study, then would start the program and the images would start to appear. The 30 line drawings were presented sequentially for 2,000 milliseconds each. The order that the line drawings appeared in was randomized across participants. Every line drawing was shown only once.

Familiarization Period. After the child had seen all 30 images, they were instructed that they would be shown two images at a time. The experimenter would then display two images, one from the encoding period matched with a perceptually similar lure. The children were told that they had seen one of the images before and the other one they had not, and were asked to touch the picture they had seen before. Once the child had made their selection they were introduced to the confidence scale. They were instructed to touch the image that represented how sure they were (i.e., “not sure,” “a little bit sure,” or “very sure”). To ensure children understood the task, they were asked, “Which one do you point to when you’re [very, a little bit, not] sure?”

for all three levels of confidence. If children were unable to complete this task, their data was excluded from analysis.

Retrieval Period. The child was then informed that they would answer those same two questions many more times. Each child was asked the question for the subsequent 29 line drawings shown at encoding, each paired with a perceptually similar lure, and to rate their confidence on the three point scale for each decision. The side of the screen the target was presented on was randomized across trials. If children selected the same side of the screen for every trial regardless of the location of the target, their data were excluded from analysis.

Results

Figure 1 shows the uncertainty monitoring results. To examine the effects of age and confidence on memory accuracy, a two-way ANOVA was performed. Results showed a main effect of age, F(1,1154)=27.65, p <0.001 and a main effect of confidence level, F(1,1154)=19.18, p < 0.001. There was also a significant interaction between age and confidence, F(1,1154)=9.447, p = 0.002. Specifically, 3-year-olds did not show a significant difference in average confidence between their correct and incorrect answers, t(327.8) = -0.07, p = 0.94 *(ns),* while 4-year-olds and 5-year-olds showed significantly higher confidence for their correct answers compared to their incorrect answers, with t(178.52) = 5.01, p < 0.001 and t(90.43) = 3.50, p < 0.001, respectively.

Results were then analyzed using a confidence-accuracy characteristic (CAC) plot, which is shown in Figure 2. This plots proportion correct on the y-axis as a function of confidence level (not sure, a little bit sure, very sure) on the x-axis. It is a measure commonly used in adults for calculating the relationship between one’s expressed confidence and the accuracy of their memory decisions. It is simply a reconfiguration of the previously stated data. Computing participants' average confidence for correct and incorrect answers (as in uncertainty monitoring) answers the following question: given that we know the accuracy of a participant's decision, what level of confidence are they likely to express? A CAC plot, on the other hand, answers a different question: given a certain level of expressed confidence, what is the likelihood that a participant's decision was correct? This measure is often used in the eyewitness memory domain, because in the real world you often only have the confidence level an eyewitness provides, and have to estimate the likely accuracy of their decision based on that level of confidence. Participants were divided by age for this analysis, and the results are shown in Figure 2. 3-year-olds showed no relationship between their confidence and their memory accuracy. Their low confidence decisions were 63% accurate and their high confidence decisions were 66% accurate. 4-year-olds showed a much stronger relationship between their confidence and their accuracy. Their low confidence decisions were 39% correct (not significantly below chance performance of 50%) and their high confidence decisions were 76% correct. 5-year-olds also showed a relatively strong relationship between their confidence and their memory accuracy. Their low confidence decisions were 58% correct and their high confidence decisions were 86% correct. These results suggest that the confidence-accuracy relationship for object stimuli develops between 3 and 5 years old.

Experiment 2

Methods

Participants

Fourteen 3-year-olds (M= 42.42 months, SD = 3.44), fourteen 4-year-olds (M= 52.59 months, SD = 3.05), and fourteen 5-year-olds (M= 62.49 months, SD = 2.47) participated in this study for a total of 42 children. An additional 5 participants were excluded because they did not complete the entire study and 1 participant because they showed a complete bias for answers presented on one side of the screen. Participants were recruited from local preschools and the local science museum in a primarily urban setting. While specific demographic information was not collected from individual participants, demographics of the recruitment locations suggest the participants were predominately white (44.5%) and middle-class (median household income of $73,900).

Materials

Stimuli were 64 faces from the Chicago Face Database (faculty.chicagobooth.edu/bernd.wittenbrink/cfd/index.html). Four races of faces were used: White/Caucasian, Hispanic/Latino, Black/African American, and Asian. There were 16 faces of each race. Half of the faces were male and half were female. All of the faces had a neutral expression and all pictures had an identical background. Each child saw 32 faces at encoding while the other 32 served as lures at retrieval. During retrieval, each face was paired with a face of the same race and gender, to mimic the perceptual similarity of the target and lure in Experiment 1. The same confidence scale was used from the first experiment.

Procedure

The procedure of Experiment 2 was identical to Experiment 1. The only change was to the stimuli.

Results

Figure 3 shows the results plotted in terms of uncertainty monitoring. To examine the effects of age and confidence on memory accuracy for face stimuli, a two-way ANOVA was performed. Results showed no main effect of age, F(1,1389) = 1.06, $p = 0.30$ *(ns)* and no main effect of confidence level, F(1,1389) = 0.74, $p = 0.39$ *(ns)*. There was also no interaction between age and confidence, F(1,1389) = 2.22 , *p* = 0.14 *(ns)*. 3-year-olds did not show a significant difference in average confidence for their correct and incorrect answers, *t*(493.77) = 0.11, *p* = 0.91 *(ns)*. 4-year-olds did snow a significant difference in average confidence for their correct and incorrect answers, *t*(342.73) = 2.64, *p* < 0.01. 5-year-olds showed a marginal difference in confidence for their correct and incorrect answers, *t*(458.52) = 1.92, *p* = 0.056.

As in Experiment 1, results were also analyzed using CAC analysis, and the results are shown in Figure 4. Confidence level is plotted on the x-axis and accuracy is plotted on the y-axis. Results are divided by age. Unlike in Experiment 1, 3-year-olds, 4-year-olds, and 5-year-olds all showed little to no confidence-accuracy relationship for face stimuli. 3-year-olds' low, medium, and high confidence decisions were not significantly different than chance performance (50% accurate) with *t*(115) = -0.37, *p* = 0.71, *t*(101) = 0.592, *p* = 0.55, and *t*(277) = -0.12, *p* = 0.90, respectively. 4-year-olds' low and medium confidence decisions were not significantly different than chance performance with *t*(63) = -1.26, *p* = 0.21 and *t*(53) = 0.27, *p* = 0.79, respectively. 4-year-olds' high-confidence decisions were significantly higher than chance with *t*(284) = 3.44, *p* < 0.001. Even then, their high confidence decisions were only 60% accurate. 5-year-olds' low and medium confidence decisions were not significantly different than chance performance with *t*(74) = 0.57, *p* = 0.56 and *t*(85) = 0, *p* = 1, respectively. 5-year-olds' high-confidence decisions were significantly higher than chance with *t*(334) = 2.82, *p* < 0.01. Their high confidence decisions were 58% accurate. These results suggest that while the confidence-accuracy relationship develops for object stimuli during the preschool years, memory performance for faces is too poor to detect a confidence-accuracy relationship for faces in this age group.

Discussion

The current study examined whether a relationship between expressed confidence and recognition memory accuracy develops during the preschool years (e.g. between 3 and 5 years old.) Findings indicate that 4- and 5-year-olds show a confidence-accuracy relationship for object stimuli both when the data were analyzed as average confidence for correct vs. incorrect answers and as CAC curves. 3-year-olds show no confidence-accuracy relationship for object stimuli with either analysis. These results match what was shown by Hembacher & Ghetti (2014) for 3-year-olds and 5-year-olds. However, for 4-year-olds our results showed that they expressed significantly higher confidence for their correct answers compared to their incorrect answers for items shown once (all items were only shown once in our experiment). Hembacher & Ghetti (2014) showed that 4-year-olds only expressed higher confidence in their correct answers compared to their incorrect answers for items presented multiple times. Based on our results, 4-year-olds show similar uncertainty monitoring to 5-year-olds for object stimuli.

For face stimuli, 3-, 4-, and 5-year-olds showed virtually no confidence-accuracy relationship. This could be due to the poor memory performance across all age groups. Only 4- and 5-year-olds' high confidence decisions were significantly above chance performance. When adults show chance memory performance, they also show no relationship between their expressed confidence and their memory accuracy (Nguyen, Pezdek & Wixted, 2016; Weber & Brewer, 2003), so it is not unreasonable that children show a similar effect, especially when children of this age already show a weaker confidence-accuracy relationship than adults and weaker memory performance overall. As in Experiment 1, a similar relationship between confidence and accuracy was shown when average confidence for correct vs. incorrect answers was compared to CAC analysis.

There was an unexpected disparity in performance between object recognition and face recognition. In adults, the relationship between confidence and accuracy remains strong regardless of the type of stimuli used. Findings from this study indicate that children show far weaker memory accuracy for face stimuli compared to object stimuli, as well as a far weaker confidence-accuracy relationship. Apparently, something makes faces in this study more difficult for children to remember than the objects. One prior study showed that preschool-aged children show no differences in their recognition accuracy for objects and faces (Picozzi, Cassia, Turati & Vescovo, 2009). Thus, it may something unique to the paradigm of this study that caused the disparity we observed. There are many possible explanations for this. One explanation may be related to the racial diversity of the faces. Adults show significantly lower memory accuracy for faces of another race compared to faces of their own race (e.g. Nguyen et al., 2016). If the majority of the faces in the study did not match the race of the majority of the participants, this may have worsened their overall memory performance. However, if their performance would have been above chance, it is possible that the confidence-accuracy relationship would still be strong even for cross-race faces, as it is for adults (Nguyen et al., 2016). Another possibility is that faces are more difficult to remember than objects because they are encoded without any semantic information or context. When children are asked to remember a ball, then a cat, then a house, they are able to tie each of those stimuli to a word and a unique set of semantic cues. The face stimuli, on the other hand, were presented without any names or other context. There are no simple, semantic cues that will help the child recall them. Their ability to recognize the faces comes down to their ability to identify features that separate the faces they saw from the faces they did not see. A way to test this explanation would be to associate each face with a familiar semantic label (e.g. an occupation) at encoding (e.g. when the face appears on the screen, a hypothetical occupation that person holds is written below it) but then at the recognition test only include the face and not the occupation. Identifying the face by occupation is advantageous over identification by a name because adding a name would just add more novel information for the child to remember. This type of test could also be done with objects (e.g. the name of the object is placed underneath the picture at encoding), and the results compared. If participants then show similar results for both objects and faces, it is evidence that semantic information helps the children encode the stimuli. Another way to test this would be to have all of the objects be from the same semantic category (i.e. all chairs or all dogs) to remove any useful semantic cues. This is similar to the paradigm used by Picozzi et al. (2009). They compared recognition of 26 faces to recognition of 26 shoes or 26 cars.

The results of Experiment 2 hold implications for the ability of preschoolers to serve as eyewitnesses in police investigations. As mentioned previously, adults are considered "reliable" eyewitnesses because their initial identification of a suspect from a lineup is indicative of the accuracy of that decision (e.g. the likelihood that the suspect is guilty) (Wixted et al., 2015; Wixted & Wells, 2017). In Experiment two, preschool-aged children showed both low memory accuracy overall and no relationship between their expressed confidence and their memory accuracy for face stimuli. Thus, they would likely be considered "unreliable" eyewitnesses because their expressed confidence in their identification of a face is not indicative of whether or not they have encountered that face previously. To further explore the reliability of young children as eyewitnesses, preschool-aged children should be tested using an eyewitness-memory-style paradigm, where they are shown a video of a person then asked to choose a picture of the person from the video out of an array of faces. If preschooler's poor performance in Experiment 2 is a result of too much cognitive load from the amount of faces they were asked to remember, it is possible they would show improved performance in this paradigm because there is only one face for them to remember. However, if their poor performance in Experiment 2 is due to a lack of semantic information, their performance may not improve in an eyewitness paradigm.

Many open questions remain about children's recognition memory performance and the relationship between their expressed confidence and recognition memory accuracy. Preschoolers show good recognition memory performance (though, not as good as adults) for objects, but poor recognition memory performance for faces. This suggests that the confidence-accuracy relationship is relatively stimulus-dependent. In addition to the explanations for the disparity provided above, this calls to question when children develop comparable recognition memory accuracy for faces. Children will need to show performance above chance for face stimuli before conclusions can be drawn about the relationship between their confidence and their recognition memory accuracy for faces. Our results differ from prior research that has shown that typically developing older children, between 9 and 18 years old, do show a strong confidence-accuracy relationship for face stimuli presented in an old/new recognition paradigm (Wilkinson et al., 2010). This previous study used 24 face stimuli at encoding, and the only demographic information noted information about the faces is that they were all female. Thus, our results may differ due to the increased number of faces encoded in our study or the varying races of the faces. It is also possible that the confidence-accuracy relationship for faces develops between 5 and 9 years old rather than 3 and 5 years old. Our results do provide additional evidence for Hembacher & Ghetti (2014)'s conclusions that preschool-aged children are able to recognize objects using a two-alternative forced-choice paradigm and 4- and 5-year-olds can use their expressed confidence to indicate the likely accuracy of their memories for objects. However, only one stimulus set of objects has been tested.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Killeen, Isabella M.; Walker, Caren M.; Wixted, John T. The dissertation author was the primary investigator and author of this material.

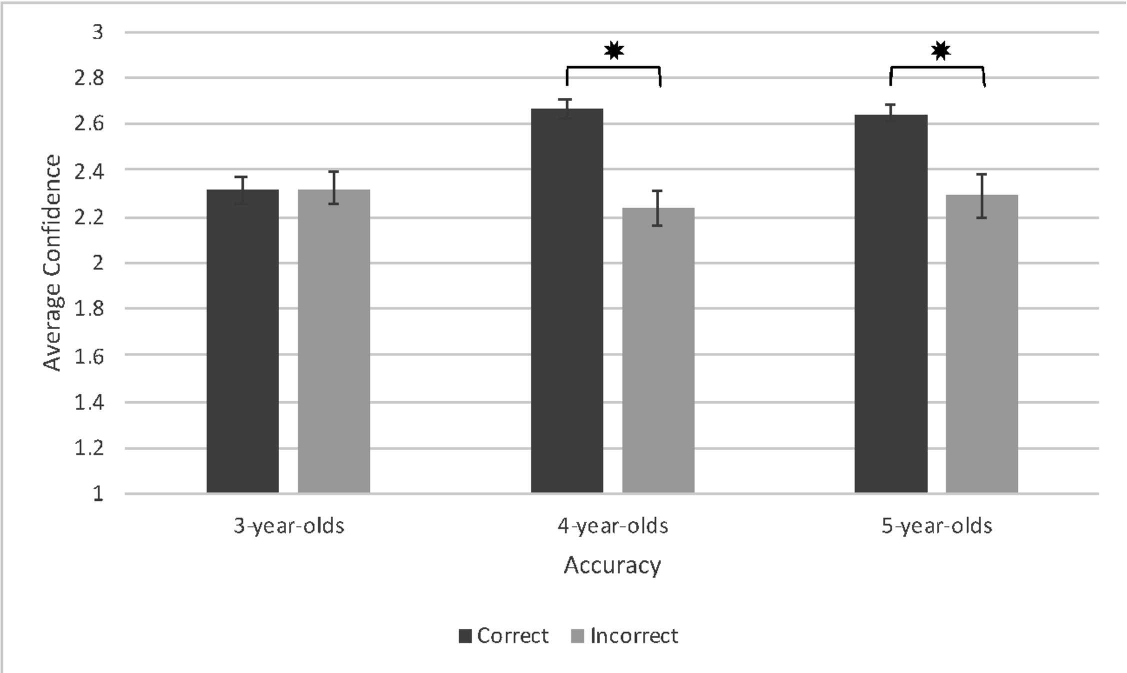

Figure 3.1. Average confidence for correct and incorrect answers for line drawings of objects (Experiment 1), separated by age. * = *p(T) < 0.001.* Error bars represent standard error.

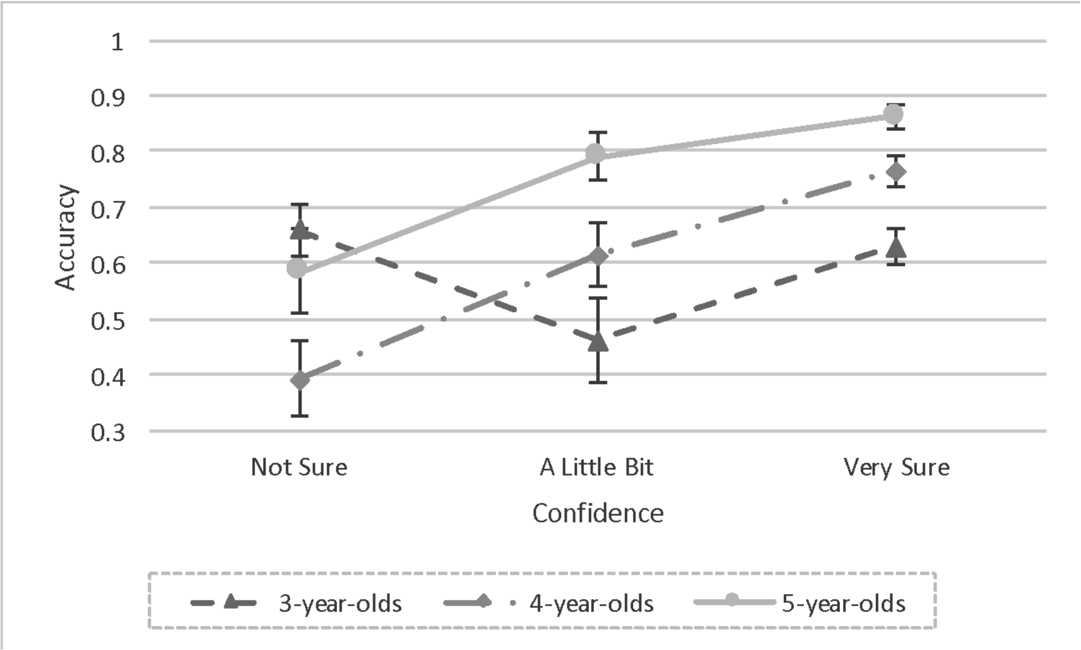

Figure 3.2. Confidence-Accuracy Characteristic curves for line drawings of objects (Experiment 1), separated by age. Error bars represent standard error.

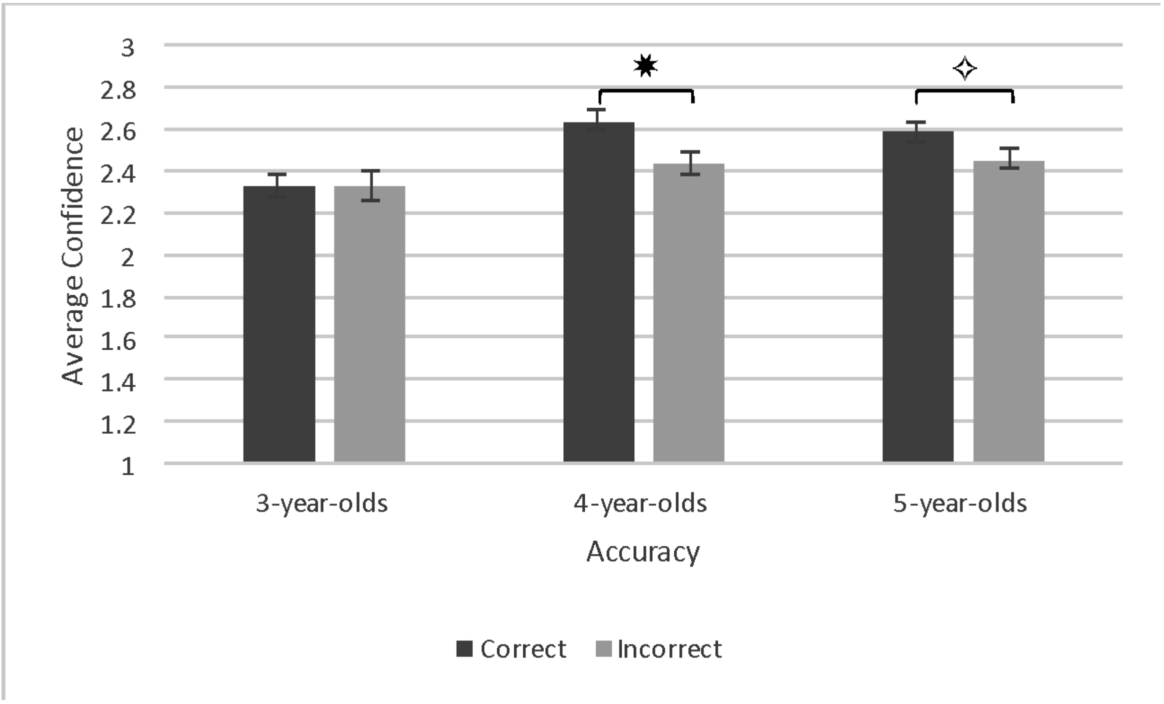

Figure 3.3. Average confidence for correct and incorrect answers for faces (Experiment 2), separated by age. * = statistically significant with *p(T) < 0.001.* ◊ = marginally significant with *p(T) < 0.06.* Error bars represent standard error.

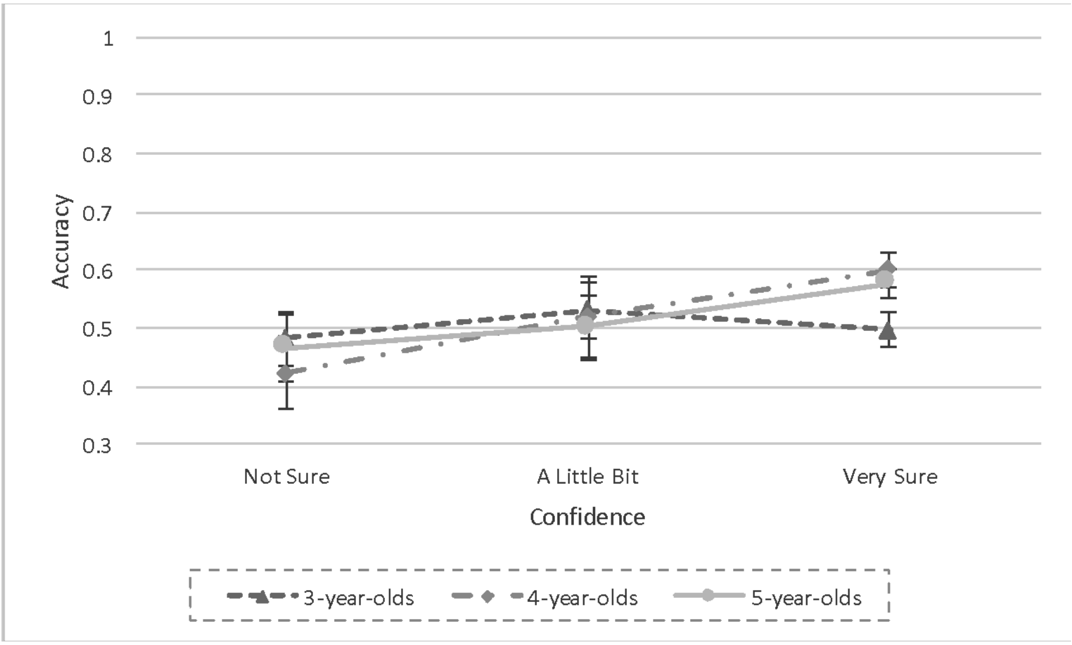

Figure 3.4. Confidence-Accuracy Characteristic curves for faces (Experiment 2), separated by age. Error bars represent standard error.

References

Berch, D.B. & Evans, R.C. (1973). Decision processes in children's recognition memory. *Journal of Experimental Psychology, 16*, 148-164.

Flavell, J.H., Green, F.L., & Flavell, E.R. (1995). Young children's knowledge about thinking. *Monographs of the Society for Research in Child Development, 60*, 1–96.

Hembacher, E. & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*(9), 1768-1776.

Hiller, R.M. & Weber, N. (2013). A comparison of adults' and children's metacognition for yes/no recognition decisions. *Journal of Applied Research in Memory and Cognition, 2*, 185-191.

Koriat, A., Goldsmith, M., Schneider, W., & Nakash-Dura, M. (2001). The credibility of children's testimony: Can children control the accuracy of their memory reports? *Journal of Experimental Child Psychology, 79*, 405–437.

Lyons, K.E. & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development, 82*(6). 1778-1787.

Marazita, J.M. & Merriman, W.E. (2004). Young children's judgments of whether they know the names for objects: The metalinguistic ability it reflects and the process it involves. *Journal of Memory and Language, 51*(3), 458- 472.

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.

Nguyen, T.B., Pezdek, K., & Wixted, J.T. (2016). Evidence for a confidence-accuracy relationship in memory for same- and cross-race faces. *Quarterly Journal of Experimental Psychology, 70*(12), 2518-2534.

Picozzi, M., Cassia, V.M., Turati, C. & Vescovo, E. (2009). The effect of inversion on 3- to 5-year-olds' recognition of face and nonface visual objects. *Journal of Experimental Child Psychology, 102*(4), 487-502.

Plude, D.J., Nelson, T.O., & Scholnick, E.K. (1998). Analytical research on developmental aspects of metamemory. *European Journal of Psychology of Education, 13*, 29–42.

Tekin, E. & Roediger, H.L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*(1), 49.

Weber, N. & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology, 88*(3), 490-499.

Wilkinson, D.A., Best, C.A., Minshew, N.J., & Strauss, M.S. (2010) Memory awareness for faces in individuals with autism. *Journal of Autism and Developmental Disorders, 40*, 1371-1377.

Chapter 4

A Signal-Detection-Based Comparison of Simultaneous and Elimination Lineups Using Receiver Operating Characteristic Analysis

Abstract

While the simultaneous and sequential lineup methods are those most commonly used in adults, another lineup procedure, called the “elimination lineup” has been proposed as an improved procedure for children. In this procedure, the lineup identification is divided into two separate judgments. First, the eyewitness must select the most familiar face of those in the lineup. Then, after all of the other faces have been removed from view, they identify whether or not they believe that face belongs to the guilty suspect. Later work has claimed that the elimination lineup may also elicit higher diagnostic accuracy in adults. However, it has never been tested using receiver operating characteristic (ROC) analysis. Here, we compared the simultaneous and elimination lineup procedures using ROC analysis. We tested two variations on the elimination lineup: the “traditional” elimination lineup as it has been proposed previously, and a novel variation where all of the faces in the lineup remain in view for both of the judgments made. We analyzed the data in theory-free fashion by computing an area-under-the-curve measure. According to our findings, use of the “traditional” elimination lineup results in both lower discriminability and more liberal responding compared to the simultaneous lineup, while use of our novel variation results in similar discriminability compared to the simultaneous lineup, but still with more liberal responding.

Introduction

In the recognition memory field, a well-established theoretical approach for interpreting data is the signal-detection framework (Egan, 1958; Wixted, 2007). However, that framework has not been widely used for interpreting data from real-world recognition memory tasks. A well-known real-world recognition memory task is eyewitness identification from police lineups. In the past, live lineups were often used by the police, but nowadays photo lineups are almost always used. A typical photo lineup contains a photo of the suspect (the person who the police believe might have committed the crime) along 5 or more photos of physically similar individuals who are known to be innocent. Over the past 30 years, a great deal of research has been conducted to determine how to present the lineup photos to eyewitnesses in order to maximize accuracy and, more specifically, to minimize false identifications (IDs) of the innocent. However, until recently, eyewitness identification researchers generally ignored the use of the signal-detection framework, despite its widespread use in basic recognition memory research.

The value of signal detection theory is perhaps best illustrated by briefly reviewing how eyewitness ID researchers have gone about testing whether one kind of lineup format is better than another. For example, should the photos in a lineup be presented to the witness all at once (the traditional simultaneous photo lineup) or should they instead be presented one at a time (the sequential lineup)? Since 1985, many laboratory studies have been conducted to answer this question. When researching lineup performance in the lab, the typical procedure involves participants (e.g. undergraduates or users of Mechanical Turk) watching a video of a mock crime, and then, after a delay, being shown a photo line-up of six physically similar individuals. In the photo lineup, a picture of the actual suspect (e.g. the person from the original video) can either be shown along with five “filler” faces (a target-present lineup) or not shown and replaced with a sixth filler (a target-absent lineup). All of the fillers should have approximately the same level of similarity to the suspect in order to make it a “fair” lineup. Each participant watches one video and makes a single response to a lineup (sometimes providing a confidence rating). Participants can identify someone from the lineup (either the suspect or a filler), or they can reject the lineup. Performance is measured across participants in terms of the correct ID rate and the false ID rate. The correct ID (hit) rate is the proportion of participants who viewed a target-present lineup and correctly selected the actual perpetrator (e.g. the target). The false ID (false alarm) rate is measured in different ways across studies. In some studies, one of the fillers in the target-absent lineup is designated as the innocent suspect, in which case the false ID rate is the proportion of participants who viewed a target-absent line-up and incorrectly chose the designated innocent suspect. Alternatively, if the target-absent lineup is fair and no one is designated as the innocent suspect, the false ID rate can be estimated by counting all filler IDs and dividing by lineup size (6). If the lineup is fair, these two approaches to measuring the false ID rate yield the same estimate in the long run (e.g. Clark, Moreland, & Gronlund, 2014; Palmer, Brewer, Weber, & Nagesh, 2013).

Somehow, the correct ID rate and the false ID rate must be converted to a singular dependent measure to assess lineup performance. In most studies, the dependent measure has been the correct ID rate divided by the false ID rate. This measure is called the “diagnosticity ratio.” Lindsay & Wells (1985) originally used the diagnosticity ratio to measure the relative diagnostic accuracy of simultaneous and sequential lineup presentation techniques. They found that the sequential procedure yielded both a lower false ID rate and a lower correct ID rate, which produced a higher diagnosticity ratio than the simultaneous lineup. This “sequential superiority effect" due to a higher diagnosticity ratio has been replicated many times (e.g. Steblay, Dysart, Fulero & Lindsay, 2001; Steblay, Dysart & Wells, 2011), and because of this, 30% of police departments switched to using the sequential lineup procedure, believing that it is diagnostically superior to the simultaneous lineup.

Recently, though, some eyewitness identification researchers have analyzed their data using receiver-operating characteristic (ROC) analysis (Wixted & Mickes, 2012). An ROC is a plot of the full range of hit and false alarm rates associated with a procedure as responding varies from conservative to liberal. By not calculating the full ROC, it is impossible to differentiate between procedures that affect diagnostic accuracy vs. ones that differ in response bias. Two procedures may be diagnostically identical, but one may cause people to become less likely to make an identification overall, meaning that they set a stricter (more conservative) familiarity criterion for making any sort of identification. The procedure that causes people to be more conservative will yield a higher diagnosticity ratio, even if the diagnostic accuracy of the two procedures is identical. In contrast, an ROC curve is constructed by computing hit and false alarm rates across the full response bias range. For instance, after computing the overall hit and false alarm rates (the first point on the ROC), the next point to the left is computed by removing any identifications made with 10% confidence, the rating given when the participant is essentially guessing (i.e., 10% is the lowest expression confidence when a 100-point confidence scale is used). The next point to the left on the ROC is computed by removing IDs with confidence ratings of 10% and 20%, and so on. The left-most point is based on the most conservative confidence criterion and includes only IDs of 90% or above, where the participant is essentially certain that they are correct. The degree to which an ROC curve extends above the diagonal line of chance performance is a standard measure of "discriminability."

Discriminability refers to the ability of eyewitnesses to distinguish between innocent and guilty suspects.

ROC analysis has recently been used to test simultaneous vs. sequential lineups. Several studies have found that simultaneous lineups yield a higher ROC (i.e., data that fall further away from the diagonal line of chance performance) than sequential lineups (Mickes, Flowe & Wixted, 2012; Gronlund et al., 2012; Dobolyi & Dodson, 2013; Carlson & Carlson, 2014; Andersen, Carlson, Carlson, & Gronlund, 2014). Because the simultaneous lineup falls on a higher ROC than the sequential lineup, it is said to yield higher discriminability, meaning that participants are better able to discern when the suspect is or is not present in a lineup. This means that for any hit and false alarm rate achieved by the sequential procedure, the simultaneous procedure can achieve a higher hit rate and lower false alarm rate.

The Elimination Lineup

A lineup where people are instructed to choose the most familiar face before deciding whether the perpetrator is present has been advanced in the eyewitness research literature as an improvement over the standard simultaneous lineup. This lineup is called the “elimination lineup.” It was originally suggested for use in children by Pozzulo & Lindsay in 1999. The rationale for the use of this lineup method was that children may have difficulty combining both an absolute and a relative judgment into one decision. An absolute judgment is one where a participant is asked whether a single stimulus is one they have seen previously, yes or no. Conversely, a relative judgment is one where the participant selects which stimulus is the most familiar to them out of an array of many stimuli (Lindsay & Wells, 1985).

The simultaneous lineup requires combining both an absolute and a relative judgment into one decision because the child is simply asked who the suspect is out of six possible choices.

This means the child must both make a relative judgment about all six faces to select the most familiar, and then an absolute judgment about whether the face they selected was the face they had previously encountered. The elimination lineup aims to ease any difficulties children have with combining these two judgments into one task by separating them into two separate tasks. In the original study, children were first asked to select the most familiar face, and then, once they had done so, they were asked whether or not the face they selected was the culprit. This study concluded that the elimination lineup was superior to the simultaneous lineup for children because it caused them to have fewer false IDs while resulting in little to no decrease in the correct ID rate. This is essentially a verbal characterization of a pattern that is often characterized quantitatively as a higher diagnosticity ratio for superior lineup condition (the elimination lineup in this case). Their results were not analyzed using ROCs.

In 2008, Pozzulo et al. tested the elimination lineup in adults. They hypothesized that adults may also benefit from having a lineup decision separated into two judgments in a similar way to children. In this study, they again found that the elimination lineup and the simultaneous lineup had comparable correct ID rates, but the false ID rate was lower for the elimination lineup (again, another way of saying that the diagnosticity ratio was higher for the elimination lineup). In that study, the elimination lineup and the sequential lineup yielded comparable results. It should be noted that the original study in children also tested the elimination lineup in adults (Pozzulo & Lindsay, 1999) and found no differences between the simultaneous and elimination lineups, but this was assumed to have been caused by “anomalously high correct identification and correct rejection rates obtained with the simultaneous lineup.” (Pozzulo et al., 2008).

Elimination lineups have never been tested using ROC analysis. As noted earlier, when simultaneous and sequential lineups were analyzed using ROCs, the sequential superiority effect disappeared, and actually the simultaneous lineup was found to produce significantly higher discriminability. A similar effect could be observed with the elimination lineup when the results are analyzed using ROC analysis. As was noted previously, the diagnosticity ratio is computed from only one point on an ROC curve, so in order to get the most comprehensive understanding of the discriminability of the elimination lineup, ROC analysis should be used. The study we designed to test the simplest signal detection model also addresses the applied question of whether or not the elimination lineup is diagnostically superior to simultaneous lineups in adults because our design consists of two conditions: the simultaneous lineup vs. the elimination lineup.

Diagnostic Feature Detection

One of the theories for why the simultaneous lineup yields a higher ROC curve than the sequential lineup is diagnostic feature detection theory (Wixted & Mickes, 2014). When an eyewitness views a simultaneous lineup, they are able to compare features across faces and isolate which features are *diagnostic,* meaning that they are unique to the face of the perpetrator. An eyewitness can then discount all *non-diagnostic* features, or those which are shared across all of the faces in the lineup. For example, because lineups are formed based on a description of the suspect, all members of the lineup may have short, brown hair and brown eyes because those are features that the police believe that the suspect possesses. Thus, when looking at a simultaneous lineup the eyewitness can ignore the hair and eye color because they will not differentiate the suspect from any of the filler faces. They can isolate those features that are unique to the face of the suspect and make an identification based on those features. In the sequential lineup, the eyewitness is unaware that some features are shared across all of the faces, because each one is presented in isolation. They cannot be certain that any feature will be shared by all of the faces except, perhaps, later in the lineup (after having viewed multiple faces).

The elimination lineup, as proposed by Pozzulo & Lindsay (1999), may be similar to the sequential lineup in this way. While the eyewitness is first asked to select the most familiar face and can use diagnostic features to make that choice, all of the faces they do not select are removed prior to them making an identification decision. Thus, even though eyewitnesses may be equally likely to pick the suspect as the most familiar face in both the simultaneous and elimination lineup formats, they may not be as likely to accurately identify the suspect in the elimination lineup because they are no longer able to use diagnostic features to make their identification. To test this, we used two different variants of the elimination lineup in this study. The first is the elimination lineup as proposed by Pozzulo & Lindsay (1999), where the faces disappear before an absolute identification is made. The second is a variation where the eyewitness first picks the most familiar face, but the rest of the faces remain while they make an absolute identification. In both cases, once an eyewitness has made their decision about the most familiar face, they cannot change their answer to that question. The only difference is that the eyewitness can compare across faces at the time of their absolute identification in our novel variation. If diagnostic feature detection does provide a benefit to eyewitness identification decisions, our novel elimination lineup procedure should elicit similar performance to a standard simultaneous lineup, meaning the ROC curves should look similar. On the other hand, the traditional elimination lineup where the filler faces are removed should yield a lower ROC.

Confidence-Accuracy Relationship

ROC analysis is used to determine which lineup procedure best enables eyewitnesses to discriminate between innocent and guilty suspects. An entirely separate question concerns how confidence is related to accuracy for a particular lineup procedure. It is a separate question because, for example, high-confidence IDs can be equally accurate for lineup procedures that yield significantly different ROCs.

Until recently, it was believed that there was, at best, only a modest relationship between a person's confidence in their identification of a suspect and the accuracy of that identification. However, that conclusion was based on the use of the point-biserial correlation coefficient to measure the relationship. Juslin et al. (1996) showed that the correlation can be low even when the confidence-accuracy relationship is as strong as it can possibly be (e.g., when 100% confidence implies 100% accuracy, 90% confidence implies 90% accuracy, and so on). More recent calibration studies have shown that that there is in fact a strong relationship between initial confidence and accuracy in lineup performance (e.g., Brewer & Wells, 2006).

Additionally, it has been argued that even calibration curves are not the most informative way to calculate the confidence-accuracy relationship, because they combine both suspect and filler IDs into the "choosers" group, while suspect and filler IDs indicate completely different things about the guilt of the suspect. If an eyewitness makes a high-confidence suspect ID, this is evidence of guilt, because high-confidence suspect IDs rarely occur when the suspect is innocent (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015). On the other hand, if an eyewitness makes a high-confidence filler ID, this is actually slight evidence of innocence because this kind of ID occurs more often when it is a target-absent lineup compared to when it is a target-present lineup (Wells & Olson, 2002). This is why more information value lies in the confidence-accuracy relationship calculated in terms of the suspect ID rate (excluding filler IDs) as a function of confidence. When measured that way, confidence is almost always found to be highly indicative of accuracy, and high-confidence accuracy is usually very high (Wixted et al., 2015). A key question for applied purposes is whether high-confidence accuracy differs for simultaneous vs. elimination lineups.

In two experiments, we compared the simultaneous lineup to the elimination lineup using ROC analysis. From a theoretical standpoint, the experiments tested whether the simplest signal-detection model accurately maps onto how people make decisions when presented with simultaneous lineups. According to that model (e.g., Macmillan & Creelman, 2005), eyewitnesses first locate the most familiar face and then identify that face if its signal strength exceeds a decision criterion. If participants are already completing those two steps in a traditional simultaneous procedure, then making the steps explicit in an elimination procedure should not affect the area under the ROC. From an applied standpoint, the two experiments tested whether using the elimination lineup results in a higher level of discriminability than the simultaneous lineup for adults (in which case it would be the diagnostically more accurate procedure in a practical sense). Additionally, because ROC analysis requires the collection of data on confidence, we were able to measure the confidence-accuracy relationship for both the simultaneous and elimination lineups.

Experiment 1

Method

Participants

1,360 undergraduate students at the University of California San Diego participated in this study for optional class credit.

Participants were randomly assigned to either the experimental condition with the elimination lineup or the control condition with the simultaneous lineup. Within each condition participants were randomly assigned to a target-present or target-absent lineup.

Materials

The study was conducted online over the University of California San Diego servers. It took approximately 10-15 minutes to complete. The study stimulus was a brief video of a mock crime where the culprit, a young, white, blonde-haired female, is seen spray-painting a wall, looking directly into the camera, and then walking away. A six-person line-up, either elimination or simultaneous, was presented in a 2x3 array and either contained a photo of the culprit (target-present) or did not (target-absent). The "filler" faces (5 of them in the target-present condition and 6 in the target-absent condition) were randomly drawn from a database of 114 description-matched faces. The database was created by having a separate group of 20 participants watch the video and then complete a form listing the perpetrator's physical attributes, including gender, eye color, hair color, ethnicity, height, and weight. We then entered the range of values for each of these attributes (and an age range of 20 to 30 years) into the Florida Department of Corrections Offender Network database (http://www.dc.state.fl.us/AppCommon/) to retrieve description-matched photographs. The culprit and filler faces were presented in greyscale and randomly positioned per line-up.

Procedure

Participants were instructed to watch a video and to be prepared to answer questions about the video at a later time. The video was followed by a distractor task where the participants played Tetris for 5 minutes. Next, participants were presented with the lineup. In the simultaneous condition, participants either selected the photo they believed was the culprit or selected a response option indicating that the culprit was not in the lineup. Once they selected a face, they could not change their answer. In the elimination condition, participants were instructed to first click on the most familiar face, which then appeared alone (i.e., the other faces disappeared), and then decide whether or not that face belonged to the culprit. After making a decision, all participants rated their confidence on a scale from 0-100 in increments of 10, with 0 indicating that they were guessing and 100 indicating that they were absolutely certain. After making a line-up decision and supplying a confidence rating, all participants were asked the following multiple choice validation questions: “What crime was committed?” (graffiti) “What was the weather like?” (cloudy) and “What color were the chairs?” (there were none). All of the questions appeared simultaneously on a new screen after the conclusion of the lineup identification. These questions were asked to ensure that they had paid attention to the original video.

Results

The alpha level was .05 for all statistical tests. Table 3.1 shows the number of participants (and the number of responses, because there was one response per participant) for each condition in the analysis.

ROC Analysis

The results were first analyzed by plotting ROC data points for each condition (see Figure 1). Our analysis focused on correct and incorrect suspect IDs (ignoring filler IDs- IDs of a face other than the suspect from a target present lineup- and no IDs) because the question of primary interest concerns the ability of witnesses to discriminate innocent from guilty suspects. Only suspect IDs imperil the identified individual because the fillers are already known to be innocent of the crime in question. Thus, if an eyewitness identifies a filler from a target-present lineup, their identification will not be used to imperil someone in a court of law because the police have already verified the innocence of the person the eyewitness identified. Correct suspect IDs (SIDs) consisted of suspect IDs from target-present lineups (SID_{TP}), and incorrect suspect IDs consisted of suspect IDs from target-absent lineups (SID_{TA}). As noted earlier, some studies pre-designate a filler in target-absent lineups to serve the role of the innocent suspect. However, most studies do not have a designated innocent suspect and estimate the number innocent suspects by counting all filler IDs (FIDs) from target-absent lineups and dividing by lineup size (6). That is, $\sim SID_{TA} = FID_{TA} / 6$. We used this common method of estimating innocent suspect IDs (Palmer et al., 2013).

For each level of confidence, the correct ID rate (also known as the hit rate) was equal to correct suspect IDs made with that level of confidence or less divided by the number of target-present lineups (nTP), and the false ID rate (also known as the false alarm rate) was equal to estimated incorrect suspect IDs made with that level of confidence or less divided by the number of target-absent lineups (nTA).

The ROC data are shown in Figure 1. The elimination lineup is shown in with a dashed line and the simultaneous lineup is shown with a solid line. The right-most data point for each condition represents the overall correct and false ID rate. These are the data points that have traditionally been used to compute the diagnosticity ratio (correct ID rate / false ID rate). The additional points are computed using the correct and false ID rate for a stricter confidence criterion. The problem with relying on the diagnosticity ratio as a dependent measure is that every point on the ROC has its own diagnosticity ratio, and its value increases continuously as a more conservative decision criterion is used (Wixted & Mickes, 2014). A more useful dependent measure is the area under the ROC because the procedure associated with higher value can achieve both a higher correct ID rate and lower false ID rate than the competing procedure. Visibly, the simultaneous condition appears to have a higher ROC than the elimination condition.

Unlike standard ROC analysis involving the full range of hit and false alarm rates from 0 to 1, partial ROC (pROC) analysis is appropriate here because the maximum false ID rate is less than 1. That is, random guessing from a fair target-absent lineup would yield an estimated false ID rate of 1 / 6 = .17. Instead of computing the full Area Under the Curve (AUC) as would be done for a standard ROC, Partial Area Under the Curve (pAUC) values were computed and compared using the statistical package pROC (Robin et al., 2011). For each ROC analysis, we selected a FAR range from 0 to *q*, where *q* was set to a value equal to the overall false alarm rate obtained for the simultaneous ROCs (i.e., 0.07). Thus, this analysis asks whether the area under the ROC curve for simultaneous lineups in the false alarm rate range of 0 to 0.07 differs from the area under the ROC curve for elimination lineups in that same range.

The pAUC for the simultaneous condition was significantly larger than the pAUC for the elimination condition, $pAUC$(0.542), D = 3.310, $p < 0.001$. This result means that for any given false ID rate, participants made more correct IDs when shown a simultaneous line-up compared to when they were shown an elimination line-up. In other words, in a practical (applied) sense, simultaneous lineups in this experiment were diagnostically superior to elimination lineups. On the surface, this result appears to conflict with the simplest signal detection model of simultaneous lineup performance. Because the elimination lineup simply makes explicit the two decision-making steps that are envisioned by that model (namely, first locate the most familiar face, then identify that face if its signal strength exceeds a decision criterion), the ROCs for the two procedures should have been the same. However, in a simultaneous lineup, after the most familiar face is theoretically located, the other faces do not disappear before making a decision about that face (as they do in the standard elimination procedure). Experiment 2 further investigates this issue by leaving the other faces on the screen in the elimination procedure after the most familiar face is located.

Confidence-Accuracy Relationship

An issue that is distinct from ROC analysis concerns the confidence-accuracy relationship. Just because the ROC is lower for one condition compared to the other does not necessarily mean that accuracy for any given level of confidence will differ across the two conditions. The confidence-specific accuracy measure of most interest is *suspect ID accuracy* (Mickes, 2015), which is simply the probability that a suspect ID made with a particular level of confidence is accurate. More specifically, suspect ID accuracy for a given level of confidence, *c*, is equal to $SID_{TP\text{-}c}$ / ($SID_{TP\text{-}c}$ + $\sim SID_{TP\text{-}c}$), where $SID_{TP\text{-}c}$ is the number of correct suspect IDs made from target-present lineups with confidence *c*, and $\sim SID_{TP\text{-}c}$ is the estimated number of incorrect suspect IDs made from target-absent lineups with confidence *c*. A plot of suspect ID accuracy as a function of confidence is known as a confidence-accuracy characteristic (CAC) curve (Mickes, 2015). Whereas ROC analysis provides the most relevant information to policymakers who have control over the kind of lineup procedure that is used during police investigations (i.e., all else equal, policymakers should use the lineup procedure that yields the highest discriminability), CAC analysis provides the most relevant information to judges and juries, who simply want to know how likely it is that a suspect ID made with a particular level of confidence is accurate.

The confidence-accuracy relationship for both conditions is shown in Figure 3.2. Again, the elimination lineup is shown with a dashed line and the simultaneous lineup is shown with a solid line. Plotted is the proportion of suspect IDs that are correct as a function of confidence on a 3-point scale (low confidence = 0%-60%, medium confidence = 70%-80%, and high confidence = 90% - 100%). The standard errors were estimated as described in the Appendix. The standard errors are clearly overlapping, which indicates that there were no significant differences in the confidence-accuracy relationship between the simultaneous and the elimination condition, though a small advantage for the simultaneous procedure is evident. However, there is an obviously strong relationship between expressed confidence and identification accuracy across conditions.

Experiment 2

As noted above, in the elimination condition Experiment 1 the faces disappeared after the most familiar one was selected. This matched the "fast elimination" technique used in Pozzulo & Lindsay (1999). However, based on diagnostic feature detection theory (Wixted & Mickes, 2014), the removal of the faces could reduce discriminability. Once the filler faces disappear the participant may focus not only on the diagnostic features that allowed them to differentiate the face they chose from the others, but also the non-diagnostic features of that face. When the other faces disappear, they may not be able to continue to make the distinction between diagnostic and non-diagnostic features, similar to in a sequential lineup or show up. To test whether this is the case, in this experiment, after the participant selected the most familiar face the other faces remained on the screen.

Method

Participants

2,335 undergraduate students at the University of California- San Diego participated in this study for optional class credit.

Participants were randomly assigned to either the experimental condition with the elimination lineup or the control condition with the simultaneous lineup. Within each condition participants were randomly assigned to a target-present or target-absent lineup.

Materials

A different mock crime video was used than in Experiment 1. This video showed a young, Caucasian male snatching a purse. Filler photos were taken from a database of 50 description matched faces.

Procedure

The procedure of this study primarily matched the procedure in Experiment 1. The only change made was when the lineup was presented. Just as in Experiment 1, the participants were first asked to select the most familiar face. However, rather than the faces disappearing, this time the faces remained on the screen for while the subsequent questions were presented. As in Experiment 1, the participants were then asked whether the face they selected was the perpetrator, and for their confidence on a scale from 0% to 100%. Even though all of the faces stayed on the screen, participants could not change their answer once they had selected the most familiar face.

Results

The alpha level was .05 for all statistical tests. Table 3.2 shows the number of participants (and the number of responses, because there was one response per participant) for each condition in the analysis.

ROC Analysis

The ROC data are shown in Figure 3.3. The novel elimination lineup is shown with a dashed line and the simultaneous lineup is shown with a solid line. The right-most data point for each condition represents the overall correct and false ID rate. Visibly, the simultaneous condition and the elimination lineup have very similar ROC curves.

Partial Area Under the Curve (pAUC) values were again computed and compared using the statistical package pROC (Robin et al., 2011). The pAUC for the simultaneous condition was not significantly different than the pAUC for the elimination condition, *pAUC*(0.51), D = -0.804, *p* = 0.42 (*ns)*. This result means that for any given false ID rate, participants made a similar number of correct IDs when shown a simultaneous lineup compared to when they were shown an elimination lineup. In other words, simultaneous lineups in this experiment were diagnostically similar to elimination lineups, unlike in Experiment 1 where the simultaneous lineup was diagnostically superior.

Note that we originally ran this experiment using approximately the same number of subjects in Experiment 1. After finding no significant difference between the two procedures, we doubled the number subjects. Even then, the difference between the two procedures was small and did not come close to being significant.

Confidence-Accuracy Relationship

The confidence-accuracy relationship for both conditions is shown in Figure 3.4. The novel elimination lineup is shown with a dashed line and the simultaneous lineup is shown with a solid line. Plotted is the proportion of suspect IDs that are correct as a function of confidence on a 3-point scale (low confidence = 0%-60%, medium confidence = 70%-80%, and high confidence = 90% - 100%). The standard errors were estimated as described in the Appendix. The standard errors are clearly overlapping for the highest level of confidence, which indicates that high confidence identifications made with the novel elimination lineup are just as reliable as high confidence identifications made with the simultaneous lineup. There are differences for low and medium confidence, with the data from the simultaneous lineup showing slightly higher accuracy. This might be explained by the more liberal response bias observed in the novel elimination lineup condition, which is evident in the fact that the rightmost ROC point falls farther to the right than the corresponding ROC point from the simultaneous condition. In the more conservative condition (in this case, the simultaneous condition), eyewitnesses set stricter criteria for their identifications across all levels of confidence, so they are more likely to be accurate in all cases. In the more liberal condition (in this case, the novel elimination condition) eyewitnesses set lower criteria for their identifications across all conditions, so there are going to be more false alarms made with low and medium confidence. This will result in lower accuracy for those levels of confidence. However, a difference between the two lineup procedures in the CAC plot in Experiment 1 (Figure 3.2) was much smaller even though the elimination procedure yielded more liberal responding in that case, too (compare the rightmost ROC points in Figure 1). Thus, the reason for the difference at the lower end of the CAC curves in Figure 3.4 is not entirely clear. In any case, in both conditions of Experiment 2, there is a strong relationship between expressed confidence and identification accuracy.

Discussion

The implications of this study are best indicated by the atheoretical pAUC analysis of the raw ROC data shown in Figures 3.1 and 3.3. The raw ROC data depict the ability of eyewitnesses to sort innocent and guilty suspects into their proper categories without appeal to any theoretical consideration. For adults, our results suggest that the traditional elimination lineup (where the faces disappear after the most familiar face is chosen) is an inferior method compared to the simultaneous lineup. In this case, the elimination lineup yields data that fall on a lower ROC. This finding indicates that adults are less able to distinguish between innocent and guilty suspects when this elimination procedure is used compared to when the simultaneous procedure is used. When the elimination lineup is used but the faces remain on the screen after the most familiar face is chosen, participants show similar discriminability compared to the simultaneous lineup. This is consistent with diagnostic feature detection theory (Wixted & Mickes, 2014), which states that one of the reasons simultaneous lineup yields higher discriminability is because one is able to compare the familiarity of multiple faces at once and thus discount features that are non-diagnostic with regards to the identification decision. In the elimination condition where the faces were removed after the most familiar was picked, participants showed reduced discriminability compared to the simultaneous lineup because they could not discount non-diagnostic features from the most familiar face while making their identification decision. However, in the elimination condition where the faces stayed on the screen the whole time, participants showed similar discriminability to the simultaneous lineup because in both cases they were able to compare features across faces and discount those that were non-diagnostic while making the identification decision.

The conclusions from this research are contrary to Pozzulo et al.'s (2008) suggestion that the elimination lineup is diagnostically superior to the simultaneous lineup. Their conclusion was based on the same reasoning that has been used for many years when comparing simultaneous and sequential lineups. In their study, contrary to what was observed here, both elimination lineups were associated with more conservative responding than simultaneous lineups (the opposite was observed here in both the case where the faces disappeared and the case where the faces remained). The effect on response bias in that study resulted in the pattern of data that has long been used to mistakenly argue that sequential lineups are diagnostically superior to simultaneous lineups. That pattern consists of non-significantly lower correct ID rates and significantly lower false ID rates when the elimination procedure is used. This is just another way of saying that responding was more conservative and the diagnosticity ratio was higher for the elimination lineup. However, a conservative response bias, which invariably results in a higher diagnosticity ratio, is not inherently indicative of diagnostic superiority.

Figures 3.1 and 3.3 actually show that, in the present study, more conservative responding was observed for the *simultaneous* procedure compared to both the traditional elimination procedure and our novel elimination procedure. It is not clear why a different result was found here in terms of response bias, but the mere fact that simultaneous lineups led to more conservative responding (and, therefore, a higher diagnosticity ratio) is not what led us to conclude that simultaneous lineups are diagnostically superior. A more appropriate way to measure diagnostic accuracy is to use ROC analysis because relying on the diagnosticity ratio alone cannot distinguish between an effect on discriminability vs. an effect on response bias. However, based on our results, participants in the traditional elimination lineup condition showed lower discriminability, and that is the key finding. Note that an effect on response bias one way or the other is not particularly important because more conservative responding (if it is desired) can be achieved merely by setting a higher confidence criterion for counting a suspect ID. That is, whether one uses the simultaneous procedure or the elimination procedure, very conservative responding can be easily achieved. For example, a police department could set a policy according to which suspect IDs accompanied by low confidence are treated as effective non-IDs. What is essential is to use the lineup procedure that yields the highest ROC. The traditional elimination lineup appears to take us in the opposite direction (just as previous research suggests about sequential lineups), and our novel elimination lineup does not provide any additional diagnostic benefit above and beyond what the standard simultaneous lineup already provides.

Finally, our results lend some support to the standard signal detection model of simultaneous lineup performance. As indicated earlier, according to that model (e.g., Macmillan & Creelman, 2005), when making an ID from a simultaneous lineup, eyewitnesses first locate the most familiar face and then identify that face if its signal strength exceeds a decision criterion. Our novel elimination procedure in Experiment 2 is simply a standard simultaneous procedure with those two theoretical steps (i.e., first, locate the most familiar face, and, second, decide whether or not to identify that face) made explicit. If participants are already completing those two steps in a traditional simultaneous procedure, then making the steps explicit should not affect the area under the ROC. In fact, as shown in Figure 3.3, the ROC curves are very similar, although more liberal responding is evident in the novel elimination procedure. These findings indicate that there is no reason to question the standard theoretical assumption of signal detection models of lineup performance, according to which eyewitnesses first locate the most familiar face and then identify that face if its signal strength exceeds a decision criterion.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Killeen, Isabella M.; Wilson, Brent M.; Wixted, John T. The dissertation author was the primary investigator and author of this material.

Table 4.1. Number of participants tested in each condition for Experiment 1.

	Target-Absent	**Target-Present**	**Total**
Simultaneous	292	299	591
Elimination	266	271	537
Total	558	570	1128

Table 4.2. Number of participants tested in each condition for Experiment 2.

	Target-Absent	**Target-Present**	**Total**
Simultaneous	593	581	1174
Elimination	572	589	1161
Total	1065	1170	2335

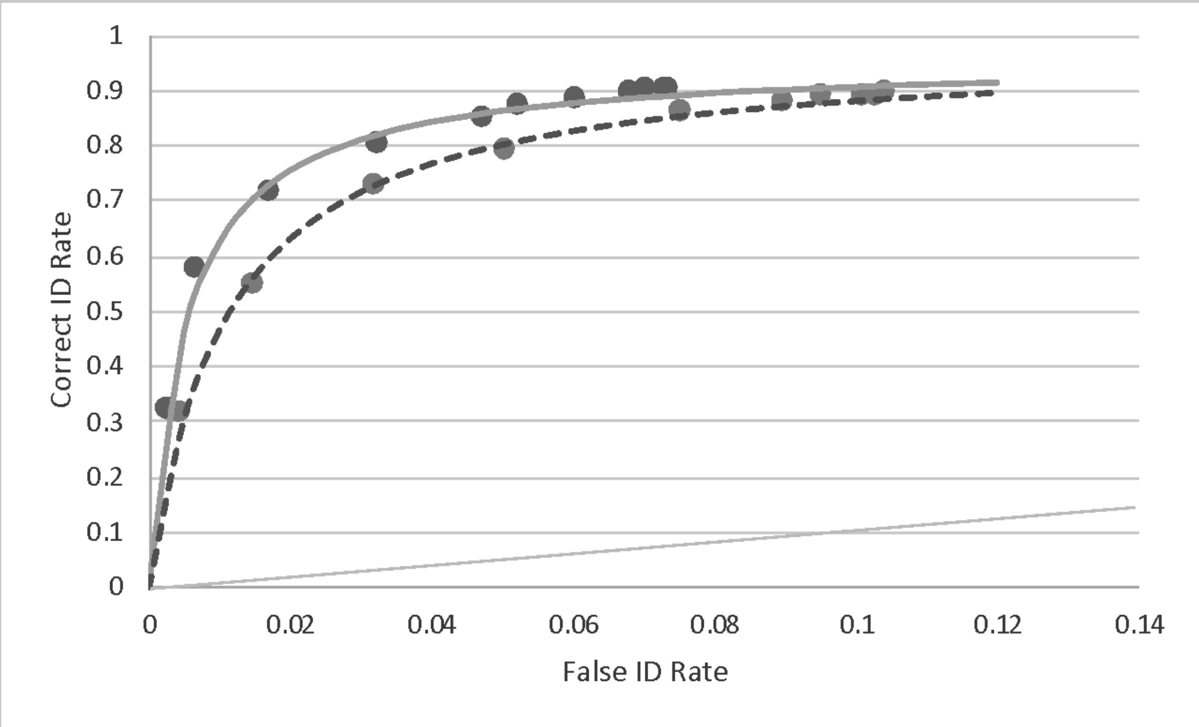

Figure 4.1. ROC data for the simultaneous and traditional elimination lineup procedures with fitted curves. Each filled in circle represents a confidence criterion. The simultaneous lineup condition is shown with a solid line and the traditional elimination lineup condition is shown with a dashed line. The smooth curves drawn through the data represent atheoretical least-squares fits of a hyperbola (which were included to help illustrate the trajectory of the ROC data).

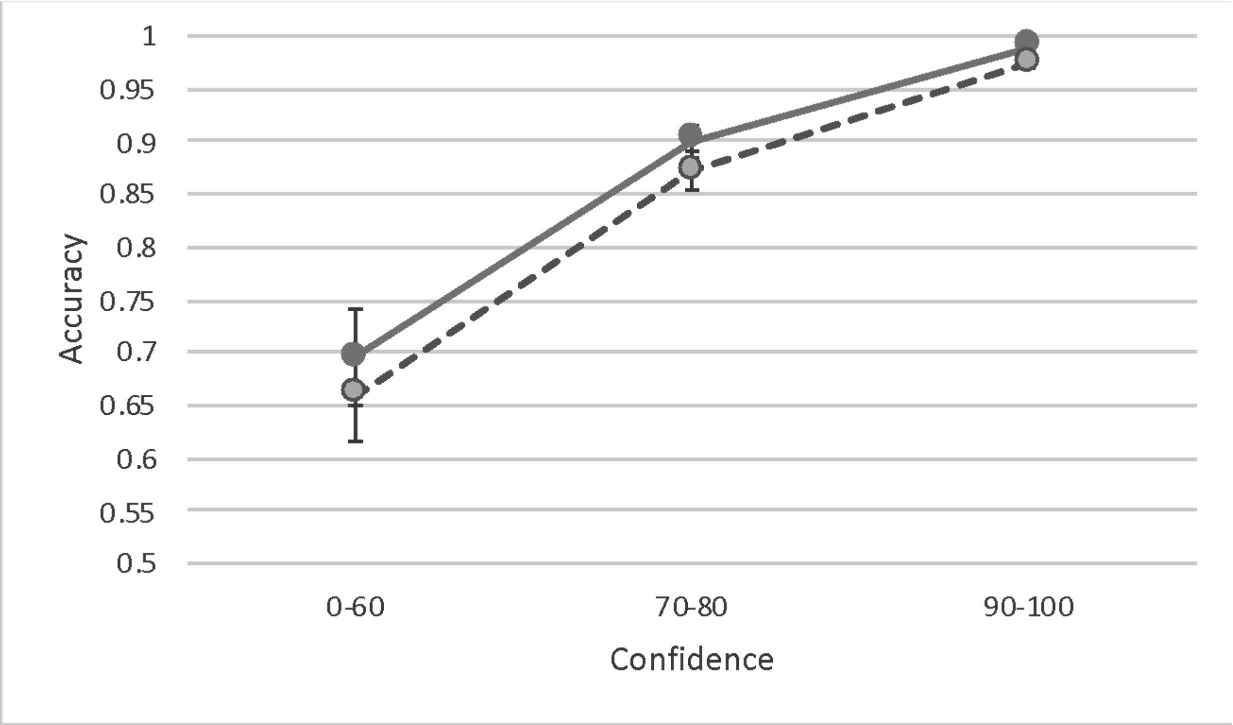

Figure 4.2. Confidence-accuracy data for simultaneous and traditional elimination lineups. The simultaneous condition is shown with a solid line and the traditional elimination condition is shown with the dashed line. Error bars represent the standard error (see Appendix).

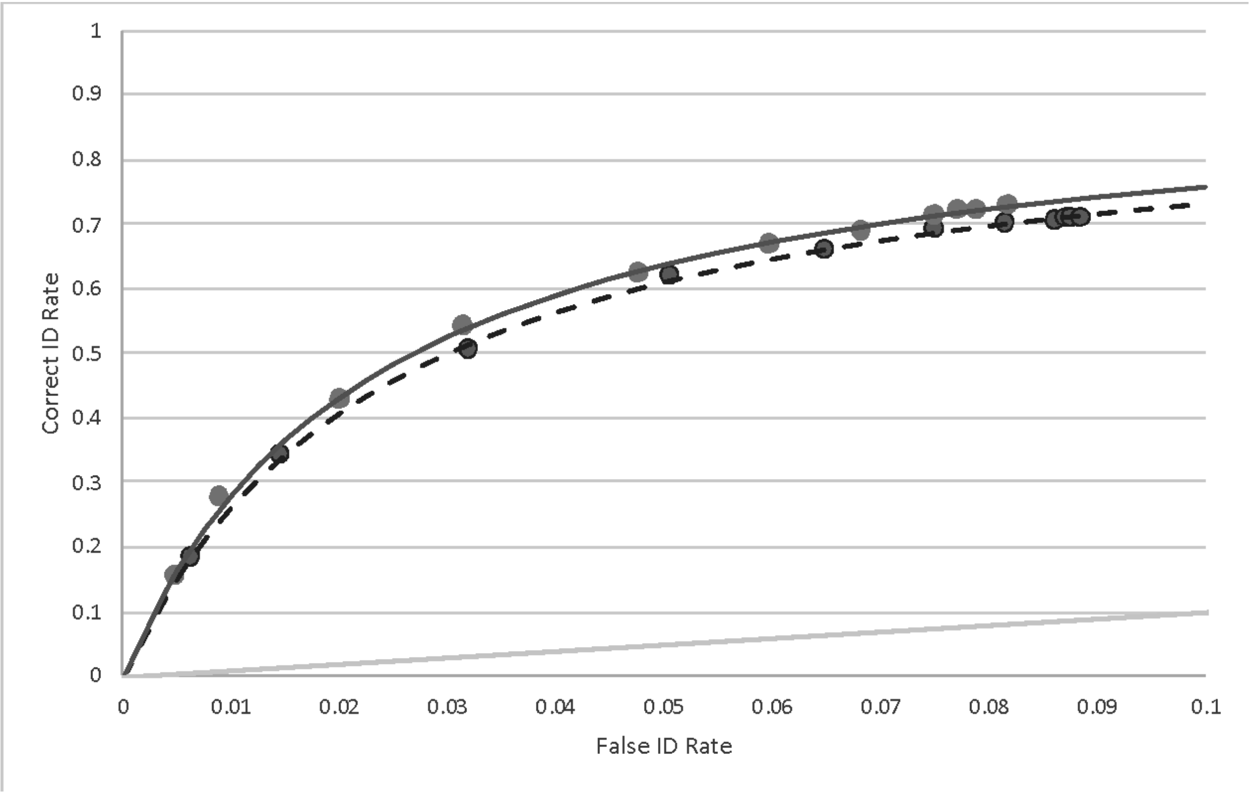

Figure 4.3. ROC data for the simultaneous and novel elimination lineup procedures with fitted curves. Each filled in circle represents a confidence criterion. The simultaneous lineup condition is shown with a solid line and the traditional elimination lineup condition is shown with a dashed line. The smooth curves drawn through the data represent atheoretical least-squares fits of a hyperbola (which were included to help illustrate the trajectory of the ROC data).

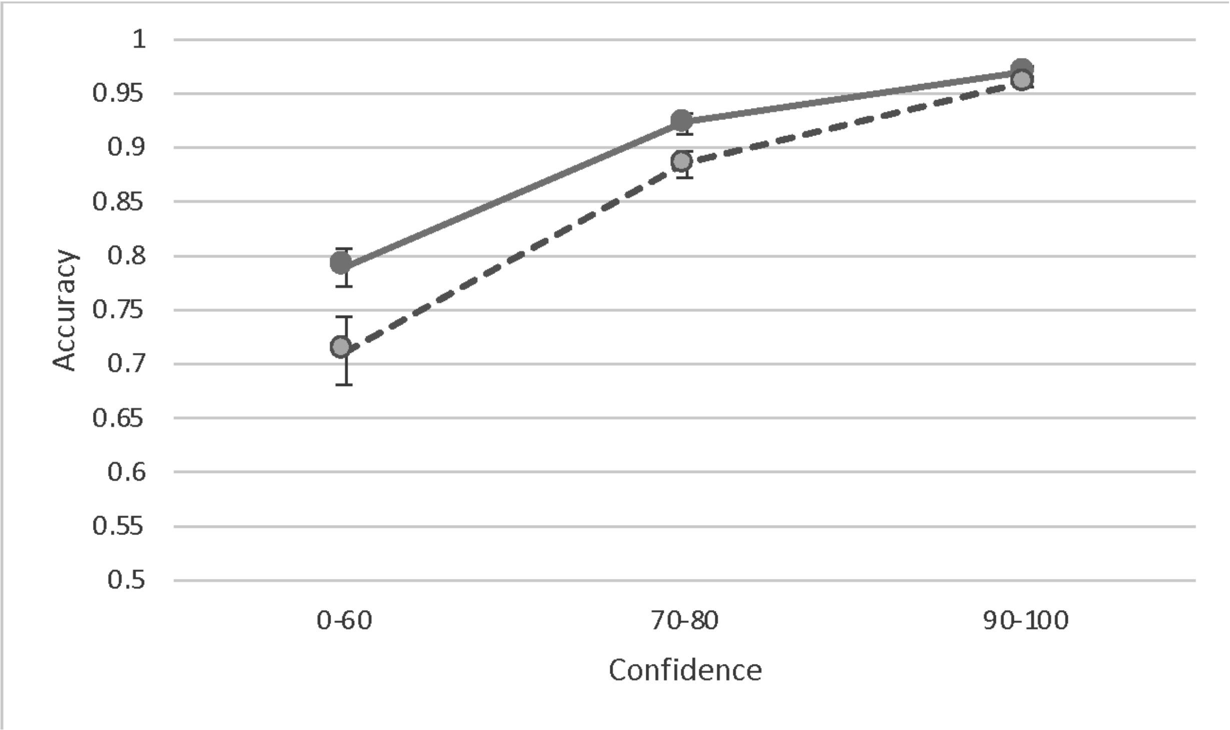

Figure 4.4. Confidence-accuracy data for simultaneous and novel elimination lineups. The simultaneous condition is shown with a solid line and the novel elimination condition is shown with the dashed line. Error bars represent the standard error (see Appendix).

References

Brewer, N. & Wells, G.L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*(1), 11-30.

Clark, S.E., Erickson, M.A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35*(5), 364-380.

Clark, S. E., Moreland, M. B. & Gronlund, S. D. (2014). Evolution of the empirical and theoretical foundations of eyewitness identification reform. *Psychonomic Bulletin & Review, 21*, 251-267.

Carlson, C. A. & Carlson, M. A. (2014). An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *Journal of Applied Research in Memory and Cognition, 3*, 45–53.

Dobolyi, D. G. & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied, 19*, 345–357.

Egan, J. P. (1958). *Recognition memory and the operating characteristic.* (Tech Note AFCRC-TN-58-51). Bloomington, IN: Indiana University, Hearing and Communication Laboratory.

Gronlund, S.D., Carlson, C.A., Neuschatz, J.S., Goodsell, C.A., Wetmore, S.A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition, 1*(4), 221-228.

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

Lindsay, R.C.L. & Wells, G.L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*(3), 556-564.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness. *Journal of Experimental Psychology: Applied, 18*, 361-376.

Mickes, L., Flowe, H.D., & Wixted, J.T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 325-339.

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.

Pozzulo, J.D., Dempsey, J., Corey, S., Girardi, A., Lawandi, A., & Aston, C. (2008). Can a lineup procedure designed for child witnesses work for adults? Comparing simultaneous, sequential, and elimination lineup procedures. *Journal of Applied Social Psychology, 38*(9), 2195-2209.

Pozzulo, J.D. & Lindsay, R.C.L. (1999). Elimination lineups: An improved identification procedure for child eyewitnesses. *Journal of Applied Psychology, 84*(2), 167-176.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., & Muller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(77).

Steblay, N.K., Dysart, J.E., Fulero, S., & Lindsay, R.C.L. (2001). Eyewitness accuracy rates in simultaneous and sequential lineups: A meta-analytic comparison. *Law and Human Behavior, 25*(5), 459-473.

Steblay, N.K., Dysart, J.E., & Wells, G.L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17*(1), 99-139.

Wells, G. L., & Olson, E. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied, 8*, 155-167.

Wixted, J.T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152-176.

Wixted, J.T. & Mickes, L. (2012). The field of eyewitness memory should abandon "probative value" and embrace Receiver Operating Characteristic analysis. *Perspectives on Psychological Science, 7*, 275-278.

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., Mickes, L., Clark, S.E., Gronlund, S.D., & Roediger H.L. (2015). Initial eyewitness confidence reliably predicts identification accuracy. *American Psychologist, 70*, 515-526.

Appendix

Estimating Standard Errors for Suspect ID Accuracy Scores

The standard errors for suspect ID accuracy scores were estimated using a 10,000-trial bootstrap procedure. On each trial, the observed data from target-present lineups were randomly sampled with replacement to obtain a bootstrap sample of suspect IDs for that trial. For example, if for the observed TP data there were 150 high-confidence suspect IDs out of 500 lineups, the observed high-confidence suspect ID hit rate = 150 / 500 = .30. Thus, on each bootstrap trial, a high-confidence suspect ID was registered with probability .30 for each of 500 lineups (i.e., a high-confidence suspect ID would be registered approximately every third lineup, on average). The first bootstrap trial might yield 157 suspect IDs, the next bootstrap trial might yield 141 suspect IDs, and so on. Similarly, on each bootstrap trial, the observed data from target-absent lineups were randomly sampled with replacement to obtain a bootstrap sample of filler IDs for that trial. For example, if for the observed TA data there were 100 high-confidence filler IDs out of 500 lineups, the observed high-confidence filler ID hit rate = 100 / 500 = .20. Thus, on each bootstrap trial, a high-confidence filler ID was registered with probability .20 for each of 500 lineups (i.e., approximately every fifth lineup yielded a high-confidence filler ID). The first bootstrap trial might yield 94 filler IDs, the next bootstrap trial might yield 101 filler IDs, and so on. After obtaining a bootstrap sample of suspect IDs and filler IDs on a given bootstrap trial, a suspect ID accuracy score was computed in exactly the same manner it was computed for the observed data using Equation 7. Thus, for example, if there were 157 suspect IDs and 94 filler IDs on the first bootstrap trial, then suspect ID accuracy for the first bootstrap trial = 157 / (157 + 94/8) = .930. Note that the bootstrap sample of 94 filler IDs was divided by lineup size (6) to estimate innocent suspect IDs from target-absent lineups. Similarly, if there were 141 suspect

IDs and 101 filler IDs on the second bootstrap trial, then suspect ID accuracy for the second bootstrap trial = 141 / (141 + 101/8) = .918. This process was repeated for 10,000 bootstrap trials, and the standard deviation of the 10,000 bootstrap suspect ID scores provided the estimated standard error.

CHAPTER 5

Confidence Scale Use in Preschool-Aged Children: Effects of Disconfirming Evidence

Abstract

Although young children often express overconfidence in their judgments, recent research has demonstrated that children as young as 3-4 years of age may be able to reliably use a confidence scale to discriminate between their own correct and incorrect responses. However, these previous studies give children the opportunity to calibrate their confidence across a large number of trials. The current research introduces a novel paradigm to facilitate children's ability to reflect on and report their own levels of confidence, based on a brief training that relies upon the presentation of disconfirming evidence. This paradigm presented 3-, 4- and 5-year-olds with "windows" that range in occlusion (none, partial, and full occlusion). Over three trials, children were prompted to use one of two 3-point scales to assess their level of confidence that a target shape is located behind each window. All children received disconfirming evidence for their initial belief, either immediately, during the first trial, or later, during the second trial. Results suggest that when evidence is revealed that violates children's expectations about the presence of the target shape on the first trial, this disconfirming evidence results in improvements in children's ability to accurately calibrate their confidence on future trials.

Introduction

Children under the age of 8 are typically described as “eternal optimists” who generally express overconfidence in their decisions (Mickes, Hwe, Wais, & Wixted, 2011). In fact, children often express high confidence even in cases in which they are likely to be incorrect, based on the level of uncertainty present (Roebers, 2002). Despite this overconfidence, previous research has proposed that children begin to engage in “uncertainty monitoring” around the preschool years (3-4 years of age). Uncertainty monitoring is defined as an introspective process by which a learner considers whether a decision made under unreliable conditions is likely to be correct (Koriat & Goldsmith, 1996), and is measured as the difference in average confidence for correct versus incorrect judgments. Specifically, if children show significantly higher average confidence for correct judgments as compared to incorrect ones, they are thought to be able to monitor their own uncertainty for a specific task (Lyons & Ghetti, 2011).

That said, prior studies examining uncertainty monitoring in preschool-aged children have all relied upon the presentation of multiple trials (e.g., 10 to 30). As the trials progress, children become more sensitive to the differences between confidence levels, calibrating their scale use over time, through trial and error. It remains unknown whether preschool-aged children can use a confidence scale correctly in the first few trials, *before* they have the opportunity to calibrate their responses. Previous developmental work also suggests a potential dissociation between children’s ability to introspectively monitor their own uncertainty and their ability to overtly express confidence in their decisions. For example, although children show metacognitive monitoring as early as 2 years (Lyons & Ghetti, 2011; Marazita & Merriman, 2004), they cannot differentiate correct and incorrect responses using a confidence scale until around 4 years (Hembacher & Ghetti, 2014).

These difficulties in the early expression of confidence therefore appear to be two-fold: (1) children are generally overconfident in their decisions, and (2) they have difficulty mapping their level of certainty onto fixed scale points. Here we examine whether the presentation of disconfirming evidence may be used to support and scaffold each of these related skills. We also examine, for the first time, whether children are able to learn to use a confidence scale in the absence of repeated trials.

Disconfirming Evidence and Belief Change

One reason to expect that even very young children may be capable of reasoning about uncertainty is related to their sophisticated causal reasoning abilities. Like adults, young children learn by interpreting and integrating new evidence into their existing hypothesis space, updating those hypotheses when necessary (e.g., Gopnik et al., 2014). In fact, when making a decision under conditions of uncertainty, children will often increase their tendency to explore and test new hypotheses in an effort to explain the inconsistencies that they observe (e.g., Legare, 2012; Schulz & Bonawitz, 2007).

In some cases, observing evidence that contradicts a currently-held theory initiates belief revision and conceptual change (e.g., Bonawitz, van Schijndel, Friel, & Schulz, 2012; Rhodes & Wellman, 2013). Past work has also demonstrated that when preschool-aged children are unable to rationally update their beliefs in light of new evidence spontaneously, a variety of scaffolds and training tasks can be used to facilitate the process of belief revision (e.g., Bonawitz, Fischer, & Schulz, 2012; Walker, Lombrozo, Legare, & Gopnik, 2014; Walker, Lombrozo, Williams, Rafferty, & Gopnik, 2016). It is therefore possible that this tendency for learners to update their existing hypotheses in light of contradicting evidence may be applied to facilitate the early recognition of uncertainty, supporting judgments of confidence. In particular, by manipulating the evidence that children observe, it may be possible to improve uncertainty monitoring in children.

Current Study

In the current study, we explore the role of disconfirming evidence in the improvement of confidence scale use and uncertainty monitoring using a novel set of stimuli that are designed to provide varying amounts of evidence in support of (or against) an existing hypothesis. To do so, we first asked that children indicate their level of certainty in an ambiguous context, and then present disconfirming evidence that violates that hypothesis over two trials. We predicted that if the presentation of disconfirming evidence improves confidence scale use, then children should adjust their (initially inflated) confidence ratings to better utilize the different scale values available (i.e., “not sure,” “a little bit sure,” or “very sure”), according to the different levels ambiguity in the evidence that they observe. For example, when confronted with a stimulus providing no evidence, or minimal evidence, children should be more likely to rate their confidence as “not sure” or only “a little bit sure.” On the other hand, when confronted with a stimulus that provides direct evidence, children should be more likely to rate their confidence as “very sure.”

Few studies have measured uncertainty monitoring in young children, and even fewer have compared the different types of confidence scales that are available for use in children. Most studies have relied upon a standard “smiley face” scale, which is also widely used in other domains (Hicks, von Baeyer, Spafford, van Korlaar, and Goodenough, 2001). However, the *point* of a confidence scale is to help children to scaffold the unique and subjective experience of their own levels of confidence onto an external representation. There may therefore be a tradeoff between familiarity (and ease of use) with the likelihood of interference resulting from past exposure to a particular scale in a different domain.

In an effort to address this issue, novel scales have been developed exclusively for the assessment of confidence. However, these scales come with their own set of limitations. For example, given the assumption that children lack the metacognitive sensitivity to distribute their confidence across multiple categories, one scale relies upon a binary judgement (i.e., “sure” vs. “not sure”) (Berch & Evans, 1975). However, despite these gains in simplicity, a binary scale necessarily limits the amount of information that can be produced: The fewer options that a learner has for expressing their confidence, the less likely it is that confidence judgements will successfully track accuracy. Hembacher and Ghetti (2014) worked to bypass these issues by developing a cartoon confidence scale with three points (“very sure,” “a little bit sure,” and “not sure”). This scale provides a greater amount of precision than the binary scale, and findings indicate that preschool-aged children are able to use all three points (Hembacher & Ghetti, 2014), but use the middle point with less frequency and more variability than the two end points. This may be due to the high degree of similarity between the cartoon images used, so we developed a new three-point scale for the current study using photographs that more clearly differentiate among the three levels of confidence. Therefore, in addition to assessing whether the presence of disconfirming evidence might support uncertainty monitoring in preschool-aged children, we also compared their performance using this new scale to the standard scale.

Methods

Participants

A total of 117 children participated in this study, including 38 3-year-olds (*M*=42.29 months, *SD*=4.03, range: 36-47 months), 35 4-year-olds, (*M*=54.14 months, *SD*=3.27, range: 48-

59 months), and 40 5-year-olds (*M*=65.12 months, *SD*=3.83, range: 60-71 months). An additional 15 children were excluded, due to failure to pass an attention check (7), experimenter error (3), caretaker interference (2) or failure to complete the entire task (3). Children were recruited from local preschools and science museums in a primarily urban setting. While specific demographic information was not collected from individual participants, demographics of the recruitment locations suggest the participants were predominately white (44.5%) and middle-class (median household income of $73,900) based on US Census Data.

Materials

Two confidence scales were used in this study. Half of participants were randomly assigned to receive a newly developed 3-point confidence scale (Figure 4.1.a), and half received a standard, 3-point "smiley face" scale (Figure 4.1.b). The newly developed scale featured three photographs of a child using a combination of facial expressions, hand gestures, and body postures associated with each level of confidence being expressed.

Other task materials included three "windows," constructed of blue paper and clear plastic sheets of varying amounts of occlusion (Figure 4.2). One window was open, with a frame only around the perimeter. A second window was partially occluded, with a cross-shaped frame. The third window was fully occluded. Three paper shapes (a heart, a rectangle, and a star) could be placed behind each of the windows by sliding them in and out of a clear sheet protector. Additional "disconfirming" shapes (Figure 4.3) included cutouts that were designed to be completely concealed when it was slid inside the partially occluded window, making them *appear* to be identical to the target shapes. A separate cut-out of each target shape that was of the same size and color as those behind the windows was used for the child to reference. Three sets of windows were created, one set for each target shape. The shapes were placed behind the windows prior to the task, so the child would not know the location of the disconfirming shape. The disconfirming shape could either be placed behind the partially occluded window or the fully occluded window for each trial. The remaining windows always contained the target shape.

Procedure

Training Period Children were tested one-on-one with the experimenter. In a brief training period, the experimenter introduced children to each of the three "windows" in turn, explaining that shapes could be placed behind them. A training shape (a circle) was used to demonstrate how a shape could be slid in and out of each window. The experimenter would first show the child the three circles, then, one by one, slide them behind the blue frames in the sheet protector. Children could therefore observe how the circles appeared behind each window.

Next, the windows were removed, and children were introduced to one of the two confidence scales, depending upon their assigned condition. They were instructed to point to the image that represented how sure they were (i.e., "not sure," "a little bit sure," or "very sure"). To ensure children understood the task, they were asked, "Which one do you point to when you're [very, a little bit, not] sure?" for all three levels of confidence. If children were unable to complete this task, their data was excluded from analysis (*n*=7).

Test Trials Following this training period, the experimenter produced a new set of three windows, each containing one of the target shapes, and placed them in front of the child. The experimenter asked the child to use the confidence scale to indicate how sure they were that the target shape was behind each window, saying "Are you very sure, a little bit sure, or not sure?" while pointing to the corresponding image on the scale. The child was instructed to respond by pointing to one of the images on the scale. After the child had produced confidence judgments for all three windows, the shapes behind the windows were revealed, one by one. Children then had the opportunity to observe the "disconfirming shape" behind either the partially occluded or the fully occluded window.

To examine potential effects of order of presentation on the calibration of confidence judgments, half of the children were presented with the disconfirming shape behind the partially occluded window on the first trial, while the other half were presented with the disconfirming shape behind the fully occluded window on the first trial. The remaining windows always contained the target shape. The placement of the disconfirming shape was then reversed for the subsequent trial. After the first test trial, the windows were replaced with three new windows for both the second and third trials, concealing the second and third sets of shapes. The same procedure was used for these remaining trials. At the conclusion of the third trial, the child was thanked for their help and dismissed.

Results

Confidence Scale Type

No significant differences in average confidence were found between the novel 3- point confidence scale and the standard "smiley face" scale for any of the three windows, with $t(336.4)=0.70$, $p = 0.48$ for the clear window, $t(323.4)=-0.83$, $p = 0.41$ for the partially occluded window, and $t(336.53)=1.35$, $p = 0.18$ for the fully occluded window. We therefore combined these scales for all subsequent analyses. These results appear in Figure 4.4.

Disconfirming Evidence

These results appear in Figures 4.5.a and 4.5.b. These results are collapsed across age as well as scale type. Consistent with our predictions, when the disconfirming shape was revealed from behind the partially occluded window in the first trial children showed similar confidence for the clear window, $t($92.24) = -1.62, p = .11, significantly lower confidence for the partially occluded window, t(98.49) =2.9, p = .005, and significantly lower confidence for the fully occluded window, t(104.6) = 2.43, p = .02 on the third (final) trial. However, when the order was reversed, and the disconfirming shape was revealed from behind the partially occluded window in the *second trial,* children did *not* show significant changes in confidence scale use between the first trial and the third trial (clear: p(T) =.69, partially occluded: p(T) = .26, fully occluded: p(T) = .53).

Confidence Judgments by Age

In addition to the role of disconfirming evidence in children's confidence scale use, we were also interested in the effect of age. 3-year-olds and 4-year-olds showed no differences in average confidence across windows on Trial 1. On Trial 3, 3-year-olds still showed no significant differences in average confidence across windows. These results are shown in Figure 4.6. On Trial 3, 4-year-olds showed no difference in average confidence between the clear window and the partially occluded window, but they did show significantly lower confidence for the fully occluded window compared to the clear window, t(65.66) = 3.22, p = 0.002. There was no significant difference in average confidence between the partially occluded window and the fully occluded window. Thus, 4-year-olds showed a *binary* understanding of confidence on Trial 3, meaning they show an understanding of the difference between "sure" and "not sure," but not the difference between "sure" and "a little bit sure" or "a little bit sure" and "not sure." These results are shown in Figure 4.7. 5-year-olds, on the other hand, showed a binary understanding of confidence on both Trial 1 and Trial 3, with t(66.13) = 4.0, $p < 0.001$ and t(65.18) = 5.0, $p <$ 0.001, respectively. These results are shown in Figure 4.8.

Discussion

The current study examined whether disconfirming evidence might be used to scaffold uncertainty monitoring in preschool-aged children. Findings indicate that 5-year-olds, but not 3- or 4-year-olds are able to spontaneously produce binary confidence judgments (e.g. "very sure" vs. "not sure") in response to differing levels of evidence about the state of the world, even in the absence of this additional scaffolding. This provides the first evidence to our knowledge that 5-year-olds are already able to understand when they are making a decision under unreliable conditions, and can monitor their own uncertainty with minimal training or calibration. Then, following their exposure to disconfirming evidence, performance continued to improve.

Although 4-year-olds did not similarly differentiate between their judgments on the first trial, they also benefitted from the presentation of disconfirming evidence, showing significantly higher confidence for the clear window compared to the fully occluded windows on Trial 3. These findings indicate that children's early sensitivity to evidence likely extends to impact the development of uncertainty monitoring. When children were shown the disconfirming shape behind the partially occluded window, it violated their existing (and reasonable) assumption that the window contained the target shape. After all, the disconfirming shape *looked* like the target shape when viewed through the partially occluded window. When this surprising evidence was revealed immediately, during the first trial, this experience led children to update their hypothesis space in a way that better reflected the uncertainty in the world and immediately apply this new knowledge to future trials. This novel paradigm may therefore provide a quick and effective training tool to facilitate children's ability to accurately report their own confidence in both research and applied settings. Interestingly, however, when children observed evidence which initially *confirmed* their existing beliefs on the first trial (i.e., when the partially occluded window contained the expected target shape), they showed no improvement in their use of the confidence scale. This suggests that the initial availability of evidence that confirms an existing hypothesis about a particular set, may make it more difficult to update that hypothesis in response to disconfirming evidence later on.

Additionally, these results provide initial evidence for a relationship between disconfirming evidence and children's understanding of uncertainty. It is possible that the presentation of disconfirming evidence in a surprising manner facilitated belief revision, particularly in 4-year-olds. Because the shape hidden behind the partially occluded window appeared to match the target shape prior to the reveal, it is likely that the participants hypothesized the shape would match the target shape when revealed. This may explain their initial overconfidence. But when the shape is revealed and it surprisingly does not match the target, participants revised their hypothesis to account for the inconsistency they observed. This is consistent with prior research that shows children in this age group will update their beliefs in response to evidence that is in conflict to what they believed prior (e.g., Bonawitz et al., 2012; Rhodes & Wellman, 2013). In response to this belief revision, they express lower confidence when they are once again asked whether the shapes will match.

Another possibility is that the children treat disconfirming evidence as a type of *negative feedback* (e.g. feedback that their hypothesis was wrong or incorrect). Children in this age show a negativity bias, meaning they are more likely to update their behaviors and decisions in response to negative information from their environment compared to positive information from their environment (Vaish, Grossman & Woodward, 2008). If this explanation is true, it means that children know when a hypothesis that they hold about the state of the world ends up being incorrect, and are able to use that information to update their hypothesis accordingly.

Finally, a variety of open questions remain regarding the features that are most relevant for the creation of an effective confidence scale. The current study found no evidence for improvements resulting from the richer, photographic scale. One explanation for this null result may be that both scales included three images, indicating three levels of confidence (low, medium, and high). Unlike adults, who show the same relationship between confidence and accuracy regardless of how many points are represented (Tekin & Roediger, 2017), preschool-aged children may be unsure how to incorporate the middle point. If so, the 3-point scale may not in fact increase sensitivity over a binary measure for children at this age. Future work will further examine developmental differences associated with the use of these scales.

Taken together, these results provide initial evidence that disconfirming evidence might play a role in the development of children's certainty judgments, and improve their ability to map those judgments onto a confidence scale. These findings therefore contribute to our understanding of the mechanisms underlying the early development of uncertainty monitoring.

Chapter 5, in full, is currently being prepared for submission for publication of the material. Killeen, Isabella M.; Walker, Caren M. The dissertation author was the primary investigator and author of this material.

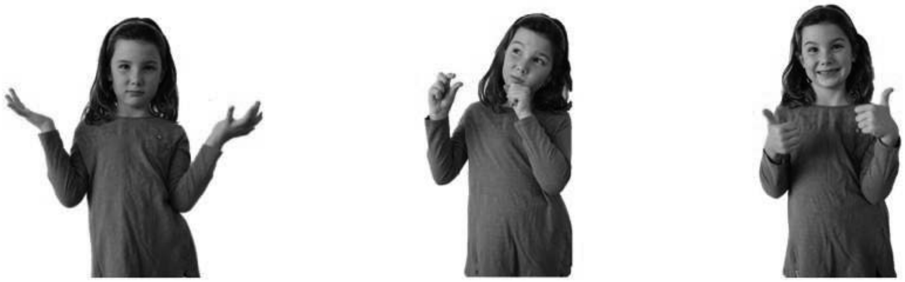

Figure 5.1.a. Novel 3-point confidence scale, from left to right: “not sure,” “a little bit sure,” and “very sure”

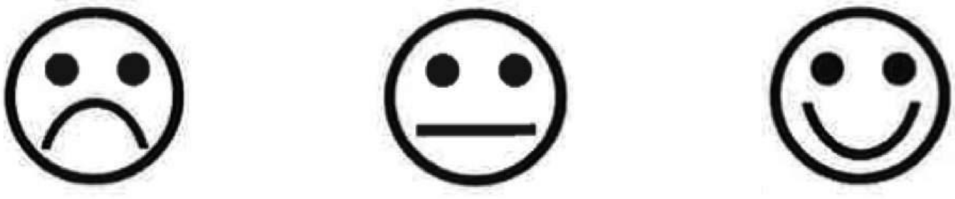

Figure 5.1.b. Standard confidence scale, from left to right: “not sure,” “a little bit sure,” and “very sure”

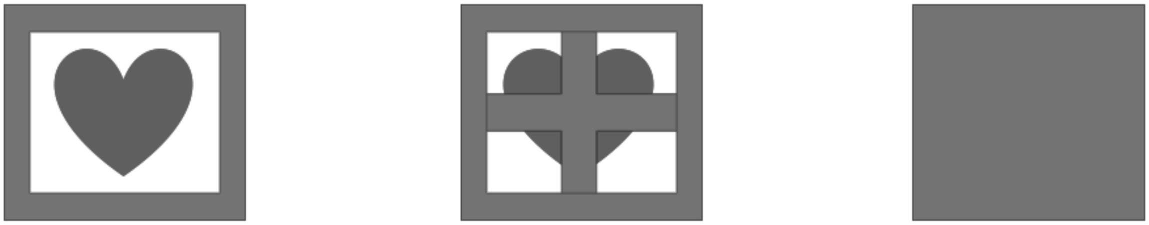

Figure 5.2. Window stimuli with a target shape (heart). From left to right: clear, partially occluded, fully occluded.

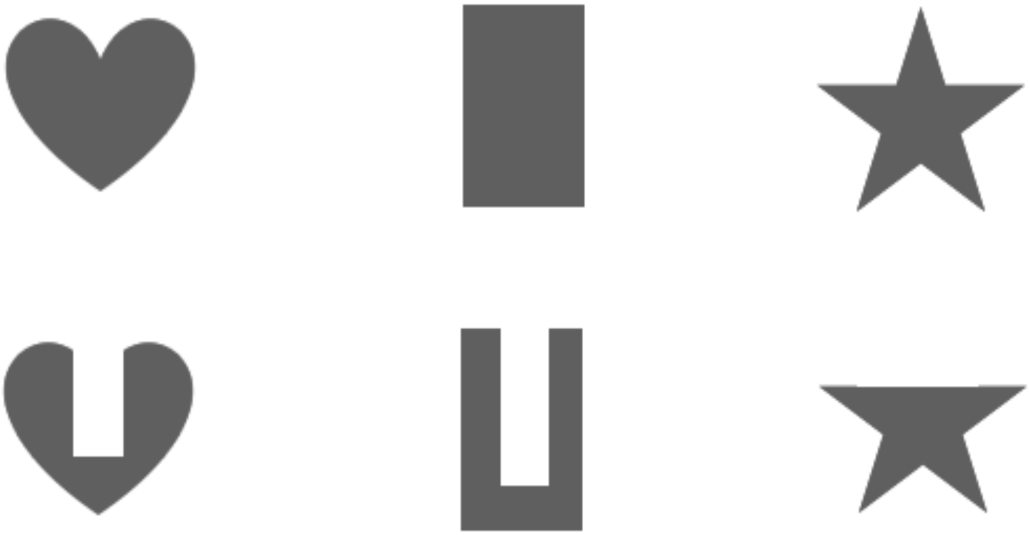

Figure 5.3. Target shapes (top row) paired by column with their corresponding “disconfirming shapes” (bottom row).

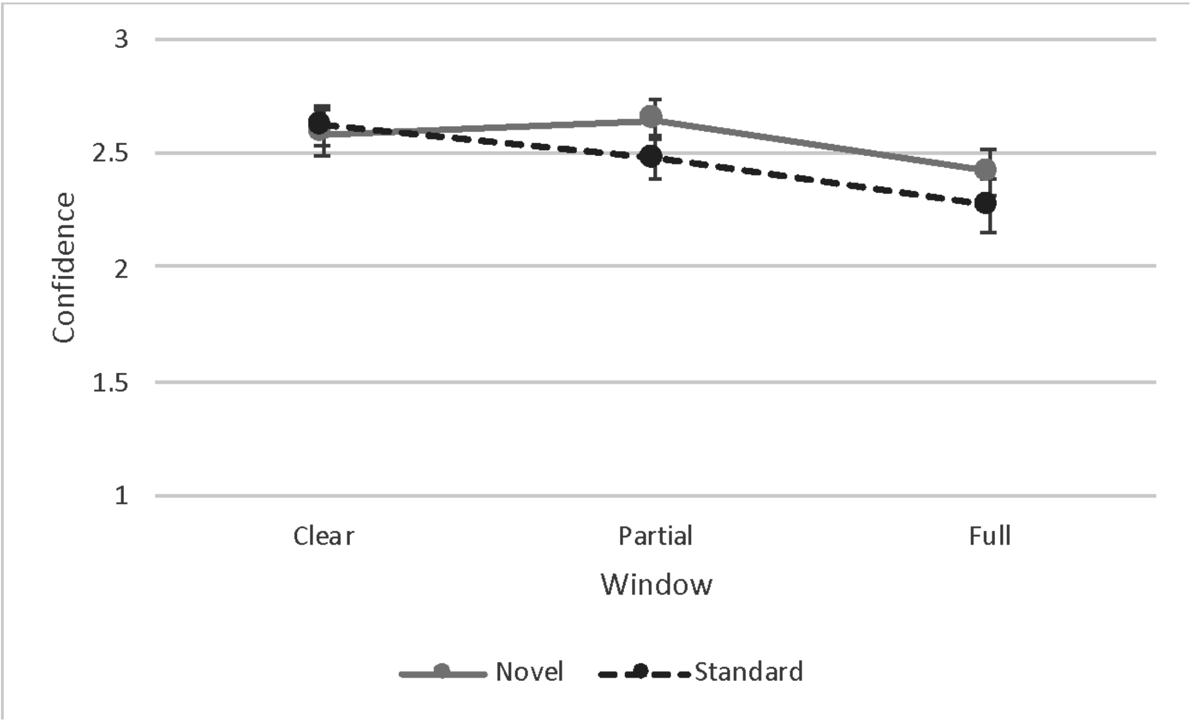

Figure 5.4. Average confidence across clear, partially, and fully occluded windows for the novel and standard confidence scales.

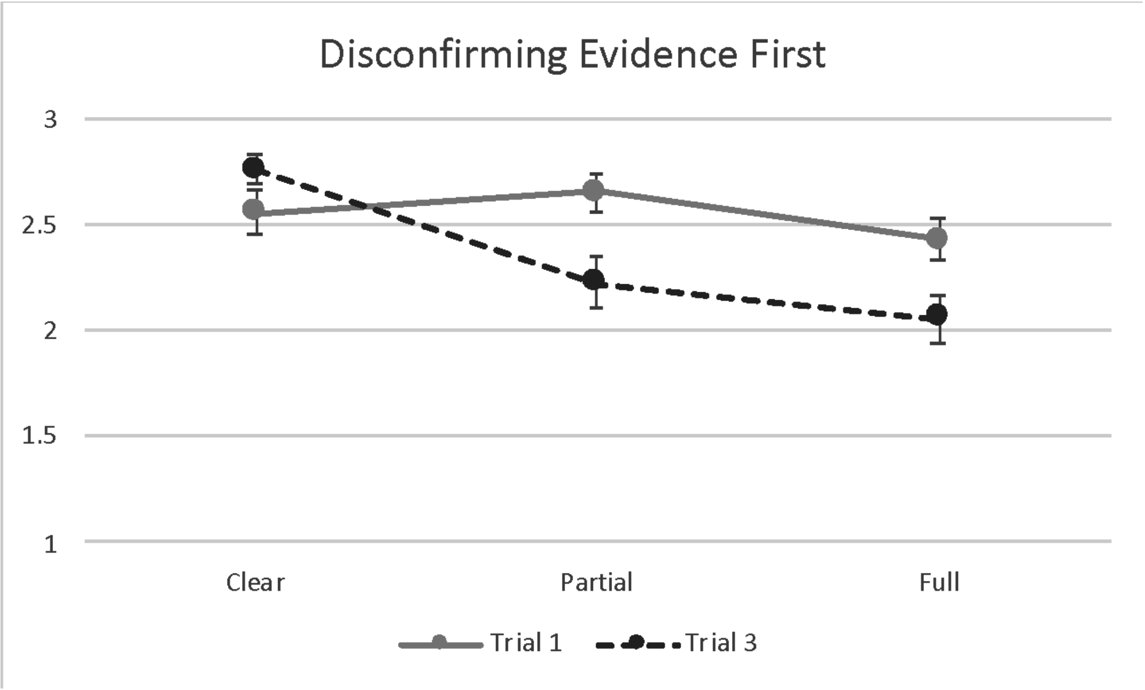

Figure 5.5.a. Average confidence as a function of window occlusion for participants who saw disconfirming evidence after the first trial.

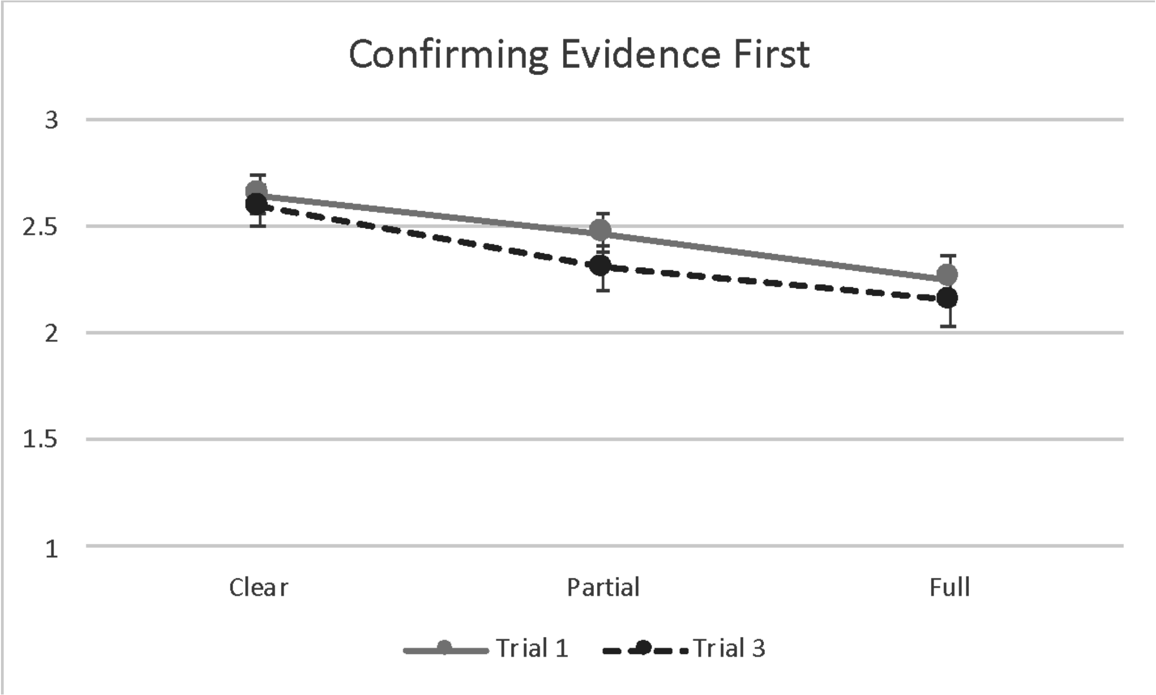

Figure 5.5.b. Average confidence as a function of window occlusion for participants who saw confirming evidence after the first trial.

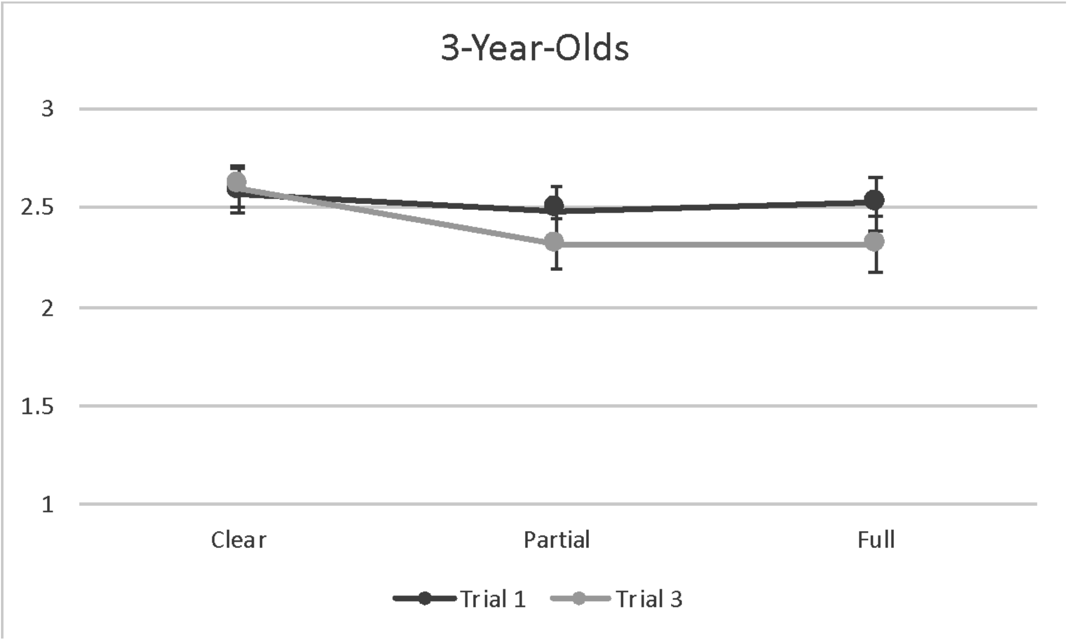

Figure 5.6. Average confidence as a function of window occlusion for 3-year-olds, separated by trial.

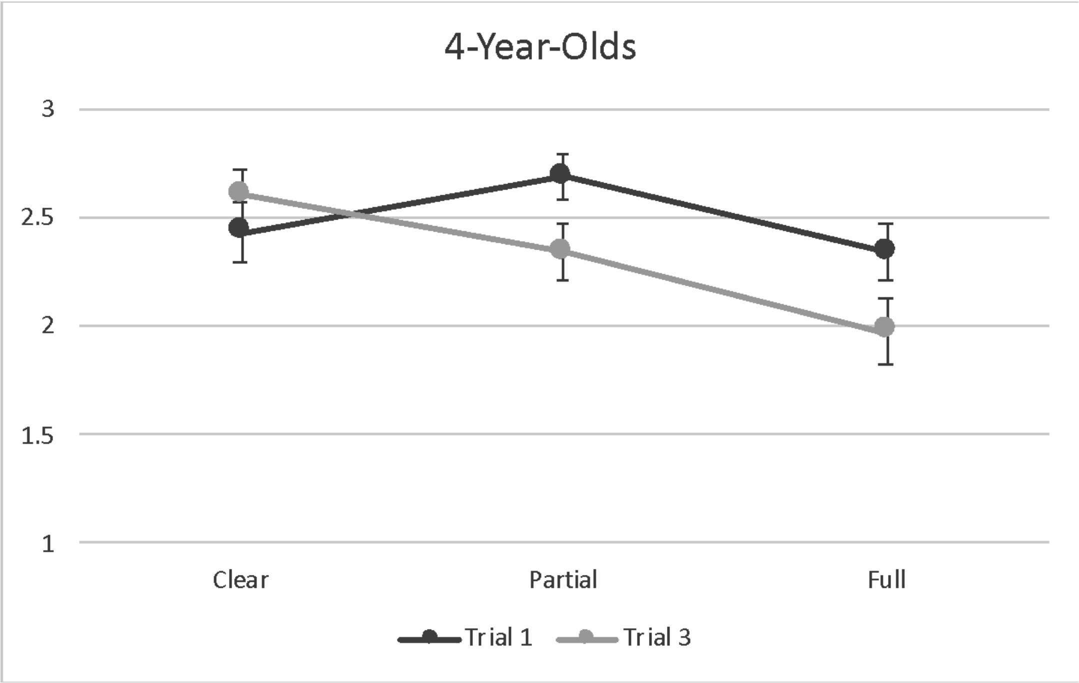

Figure 5.7. Average confidence as a function of window occlusion for 4-year-olds, separated by trial.

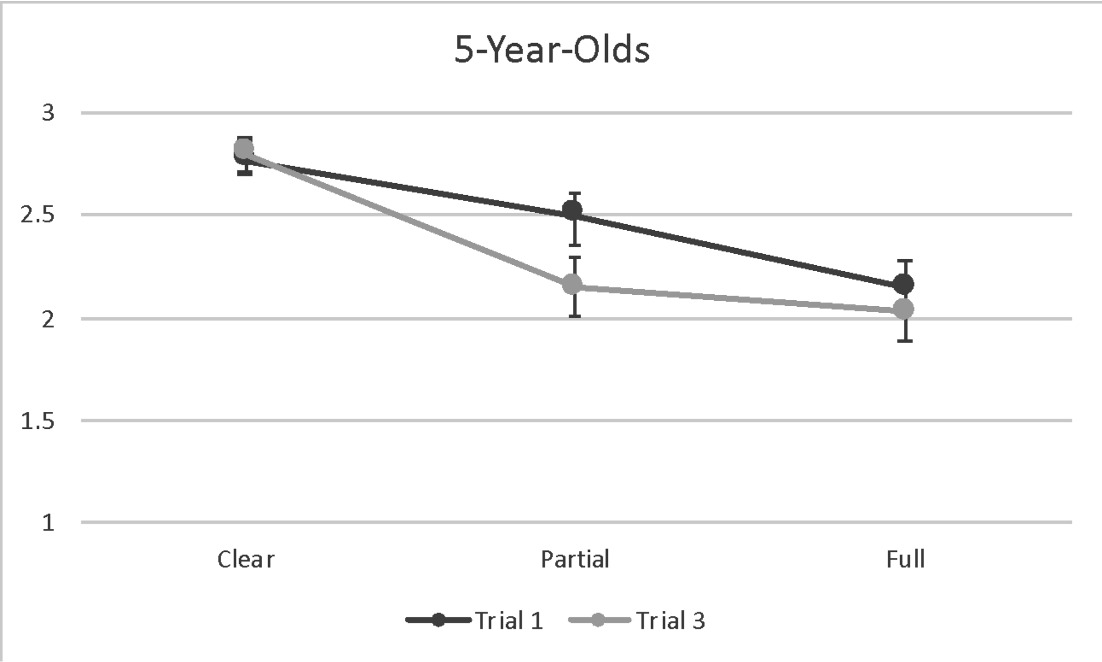

Figure 5.8. Average confidence as a function of window occlusion for 5-year-olds, separated by trial.

References

Berch, D.B. & Evans, R.C. (1973). Decision processes in children's recognition memory. *Journal of Experimental Psychology, 16*, 148-164.

Bonawitz, E.B., Fischer, A., & Schulz, L. (2012) Teaching 3.5-year-olds to revise their beliefs given ambiguous evidence. *Journal of Cognition and Development, 13*(2), 266-280.

Bonawitz, E.B., van Schijndel, T., Friel, D., & Schulz, L. (2012). Balancing theories and evidence in children's exploration, explanations, and learning. *Cognitive Psychology, 64*(4), 215-234.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 1-31.

Hembacher, E. & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*(9), 1768-1776.

Hicks, C.L., von Baeyer, C.L., Spafford, P.A., van Korlaar, I., & Goodenough, B. (2001). The Faces Pain Scale- Revised: Toward a common metric in pediatric pain measurement. *Pain, 93*(2), 173-183.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

Legare, C. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development, 83*, 173-185.

Lyons, K.E. & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development, 82*(6). 1778-1787.

Marazita, J.M. & Merriman, W.E. (2004). Young children's judgments of whether they know the names for objects: The metalinguistic ability it reflects and the process it involves. *Journal of Memory and Language, 51*(3), 458- 472.

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.

Rhodes, M. & Wellman, H. (2013). Constructing a new theory from old ideas and new evidence. *Cognitive Science, 37*(3), 592-604.

Roebers, C.M. (2002). Confidence judgments in children's and adults' event recall and suggestibility. *Developmental Psychology, 38*(6), 1052-1067.

Schulz, L. & Bonawitz, E.B. (2007). Serious fun: Preschoolers play more when evidence is confounded. *Developmental Psychology, 43*(4) 1045-1050.

Tekin, E. & Roediger, H.L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*(1), 49.

Vaish, A., Grossman, T. & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin, 134*(3), 383 403.

Walker, C.M., Lombrozo, T., Legare, C., & Gopnik, A. (2014). Explanation prompts children to privilege inductively rich properties. *Cognition, 133*, 343-357.

Walker, C.M., Lombrozo, T., Williams, J. J., Rafferty, A., & Gopnik, A. (2016). Explaining constrains causal learning in childhood. *Child Development, 88*, 229-246.

CHAPTER 6

Conclusion

In this dissertation, I have examined the relationship between expressed confidence and memory accuracy in both children and adults, as well as confidence scale use in children. Chapter 2 reviewed the existing literature on the confidence-accuracy relationship in children, in both developmental and eyewitness contexts. The work in Chapter 3 provides evidence for the emergence of the confidence-accuracy relationship for recognition memory during the preschool years. It replicated the previous work done in Hembacher & Ghetti (2014). However, the confidence-accuracy relationship was only seen for object stimuli and not face stimuli. The work in Chapter 4 shows that while a lineup procedure designed for children- the elimination lineup- yields lower or equal discriminability (depending on whether or not the filler faces disappear before the eyewitness decides whether the face they have chosen is the culprit) compared to the standard simultaneous lineup in adults, the confidence-accuracy relationship is equally strong regardless. Finally, the work in Chapter 5 describes a novel paradigm that uses disconfirming evidence to improve confidence scale use in preschoolers who initially show overconfidence. In all chapters I have used children's and adult's overt expressions of confidence to deepen our understanding of the development of confidence and its relationship to memory accuracy.

All of this work, taken together, has shown that there are differences between adults and children in both confidence scale use and the confidence-accuracy relationship. Prior work has shown that adults show a strong confidence-accuracy relationship across a wide variety of recognition memory paradigms. This includes two-alternative forced-choice recognition of faces and objects (Boutet & Faubert, 2006), as well as old/new recognition of faces (Wilkinson, Best, Minshew, & Strauss, 2011) and words (Hiller & Weber, 2013), and eyewitness memory paradigms (e.g. Wixted & Wells, 2017). In Chapter 4, we found that adults show a strong confidence-accuracy relationship even when tested with an eyewitness paradigm for children that yielded lower discriminability (e.g. it made it more difficult for adults to differentiate between the guilty suspect and filler faces). However, the results of Chapter 3 suggest that the confidence-accuracy relationship is not as strong in preschool-age children. In fact, 3-year-olds show no confidence-accuracy relationship for object stimuli or face stimuli in a two-alternative forced-choice paradigm. This suggests they are "eternal optimists" (Mickes et al., 2011). Between preschool and adolescence, children must develop an understanding of uncertainty that allows them to express confidence with ever increasing precision. By the time children reach adolescence or adulthood, they have shifted from general overconfidence to indicating confidence that precisely matches the likely accuracy of their decisions.

Disconfirming evidence may play a role in the development of a relationship between confidence and accuracy. As shown in Chapter 5, children's confidence scale use improved after the presentation of disconfirming evidence, but not confirming evidence. In this case, disconfirming evidence may facilitate belief revision due to the presentation of information that conflicts with the children's preexisting hypothesis about the state of the world (Bonawitz et al., 2012; Rhodes & Wellman, 2013). As children revised their hypotheses to better match the state of the world, they adjusted their confidence scale use accordingly. While in this case we intentionally set up a plausible hypothesis about the state of the world to then violate, this situation may resemble how children learn to account for uncertainty in the real world. Children's initial overconfidence may be due to a strong belief in their preexisting hypotheses about the world. As they encounter more and more evidence that disputes their hypotheses, and they update their hypotheses accordingly, their confidence may lower to better reflect their uncertainty about the state of the world. Children may also treat disconfirming evidence as a type of *negative feedback* (e.g. feedback that their hypothesis was wrong or incorrect). As mentioned in Chapter 5, children in this age show a negativity bias, meaning they are more likely to update their behaviors and decisions in response to negative information from their environment compared to positive information from their environment (Vaish, Grossman & Woodward, 2008). If children treat disconfirming evidence as a negative outcome, this explains why they are more likely to update their beliefs in response to it compared to confirming evidence (e.g. the corresponding positive outcome). In the paradigm used in Chapter 5, where children were not given any indication of whether their answers were correct or incorrect, they would need to assign correctness themselves. However, in the real world, children- particularly once they reach preschool and kindergarten age- are often given immediate feedback about the correctness of their decisions. Thus, if they make a decision based on a hypothesis they hold and are then immediately told they are incorrect, they may associate this new evidence that conflicts with their prior hypothesis with a negative outcome. If they also associate their corresponding confidence with a negative outcome, it may result in lower confidence on future decisions.

The work in Chapter 3 provides evidence that even though adults show an equally strong confidence-accuracy relationship across a variety of stimuli, this may not be the case for children. Future research should continue to test preschooler's memory for faces, but reduce the number of trials to reduce the cognitive load and see if overall memory accuracy improves. Because adults show no confidence-accuracy relationship when memory accuracy is at chance (Nguyen, Pezdek & Wixted, 2016; Weber & Brewer, 2003), it is plausible that if overall memory accuracy were higher, a confidence-accuracy relationship may emerge. This would be the case even for cross-race faces. It just may be that in this case, the large number of cross-race identifications the children had to make dropped their performance to chance and as such the confidence-accuracy relationship disappeared. Future research should collect demographic information from the participants and match the race of the faces in the study to the race of the majority of the participants. Or, to examine whether the confidence-accuracy relationship is just as strong for cross-race faces compared to same-race faces, use faces of both the race of the majority of the participants and faces of the next most common race. Additionally, future research should explore the role of semantic information in memory encoding and consolidation for young children. Perhaps the strong cues associated with the names and other semantic features of objects makes them easier for young children to encode, consolidate, and retrieve. Faces, on the other hand, have very little semantic information associated with them and this may make them more difficult for children to encode, consolidate, and retrieve. If faces were associated with semantic information at encoding, it may improve recognition performance. Finally, future research should test preschool-aged children on eyewitness-memory-style paradigms. These paradigms will reduce the cognitive load compared to the paradigm used in Chapter 3 and will provide a clearer indication of whether children of this age can serve as reliable eyewitnesses.

The work in Chapter 4 shows that while a lineup procedure designed for children- the elimination lineup- yields lower discriminability in adults compared to the standard simultaneous lineup, the confidence-accuracy relationship is strong regardless. These results indicate that the legal system should not switch to using the elimination lineup for adults. Future research should compare the simultaneous lineup to both the novel elimination lineup and the traditional elimination lineup in children. It is possible that children may benefit from the two-step nature of the procedure even if it does not benefit adults. Future research should also directly compare the novel elimination lineup to the traditional elimination lineup to more directly measure the difference in discriminability between the two. In Chapter 4, the two experiments used different stimulus sets. Directly comparing the novel and traditional elimination lineups using one stimulus set will more clearly define the difference between the two lineups.

The work in Chapter 5 developed a novel paradigm that uses disconfirming evidence to improve confidence scale use in preschool-aged children. Future work should examine the role of prior knowledge in the effect of disconfirming evidence on confidence scale use. Is it the case that the children who show the highest amount of overconfidence benefit the most from disconfirming evidence, or is it the case that children who show the most prior understanding of confidence and uncertainty use disconfirming evidence to further clarify their preexisting knowledge? To test this, children should be given a “pre-test” to determine their level of overconfidence and then divided into groups based on those levels. Then, the relative effect of disconfirming evidence should be measured for each group. Additionally, there may be a relationship between the amount of belief revision required and children’s confidence in their hypotheses. For example, if a child continuously has their beliefs violated (e.g. children are continuously shown that the hidden shapes do not match their target shapes), their confidence may diminish significantly. On the other hand, if children never have their beliefs violated (e.g. children are continuously shown that the hidden shapes do match the target shapes) they may become even more overconfident. To test this theory, as a follow up to Chapter 5 we will show children only one type of evidence (either confirming or disconfirming) for the entire study and again track the changes in their confidence scale use. If this theory is correct, children would show a more drastic reduction in confidence in response to *only* disconfirming evidence, and no reduction in confidence (or even an increase in confidence) in response to *only* confirming evidence. The work in Chapter 5 showed that the timing of disconfirming evidence mattered greatly, but did not explore what happens when children are only shown a single kind of evidence. Lastly, should future work continue to show that disconfirming evidence plays a role in improving confidence scale use, a training paradigm should be adapted that helps introduce children to a confidence scale that they will use on a later task. This could be important not only for research but also the legal system. If children of this age show a confidence-accuracy relationship in eyewitness paradigms, strengthening their confidence scale use will only further improve the reliability of their eyewitness identifications.

In conclusion, the work of this dissertation sought to increase our understanding of confidence and its relationship to memory accuracy. The review in Chapter 2 synthesized the existing literature on the confidence-accuracy relationship in children. The work in Chapter 3 builds upon the small amount of recognition memory studies conducted in preschool-aged children, and provides additional evidence that children as young as 5 show a confidence-accuracy relationship, at least for object stimuli. It also indicates that the confidence-accuracy relationship for object stimuli develops during the preschool years. The work in Chapter 4 shows that the elimination lineup, which was initially proposed as an improved lineup procedure for children and later claimed to be an improved procedure for adults as well, is not an improved procedure for adults compared to the standard simultaneous lineup. Thus, it should not be used by police departments for adult eyewitnesses. However, it does show that if the elimination lineup is used to make an identification, as long as that identification is made with high confidence it is likely to be accurate. The work in Chapter 5 provides a novel paradigm that can improve confidence scale use in preschool-aged children. This paradigm could be adapted for use as a training tool both in research and in the legal system. Through all of this work, my dissertation has deepened our understanding of confidence, uncertainty, and their relationship to memory accuracy.

References

Bonawitz, E.B., Fischer, A., & Schulz, L. (2012) Teaching 3.5-year-olds to revise their beliefs given ambiguous evidence. *Journal of Cognition and Development, 13*(2), 266-280.

Boutet, I. & Faubert, J. (2006) Recognition of faces and complex objects in younger and older adults. *Memory & Cognition, 34*(4), 854-864.

Hiller, R.M. & Weber, N. (2013). A comparison of adults' and children's metacognition for yes/no recognition decisions. *Journal of Applied Research in Memory and Cognition, 2*, 185-191.

Hembacher, E. & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*(9), 1768-1776.

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.

Nguyen, T.B., Pezdek, K., & Wixted, J.T. (2016). Evidence for a confidence-accuracy relationship in memory for same- and cross-race faces. *Quarterly Journal of Experimental Psychology, 70*(12), 2518-2534.

Rhodes, M. & Wellman, H. (2013). Constructing a new theory from old ideas and new evidence. *Cognitive Science, 37*(3), 592-604.

Vaish, A., Grossman, T. & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin, 134*(3), 383 403.

Weber, N. & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology, 88*(3), 490-499.

Wilkinson, D.A., Best, C.A., Minshew, N.J., & Strauss, M.S. (2010) Memory awareness for faces in individuals with autism. *Journal of Autism and Developmental Disorders, 40*, 1371-1377.

Wixted, J.T. & Wells, G.L. (2017). The Relationship between Eyewitness Confidence and Identification Accuracy: A New Synthesis. *Psychological Science in the Public Interest*, *18*, 10-65.