On the Assessment of Witnesses' Memory for Events

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Berenike Waubert de Puiseau
aus Mainz am Rhein

Düsseldorf, Januar 2016

aus dem Institut für Experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent:	Prof. Dr. Jochen Musch
Koreferentin:	Prof. Ute Bayen, Ph.D.
Tag der mündlichen Prüfung:	11. März 2016

“Remembrance of things past is not necessarily the remembrance of things as they were.”

Marcel Proust (1871–1922)

For my grandparents

Acknowledgments

First and foremost, I would like to thank my supervisor *Jochen Musch* for the support he has given me throughout the course of conducting and writing up this research. I very much appreciated the trust that he has put into me, and the freedom he has given me to pursue my research ideas and my interests. It has been an honor to be allowed to benefit from his expertise and creativity.

Moreover, I would like to thank *Ute Bayen* for serving as my second supervisor and my mentor. I appreciate her interest in my research and I am thankful for her support in finishing this doctoral thesis.

I would not be able to write these acknowledgments if it was not for *Edgar Erdfelder*, who for many years has served as a mentor to me. He has inspired and motivated me with his ideas, his support and his friendliness. I have always felt lucky knowing that I could count on his advice and I thank him for all his support in the last years.

Another person also deserves special mentioning in these acknowledgments. Since the days when we were both student research assistants at the University of Mannheim, *André Aßfalg* has spent hours discussing research ideas with me. Moreover, he has shared much laughter and many cartoons with me that have always lightened up my working days. I thank him for sharing his expertise with me, for his encouragement, his understanding, his support, and his friendship.

Many people have given me advice, have listened to my concerns, and have answered my questions throughout the past years. In particular, I would like to thank *Monika Undorf* and *James Sauer* for their support and for sharing their expertise on metacognition with me. Moreover, I thank *Derek Perkins* for his genuine interest in my research and for sharing his expertise on forensic psychology with me.

Creating a crime simulation that needed to meet what felt like a billion requirements is a complex and time-consuming endeavor. Most of all, I would like to thank *Arvid Hofmann* for his endless patience and his impressive perfectionism in finalizing the video. Moreover, I would like to thank *Carolin Meschede*, *Thorsten Meschede*, *Jana Einicke*, *Tim Eichhorn*, and *Richard Barker* for their willingness to first spend one day freezing in a rather uninspiring area of Düsseldorf and to then spend another day at our lab repeating the very same sentences over and over again. I would also like to thank *Sven Platzek*, *Anneke Weide*, *Svenja Jessica Löffler*, *Arne Stops*, and *Julia Meisters* who supported me as co-authors, thesis students, research interns, and research assistants.

Despite their own heavy workloads, two people proofread this thesis in an incredibly short amount of time. I would like to thank *Janin Rössel* and *Kate O'Brien* for their thorough revisions and insightful comments that helped me finalize this thesis.

Moreover, I would like to thank my colleagues at the Department for Psychological Assessment and Differential Psychology for their feedback on my research and for making the

time in Düsseldorf worthwhile: *Sabine Hillebrandt*, *Adrian Hoffmann*, *Frank Calio*, *Martin Papenberg*, *Meik Michalke*, *Martin Ostapczuk*, *Birk Diedenhofen*, and *Jana Sommer*.

Many people have accompanied me in the past years that have lightened up my life. I would like to thank *Anke*, *Janin*, *Mariela*, *Stefan*, *Ben*, *Christoph*, *Kate*, *Deborah*, *Simona*, *Alexandra*, *Martin*, and *Marc-Oliver* for helping me procrastinate and for helping me focus – I am very lucky to have you in my life.

Last, but not least, I thank *my parents* who have supported me and encouraged me to pursue my goals.

I dedicate this dissertation to my grandparents, who are not here anymore to see me reaching this milestone and to read the outcome of what I have put so much work and thought into during the past years. I am thankful for the memories of the time we spent together and for the love and support they gave me.

Zusammenfassung

Zeugenaussagen sind oft fehlerbehaftet, üben jedoch großen Einfluss auf juristische Entscheidungsfindungsprozesse aus. Nicht alle Faktoren, die sich negativ auf die Güte von Zeugenaussagen auswirken, unterliegen der Kontrolle durch die Beteiligten. Deshalb ist es wichtig, die Akkuratheit von Zeugenaussagen so zuverlässig wie möglich zu bestimmen. Forschung zur Güte von Zeugenaussagen beschäftigt sich vorwiegend mit der Untersuchung von Effekten, wohingegen die Entwicklung von Theorien, die den Effekten zugrunde liegen, vernachlässigt wird. Die vorliegende Arbeit beschäftigte sich mit zwei Ansätzen, die Akkuratheit von Zeugenaussagen zuverlässig vorherzusagen: einerseits wurden subjektive Sicherheitsurteile als Prädiktor für die Akkuratheit von Zeugenaussagen untersucht, andererseits die Überlegenheit von aggregierten im Vergleich zu individuellen Aussagen bestimmt. Beiden Ansätzen wurden Gedächtnismodelle zugrunde gelegt. Da nur wenig über die Wahrhaftigkeit des subjektiven Sicherheitsurteils bei Ereignisgedächtnis bekannt ist, wurde zunächst eine Meta-Analyse über acht Studien mit insgesamt 24 unabhängigen Bestimmungen der Kalibrierung, Unter-/Überschätzung sowie Diagnostizität des subjektiven Sicherheitsurteils durchgeführt (Studie 1). Die durchschnittliche subjektive Sicherheit der Zeugen [i] korrespondierte mit dem tatsächlichen Anteil korrekter Antworten (gute Kalibrierung), jedoch neigten Zeugen dazu ihre Kompetenz leicht zu überschätzen. Die Fähigkeit anhand des subjektiven Sicherheitsurteils zwischen korrekten und falschen Aussagen zu differenzieren war sehr gering ausgeprägt (niedrige Diagnostizität). Die Meta-Analyse ergab ferner, dass die meisten der in der Meta-Analyse berücksichtigten Studien die gleiche Methodik in Form des gleichen Stimulusmaterials und des gleichen Itemtypus verwendeten. Eine Moderatoranalyse ließ den Schluss zu, dass sich die Methodik systematisch auf die Wahrhaftigkeit von Sicherheitsurteilen auswirkte, was die Generalisierbarkeit der Ergebnisse der Meta-Analyse einschränkte. Darüber hinaus wurde eine publikationsbedingte Verzerrung der Effektstärkeschätzer für Kalibrierung und Diagnostizität gefunden. Dass Menschen die Akkuratheit ihrer Entscheidungen unter Unsicherheit überschätzen, wird oft beobachtet und mehrere Theorien zur Erklärung des Phänomens wurden vorgeschlagen. Eine dieser Theorien, MINERVA-Decision-Making (MDM; Dougherty, 2001; Dougherty, Gettys, & Ogden, 1999), wurde in Studie 2 angewendet, um den Einfluss von Skripten und reduzierter Arbeitsgedächtniskapazität auf die Überschätzung der Akkuratheit der eigenen Zeugenaussagen zu erklären. Wie von dem Modell vorhergesagt, überschätzten alle Zeugen ihre eigene Leistung. Die Überschätzung war besonders stark ausgeprägt, wenn Details in den Gedächtnisspuren zuvor beobachteter Verbrechen (z. B. aus Filmen oder Büchern) vergleichbaren Details des untersuchten Verbrechen widersprachen und somit zu einem falschen Gefühl von Vertrautheit führten.

[i] Die Verwendung der männlichen Form schließt hier und im Folgenden sowohl Frauen, Männer, als auch Personen ein, die sich weder dem weiblichen, noch dem männlichen Geschlecht zuordnen.

Unter diesen Umständen gaben Zeugen falsche Antworten mit überzufällig hoher subjektiver Sicherheit. Die Ergebnisse von Studie 2 legten nahe, dass MDM die dem Zeugengedächtnis zugrunde liegenden Prozesse zuverlässig beschreibt und somit Vorhersagen über das Auftreten von Selbstüberschätzung ermöglicht. Jedoch liegen bei juristischen Ermittlungen nicht immer subjektive Sicherheitsurteile vor, anhand derer die Akkuratheit von Zeugenaussagen beurteilt werden kann. Außerdem können subjektive Sicherheitsurteile durch die Anwesenheit anderer Zeugen verzerrt werden. Vor dem Hintergrund, dass die meisten Verbrechen durch mehrere Zeugen beobachtet werden, ist dies besonders problematisch. Aus diesem Grund wurde in Studie 3 die Validität aggregierter Verbrechensrekonstruktionen untersucht. Dafür wurde die Performanz von zwei Aggregationsmethoden, der einfachen Mehrheitsregel und des General Condorcet Modells (GCM; Karabatsos & Batchelder, 2003; Romney, Weller, & Batchelder, 1986), das individuelle Unterschiede in Kompetenz und Ratetendenz sowie verschiedene Frageschwierigkeiten berücksichtigt, mit den Aussagen einzelner Zeugen verglichen. Zusätzlich wurde berücksichtigt, ob die Kompetenzen der Zeugen, deren Aussagen aggregiert wurden, heterogen waren. Aggregierte Verbrechensrekonstruktionen waren stets akkurater als einzelne Zeugenaussagen. Die Validität der Rekonstruktionen wurde maximiert, wenn die Kompetenzen der Zeugen, deren Aussagen aggregiert wurden, heterogen waren und das GCM verwendet wurde. Die Validität von auf der Mehrheitsregel basierenden Aggregationen wurde von der Kompetenzheterogenität kaum beeinflusst. Die Ergebnisse der drei Studien legen nahe, dass (a) Zeugen grundsätzlich in der Lage sind, die Akkuratheit ihrer Aussagen einzuschätzen, jedoch nicht zwischen einzelnen korrekten und falschen Antworten unterscheiden können, dass (b) MDM die Gedächtnisprozesse von Zeugen akkurat repräsentiert, woraus geschlossen werden kann, dass die Überschätzung der eigenen Leistung ein allgegenwärtiges Phänomen ist, das von automatischen Prozessen herrührt und für untypische Verbrechen maximiert wird, sowie dass (c) Gruppen von Zeugen eine bessere Rekonstruktion der untersuchten Verbrechen als einzelne Zeugen erlauben und es deshalb sinnvoll ist, die Aussagen mehrerer Zeugen zu aggregieren, insbesondere wenn ihre Kompetenzen heterogen sind und die verwendete Aggregationsmethode diese Heterogenität berücksichtigen kann, wie dies beim GCM der Fall ist. Zusammenfassend lässt sich auf Basis der vorliegenden Arbeit festhalten, dass bislang kaum für die Forschung verwendete Gedächtnismodelle wie MDM und das GCM für die Untersuchung, das Verständnis und die Beurteilung von Zeugenaussagen hilfreich sind.

Schlüsselbegriffe: Zeugengedächtnis; Wiedererkennung; Meta-Analyse; Kalibrierungsanalyse; Wahrhaftigkeit von subjektiven Sicherheitsurteilen; MINERVA-Decision-Making; Aggregation; General Condorcet Modell; Mehrheitsregel

Abstract

Witness memory has been found to be unreliable, yet highly influential with regard to legal decision-making. Many factors that impair witness memory are not under the control of actors of the legal system. It is therefore important to maximize the validity of assessments of the accuracy of witnesses' testimonies. Research investigating witness memory has largely focused on effects and has neglected theory development. The present doctoral thesis investigated two approaches to assessing the accuracy of witnesses' reports: through confidence ratings indicating witnesses' subjective certainty that their reports are correct, and through aggregating multiple witness reports. For both approaches, theories that model cognitive processes underlying witness memory were proposed. Because little is known about the magnitude of confidence calibration, under-/overconfidence, and confidence resolution – measures that are commonly referred to as the realism of confidence – in witnesses' memory for events, a meta-analysis of eight studies containing 24 independent assessments of these measures was conducted (Study 1). Witnesses' mean confidence ratings were found to correspond rather well with their testimony's accuracy (good calibration), but witnesses were also found to overestimate their own performance (some overconfidence). Moreover, results indicated that witnesses were barely capable of discriminating between correct and incorrect responses (poor resolution). Generalizability of the findings was found to be limited because almost all studies included in the meta-analysis employed the same study method (stimulus material and item type) that could be shown to be a significant moderator of calibration, under-/overconfidence, and resolution. Moreover, a small publication bias was detected for calibration and resolution. According to the meta-analysis, witnesses' mean confidence ratings exceeded the accuracy of their reports. Overconfidence has been found to be a common phenomenon when people make judgments under uncertainty and several theories explaining its causes exist. For this reason, Study 2 focused on understanding the cognitive processes underlying overconfidence in witnesses' memory for events. A theory from the judgment and decision-making domain, MINERVA-Decision-Making (MDM; Dougherty, 2001; Dougherty, Gettys, & Ogden, 1999), was applied to model the impact of witnesses' scripts and working memory load on overconfidence. In line with the model's predictions, overconfidence was found to be a ubiquitous phenomenon in eyewitnesses' memory for events and to be particularly high when previously observed crimes led to a false feeling of familiarity for the probed details that resulted in an incorrect response and a confidence rating exceeding the level of chance. MDM thus proved to be a viable model of witnesses' memory for events. Confidence ratings are however not always available to help judge the accuracy of witnesses' testimonies, and particularly when other witnesses are present, confidence ratings may be distorted. This is problematic because most crimes have been found to feature multiple witnesses. However, some studies suggest that aggregating testimonies may produce accurate reconstructions of crimes. For this reason, in Study 3, the validity of aggregated crime reconstructions was assessed. Validities of crime reconstructions based on two

aggregation rules, the simple Majority Rule and the General Condorcet Model (GCM; Karabatsos & Batchelder, 2003; Romney, Weller, & Batchelder, 1986), were compared to each other and to individual testimonies as a function of heterogeneity in witnesses' levels of competence. Other than the Majority Rule, the GCM can take differences in competences and guessing biases between witnesses and in difficulties between items into account. Aggregation-based crime reconstructions were always superior to individual crime reconstructions. The validity of aggregation-based crime reconstructions was maximized when witnesses' levels of competence were heterogeneous and the GCM was employed. The validity of crime reconstructions based on the Majority Rule was barely affected by competence heterogeneity. The results of the three studies suggest that (a) witnesses are capable of monitoring the overall accuracy of their reports, but are unable to distinguish between correct and incorrect responses, that (b) MDM is a viable model of witnesses' memory for events and respective confidence ratings, implying that overconfidence results from automatic processes, is a rather ubiquitous phenomenon, and is maximized when observed crimes are in conflict with existing scripts, and that (c) groups of witnesses provide more accurate crime reconstructions than individual witnesses and aggregating multiple witness reports is particularly useful when witnesses' competence levels are heterogeneous and when aggregation rules such as the GCM are employed that can account for individual differences between witnesses and items. In conclusion, the present doctoral thesis provided empirical evidence for the viability of memory models – such as MDM and the GCM that both have rarely been used in forensic psychological research – in understanding cognitive processes underlying witnesses' memory for events.

Keywords: Witness memory; recognition memory; meta-analysis; calibration analysis; the *realism of confidence; confidence-accuracy relation; MINERVA-Decision-Making; aggregation; General Condorcet Model; Majority Rule*

Table of Contents

Acknowledgments I

Zusammenfassung III

Abstract V

List of Tables VIII

List of Figures IX

1. Introduction 1

2. Assessing Witness Memory Through Confidence Ratings 7

2.1. Calibration Analysis 9

2.2. Study 1: A Meta-Analysis of the Realism of Confidence in Eyewitness Event Memory 12

2.3. A Theoretical Analysis of the Confidence-Accuracy Relationship 17

2.4. Study 2: How Scripts Influence the Overconfidence in Eyewitness Event Memory: A Model-Based Analysis 19

2.5. Discussion of the Confidence-Accuracy Relationship 29

3. Assessing Witness Memory by Aggregating Multiple Testimonies 32

3.1. Study 3: On the Importance of Considering Heterogeneity in Witnesses' Competence Levels When Reconstructing Crimes from Multiple Witness Testimonies 36

3.2. Discussion of the Aggregation Approach 42

4. General Discussion 44

4.1. Future Directions 47

4.2. Practical Implications 53

References 58

Appendix 77

Versicherung an Eides Statt 205

List of Tables

Table 1 *2x2 Contingency Tables and the Results of the McNemar Tests Comparing the Majority with the Consensus Reconstructions for the Homogeneous versus Heterogeneous Samples Separately for Different Sample Sizes (a: 10, b: 20, and c: 40)* 41

List of Figures

Figure 1. Mean effect size estimates and their 95% confidence intervals for a) calibration, b) under-/overconfidence, and c) resolution across all studies (blue bars), across studies that employed the video first used by Granhag (1997) and that used 2AFC items (grey bars), and across studies that employed other stimulus materials and that used T/F items (green bars). 15

Figure 2. Schematic of the MDM applied to eyewitness event memory; depiction of direct retrieval (blue) and indirect, familiarity-based retrieval (green) regarding a specific crime observed (Crime G) when the witness has previously observed and encoded six crimes (A, B, C, D, E, F) that are taken into account when determining the familiarity of a recognition item. 21

Figure 3. Accuracy and mean confidence (and their standard errors) for script-conforming, script-neutral, and script-nonconforming items (within-subjects) by working memory load (no working memory load vs. working memory load; between-subjects). 25

Figure 4. Mean overconfidence (and its standard error) for script-conforming, script-neutral, and script-nonconforming items (within-subjects) by working memory load (no working memory load vs. working memory load; between-subjects). 27

Figure 5. Mean proportions (and their standard errors) of agreement between the true answer key on the one hand and the answer key estimates that were based on the Majority Rule (majority reconstruction, green lines), the GCM (consensus reconstruction, blue lines), and the individual responses (individual reconstruction, red lines) on the other hand as a function of competence heterogeneity (a: homogeneous competences; b: heterogeneous competences) and the number of witnesses (n = 10, 20, and 40). 40

Figure 6. Schematic of MDM (blue) and 2-HTM (green) combined into one memory model for a true statement (for false statements, direct retrieval would result in a "False" response; familiarity-based retrieval with unequal familiarities would result in a "True" response if the statement appeared familiar, and in a "False" response if the statement appeared unfamiliar); d_i denotes a witness's competence, g_i denotes a witness's tendency to guess "true" when the witness does not know the correct response. 47

"Did you realize what had happened when you heard the shots? Did people lie down on the ground? You did lie down? Did Mrs. Kennedy scream on the first shot? And the President fell into Mrs. Kennedy's arms?"

Questions asked by Bill Lord, ABC Television Network, when interviewing Mary Moorman, a witness to the assassination of John F. Kennedy[1]

1. Introduction

In legal investigations, witness testimony is almost always taken into consideration as a highly relevant and important piece of evidence (Wells, Memon, & Penrod, 2006). Witness testimony is central to the reconstruction of crimes: "People pay attention to what a witness says, and from a witness's report they decide what reality is" (Loftus, 1996, p. 12). Mirroring the importance of witness testimony, a large body of research investigating witness memory has grown over the past four to five decades. This research has been characterized by two central findings. On the one hand, hundreds of studies have empirically and consistently found that witness memory is fallible. On the other hand, witness testimony has commonly been found to be perceived as a reliable and trustworthy piece of evidence.

Münsterberg's (1908) studies showing that students were unable to perform simple memory tasks marked the beginning of research on witness memory. Several decades later, Buckhout (1974) was able to replicate and extend Münsterberg's findings when he demonstrated that almost two-thirds of over 100 students who witnessed a staged assault on a university campus were unable to identify the perpetrator from a photo lineup. Today, over 100 years after Münsterberg's pioneering work and over 40 years after the rebirth of experimental eyewitness research, there is a general acceptance among psychologists that witness memory is far from perfect and that witness testimony is likely to be flawed (Loftus,

[1] Retrieved from https://youtu.be/YEavxZReo84 on January 08, 2016

1996; Memon, Mastroberardino, & Fraser, 2008; Turtle, Read, Lindsay, & Brimacombe, 2008; Wells et al., 2006; Wells & Olson, 2003).

Despite the empirical evidence that witness memory is fallible, both lay people and legal experts still believe witness memory to be highly reliable (e.g., Kassin, Tubb, Hosch, & Memon, 2001; Wise, Pawlenko, Safer, & Meyer, 2009; Wise & Safer, 2004). Among 1,838 participants of a representative telephone survey in the U.S. American population, over 60% strongly or mostly agreed that human memory works like a tape recorder. Almost half of the participants believed that memory for an event is permanent and therefore does not change once it has been stored (Simons & Chabris, 2011). Agreement with both statements was negatively related to education level. Nevertheless, of the participants who had completed graduate school 47% agreed with the tape recorder statement, and 41% believed that memory was permanent.

Given that people put so much trust in the accuracy of human memory, it is not surprising that witness reports have been found to impact legal decision-making. In one of her pioneering studies, Loftus (1975b) demonstrated the strong influence that the presence of a witness's testimony in a trial case has on jurors' decisions. In her study, 150 mock jurors were asked to decide whether a defendant was guilty of robbing a grocery store and killing the owner and the owner's grandchild. Of the mock jurors who read a case description containing no eyewitness testimony, only 18% voted to convict the defendant. In the group of participants who read a version of the case, in which a witness claimed to have seen the defendant shoot the two victims, the proportion of convictions rose to 72%. A third group also received the case description containing a witness's testimony, but in this group the witness was described as suffering from bad eyesight and, according to the case description, did not wear glasses on the day of the robbery. In this group, still 68% of mock jurors voted to convict the defendant. In sum, when a witness was added to the case, the proportion of

convictions quadrupled. This was independent of whether the witness possessed the physiological and physical requirements to correctly perceive the crime.

For many years, the criminal justice system largely ignored psychological research findings. This rejection began to recede when advances in the analysis of forensic science helped identify and overturn multiple wrongful convictions (Wells et al., 2006). Since the installment of the *Innocence Project* in the early 1990s, a U.S. American organization dedicated to the investigation of alleged wrongful convictions using DNA technology, 340 wrongfully convicted people, who spent on average 14 years in prison, have been exonerated.[2] In over two-thirds (n = 237, 69.7%) of these cases, wrongful eyewitness identification was the single or one of the main contributing factors that led to the wrongful convictions. The Innocence Project thus provides real-world data supporting both notions that witness memory is fallible and that testimony of witnesses strongly influences legal decision-making.

Given the detrimental effects that false witness testimony can have, it is important to accurately assess *testimony accuracy*, that is, the accuracy of witnesses' accounts of an observed crime. When testimony accuracy can be estimated reliably, crime reconstructions based on the respective testimonies should be more accurate. Traditionally, testimony accuracy is assessed in terms of estimator variables or system variables that may benefit or impair testimony (cf. Kassin et al., 2001; Wells, 1978). Estimator variables are factors that cannot be controlled by the legal system and include, for example, whether a weapon was present during the crime (Fawcett, Russell, Peace, & Christie, 2013; Steblay, 1992). In contrast, system variables are determined by the legal system and include, for example, interviewing conditions (Köhnken, Milne, Memon, & Bull, 1999; Memon, Meissner, & Fraser, 2010). The examination of factors that benefit or impair witness memory and testimony is important, but it has two major drawbacks. First and foremost, the list of factors

[2] http://www.innocenceproject.org/, accessed on December 05, 2015

that influence memory accuracy is unlikely to be complete. Rather, it has been suggested that research has not yet managed (and may never be able) to identify all factors that potentially improve or impair testimony accuracy (Turtle et al., 2008). Second, little is known about how these factors interact and studies proposing theories of the cognitive processes underlying the influence of these factors are scarce.

Two alternative approaches to the assessment of testimony accuracy seem promising. A first and rather popular approach to determining testimony accuracy is the assessment of confidence ratings expressing a witness's subjective certainty that a response is correct. Studies investigating the relationship between confidence and accuracy have produced mixed results. Correlation analyses suggest that the relationship between confidence and accuracy is rather weak (e.g., Bothwell, Brigham, & Deffenbacher, 1987; Cutler & Penrod, 1989; Sporer, Penrod, Read, & Cutler, 1995). In contrast, studies assessing the *realism of confidence* by employing calibration analyses that compare absolute levels of accuracy and confidence have suggested that confidence might be a rather valid predictor for eyewitness accuracy (Olsson & Juslin, 2002). Research on the confidence-accuracy relationship has however largely been driven by effects or technical aspects (e.g., whether witnesses choose or refuse to choose a suspect from a group of people they are presented with; cf. Sporer et al., 1995; Weber & Brewer, 2003; 2004; or whether an identification is made directly or in several steps; cf. Weber & Varga, 2012). Theories that corroborate the empirical results are largely lacking (Brewer, Weber, & Semmler, 2007).

A second approach to assessing testimony accuracy and to reconstructing crimes is through *aggregation of multiple testimonies*. Aggregated judgments have consistently been found to outperform individuals in various cognitive tasks (e.g., Davis-Stober, Budescu, Dana, & Broomell, 2014). This is interesting in the context of eyewitness testimony, because an average crime is generally observed by around four witnesses (Paterson & Kemp, 2006; Skagerberg & Wright, 2008). Only three studies have investigated the potential of aggregation

in forensic psychology. In all studies, aggregation was found to be beneficial to assessing testimony accuracy (Clark & Wells, 2008; Sanders & Warnick, 1982; Waubert de Puiseau, Aßfalg, Erdfelder, & Bernstein, 2012), but little is known about the conditions under which the benefit of aggregation can be maximized.

Most of the studies on the confidence-accuracy relationship and two of the three studies examining the validity of the aggregation approach (Clark & Wells, 2008; Sanders & Warnick, 1982) investigated *eyewitness identification decisions*. Identification decisions require witnesses to recognize (usually from line-ups) perpetrators that were previously observed in a crime. Witnesses may select a suspect from the line-up, or reject the line-up altogether (cf. Sporer et al., 1995). A witness makes a correct decision either when selecting the correct suspect from a line-up, in which the perpetrator is included, or when rejecting a line-up, in which the actual perpetrator is not included. *Eyewitness event memory*, in contrast, has been neglected. It refers to actions, persons, conversations, and surroundings of a crime. In a typical study assessing witnesses' memory for events, participants view a simulated crime and subsequently answer for example recognition questions about this crime. Given the obvious differences between identification decisions and the recognition of event details, it seems unclear whether findings from studies of eyewitness identification decisions can be generalized to event memory (Allwood, Knutsson, & Granhag, 2006).

The aim of the three studies presented in this doctoral thesis was to provide insights into how testimony accuracy can be estimated and how crimes can be reconstructed from witnesses' memory for events. The studies concentrated on event recognition memory because *focused questions* (i.e., questions that present witnesses with answer alternatives from which the correct one is to be chosen) about crime details have been found to be rather typical for police interviews (Fisher, Geiselman, & Raymond, 1987; George & Clifford, 1992; Peterson & Grant, 2001; Wright & Alison, 2004). In particular, the studies aimed to

contribute to the understanding of cognitive factors underlying eyewitness memory. Therefore, in two of the studies, memory models were employed.

Chapter 2 provides a discussion of the confidence-accuracy relationship in witnesses' memory for events and presents the research questions and core findings of two studies investigating the realism of confidence in eyewitness event recognition memory (Chapters 2.2 and 2.4). Chapter 3 outlines the potential of the aggregation approach in general and for eyewitness memory in particular. The research questions and core findings of the third study, in which the performance of two aggregation methods is investigated as a function of the heterogeneity of the witnesses' competence levels (i.e., the probability of witnesses providing correct responses), are presented in Chapter 3.1. The doctoral thesis concludes with a discussion of potential future directions and the practical implications of the present research (Chapter 4). The original study manuscripts are presented in the Appendix (Appendices A, B, and C).

2. Assessing Witness Memory Through Confidence Ratings

In the first two parts of this chapter (2.1 and 2.2), different methods to assess the relation between confidence and accuracy in eyewitness memory are discussed. Empirical research findings regarding the relation between confidence and the accuracy of recognition judgments about crime details are presented. The following two parts of the chapter (2.3 and 2.4) focus on the cognitive processes underlying witness recognition memory and respective confidence ratings. The chapter concludes with a discussion of the confidence-accuracy relation in eyewitnesses' recognition memory for events (2.5).

Asking witnesses to rate their confidence in the correctness of their responses and using these ratings to assess the accuracy of witness reports has considerable intuitive appeal. The expectation that confidence ratings and testimony accuracy are closely related arises from the assumption that witnesses know how much they know. It is for this reason that confidence has been suggested as an indicator of testimony accuracy in several jurisdictions, for example in the United States (cf. O'Toole & Shay, 2006; Wells & Murray, 1983).

Both lay people and legal experts have been shown to perceive confidence to be a valid indicator of testimony accuracy. Witnesses who are confident in their own testimonies are often perceived as more trustworthy than witnesses who are more critical about their testimonies (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988; McClure, Myers, & Keefauver, 2013; Potter & Brewer, 1999). In this vein, of the 1,838 U.S. American adults who participated in the study by Simons and Chabris (2011) cited in Chapter 1, one third believed that the testimony of one confident eyewitness would be sufficient to allow a defendant to be convicted.

Confidence has indeed been found to be closely correlated to performance in several domains for example for general knowledge questions (e.g., Perfect, 2002; Perfect & Hollins, 1996). However, several meta-analyses that investigated the association between mean confidence ratings and accuracy of eyewitness identifications (between-subjects correlations

computed across witnesses; cf. Smith, Ellsworth, & Kassin, 1989) have found that correlation coefficients were, at best, of medium size (up to .41; Bothwell et al., 1987; Cutler & Penrod, 1989; Deffenbacher, 1980; Sporer et al., 1995; Wells & Murray, 1987). Similarly, the only existing synthesis of research findings on the relationship between confidence and accuracy for event memory found a mean between-subjects correlation of .21 (Perfect, 2002). These findings have led to the conclusion that only a small part of the variance in accuracy of witnesses' testimonies can be explained by witness confidence.

These findings however do not allow for a concise conclusion regarding the use of witnesses' confidence ratings as a predictor of the accuracy of an individual witness's testimony. This is because correlation coefficients computed across witnesses provide information only about the extent to which mean confidence and accuracy covary. Because low variance always leads to low covariance, the size of the correlation coefficient depends on the amount of variance in confidence and accuracy in the sample under investigation. Thus, a low correlation may either result from a weak association between confidence and accuracy, from a low amount of variance in confidence ratings, from a low amount of variance in accuracies, or from any combinations thereof (Juslin, Olsson, & Winman, 1996). In standard studies of eyewitness testimony, the variances in confidence and accuracy are expected to underestimate the true magnitude of these variances and their covariance that would be expected in real-world witnesses, leading to an underestimation of between-subjects correlation coefficients. This is because eyewitness studies are commonly conducted in laboratories under highly standardized conditions using samples of students with rather similar cognitive abilities (Brewer, 2006; Lindsay, Nilsen, & Read, 2000; Lindsay, Read, & Sharma, 1998). It is for this reason that correlation coefficients were recommended to be computed within-subjects, that is, separately for all witnesses across their responses resulting in one correlation coefficient for each witness (cf. Smith et al., 1989). In a study comparing between-subjects correlations with within-subjects correlations in witnesses' memories for

events, within-subjects correlations have been found to be higher and more stable than between-subjects correlations (Robinson & Johnson, 1996). Other empirical studies investigating within-subjects correlations have found coefficients ranging from .10 (Wheatcroft, Kebbell, & Wagstaff, 2001) to over .60 (Bulevich & Thomas, 2012; Kebbell, Evans, & Johnson, 2010; Kebbell & Giles, 2000) for recognition questions about observed crime events.

However, neither between-subjects nor within-subjects correlations are informative with regard to the assessment of an individual witness's testimony accuracy: "[…] interpretation of the point biserial correlation is not straightforward in the forensic context. For example, it is not clear how knowing that the [confidence-accuracy] correlation is .23 or .37 should contribute to a juror's interpretation of the likelihood that a witness may be accurate when that witness has reported 90% confidence in their identification" (Brewer, 2006, p. 11).

2.1. Calibration Analysis

It is for this multitude of difficulties associated with correlation analyses that calibration analysis was suggested as an alternative assessment of the confidence-accuracy relationship (Lichtenstein & Fischhoff, 1977; Lichtenstein, Fischhoff, & Phillips, 1982; Olsson, 2000; Wagenaar, 1988). Calibration analysis plots absolute levels of confidence against accuracy for each witness and can therefore detect meaningful relationships between these two variables – often referred to as the *realism of confidence* – when correlation analyses suggest that they do not or only slightly covary (Juslin et al., 1996). Three measures are commonly computed to assess the realism of confidence: the calibration coefficient *C*, the under-/overconfidence index *U/O*, and the normalized resolution index *NRI* (e.g., Baranski & Petrusic, 1994; Lichtenstein & Fischhoff, 1977; Yaniv, Yates, & Smith, 1991).

Several authors have provided intelligible explanations of the measures of calibration analysis (e.g., Brewer & Wells, 2006) that the following outline is based on. The calibration coefficient *C* describes the deviation of the observed calibration curve from a perfect 45° calibration curve (plotting accuracy against confidence; the perfect calibration curve, thus, indicates a perfect match of accuracy and confidence). The coefficient ranges from 0 (no deviation and, thus, perfect calibration) to 1 (maximum deviation and, thus, very poor calibration or maximum miscalibration). To compute *C*, the confidence scale is divided into *J* class intervals (commonly 0-10, 11-20, etc.) and each class interval is inspected separately. The following formula is used to compute *C*:

$$C = \frac{1}{n}\sum_{j=1}^{J} n_j \left(c_j - a_j\right)^2, \qquad (1)$$

where n_j is the number of observations in class interval *j*, c_j is the mean confidence level in class *j*, and a_j is the proportion of correct responses in class interval *j*.

The under-/overconfidence statistic *U/O* is computed in a similar manner, except that differences between mean confidence levels c_j and the respective proportions of correct responses a_j are not squared. Consequently, *U/O* ranges from -1 to +1 with figures below 0 indicating underconfidence (witnesses' confidence is lower than their accuracy) and figures above 0 indicating overconfidence (witnesses' confidence exceeds their accuracy).

As a third measure of the realism of confidence, a witness's resolution can be computed. The resolution of witnesses' confidence refers to witnesses' ability to discriminate between their correct and incorrect recognition judgments. The *NRI*, as the measure of resolution, ranges from 0 (no discrimination between correct and incorrect recognition judgments, that is, no resolution) to 1 (perfect discrimination, that is, perfect resolution). To compute the *NRI*, a normalized sum of the squared differences between the proportions of

correct responses a_j in each class interval *j* and the grand proportion of correct responses *a* is computed:

$$NRI = \frac{\frac{1}{n}\sum_{j=1}^{J} n_j \left(a_j - a\right)^2}{a(1-a)} \qquad (2)$$

The interpretation of the effect size of *NRI* is analogous to η^2 (Baranski & Petrusic, 1994); that is, values below .06 may be considered small, and values exceeding .13 can be considered large (Cohen, 1988).

A meta-analysis summarizing 52 empirical assessments of the realism of confidence in eyewitness and earwitness identifications has found that in one third of the included assessments, witnesses were perfectly calibrated (i.e., calibration was not significantly different from 0). Overconfidence for face identifications ranged from .2 to .3. No statistical association between between-subjects correlation coefficients (r = -.11, *ns*) and the respective calibration coefficients or resolution coefficients (r = .31, *ns*) could be observed (Olsson & Juslin, 2002). Due to its obvious advantages over correlation analyses, calibration analysis has become rather popular in eyewitness identification research (e.g., Brewer, Keast, & Rishworth, 2002; Brewer & Wells, 2006; Palmer, Brewer, Weber, & Nagesh, 2013; Semmler, Brewer, & Wells, 2004).

Regarding the realism of confidence in eyewitness event memory (in contrast to identification decisions), several studies can be found in the literature. These studies investigate the influence of a multitude of experimental factors on calibration, under-/overconfidence, or resolution. For example, Allwood, Granhag, and Johansson (2003) found that witnesses were better calibrated and less overconfident when witnesses completed recognition questions and provided confidence ratings jointly with another witness compared with initial individual judgments and confidence ratings. Allwood et al. (2006) gave participants confirmatory or disconfirmatory feedback following responding. The magnitude

of overconfidence after confirmatory feedback was more than twice the magnitude of overconfidence after disconfirmatory feedback. Resolution was reduced only after disconfirmatory, but not after confirmatory feedback. Other studies manipulated whether misinformation was presented (Bonham & Gonzalez-Vallejo, 2009), whether participants selected the items for which they believed calibration to be best (Buratti & Allwood, 2012), or whether and with whom participants discussed their observation prior to being interviewed (Granhag, Jonsson, & Allwood, 2004).

Given the diversity of experimental manipulations, it appears to be difficult to evaluate the overall realism of confidence ratings for witnesses who recognize event details and who indicate their subjective certainty that their recognition judgments are correct. However, in contrast to the meta-analysis on the realism of confidence in eyewitness identification, no study has yet qualitatively or even quantitatively summarized the existing findings on witnesses' memory for events. Study 1 took up this point and aimed to provide a quantitative synthesis of the existing findings on the realism of confidence in eyewitness event recognition memory.

2.2. Study 1: A Meta-Analysis of the Realism of Confidence in Eyewitness Event Memory

Synthesizing empirical research findings facilitates their evaluation and interpretation. Meta-analytic methods are particularly appropriate to summarize research findings as they allow estimating a mean effect size (Green & Hall, 1984). When research findings from multiple studies are quantitatively summarized, results can be based on larger samples and validity of estimations can thus be increased. Meta-analyses also allow identifying potential moderators that influence effect sizes (Rosenthal & DiMatteo, 2001). Moreover, publication bias analyses can be conducted to identify whether the probability that research findings are published is linked to the observed effect size.

It is for these benefits that in Study 1 (Appendix A), a meta-analysis was conducted to determine the magnitude of confidence calibration, under-/overconfidence, and resolution as measures of the realism of confidence in eyewitness event recognition memory and to identify potential publication biases in the articles published on this topic. The studies included in the meta-analysis had to meet four criteria: (a) an objective criterion had to be available to compute the accuracy of witnesses' reports. Therefore, only studies using videos, slides, or stagings depicting criminal behavior were included, and analyses of real-world cases were excluded. (b) Studies had to employ recognition questions (e.g., two-alternative forced-choice questions, yes/no or true/false questions, multiple choice questions with more than two answer options etc.) about the crime event. Studies using free or cued recall questions were excluded, because different cognitive processes may determine performance in recall and recognition memory (cf. Allwood, Innes-Ker, Homgren, & Fredin, 2008). Moreover, recognition questions are commonly employed in police interrogations (cf. Chapter 1). (c) Studies had to employ adult samples, because there are special problems associated with child witnesses (e.g., the stimulus material that child participants can be presented with differs from stimulus materials that can be employed with adult participants; for differences in memory performance between child and adult witnesses, cf. Allwood et al., 2008; Buratti, Allwood, & Johansson, 2014). Studies using samples of elder adults were also excluded because of potential confounds with decreasing or impaired memory (cf. Dahl, Allwood, Scimone, & Rennemark, 2015). (d) Only studies that computed at least one of the three measures of the realism of confidence (calibration, under-/overconfidence, or resolution) were included.

Eight studies containing 24 independent assessments of the confidence-accuracy relationship (henceforth referred to as study units) using 803 participants were identified as suitable for inclusion in the meta-analysis. Raw means were computed as effect size estimates for calibration (*C*), under-/overconfidence (*U/O*), and resolution (*NRI*). Fixed-effect models

were computed to determine mean effect size estimates. Witnesses were slightly, but significantly miscalibrated, $k = 21$, $C = .048$ ($SE = .001$), 95% CI [.046, .051], $z = 44.689$, $p < .001$, and overconfident, $k = 20$, $U/O = .095$ (SE = .004), 95% CI [.088, .102], $z = 27.068$, $p < .001$. Resolution was low, but also significantly different from zero, $k = 14$, $NRI = .025$ ($SE = .001$), 95% CI [.024, .026], $z = 37.733$, $p < .001$.

Six of the eight studies (i.e., 17 of the 24 study units) presented participants with the same crime simulation (a video first used by Granhag, 1997), and used two-alternative forced-choice (2AFC) items. The remaining two studies (7 study units) used different crime simulations and true/false (T/F) items. To determine whether this systematic variation in study methods impacted the realism of confidence, fixed-effect models with study method as moderator variable were computed. Study method significantly impacted confidence calibration, under-/overconfidence, and resolution. All mean effect size estimates are displayed in Figure 1. Witnesses were significantly more miscalibrated, $C = .050$ vs. $C = .044$, $Q(2) = 2004.304$, $p < .001$, and more overconfident, $U/O = .105$ vs. $U/O = .064$, $Q(2) = 758.999$, $p < .001$, in the studies using the video by Granhag (1997) with 2AFC items as compared with the studies using other stimulus materials and T/F items. Participants were also significantly higher in resolution in studies using the video by Granhag (1997) with 2AFC items, $NRI = .033$ vs. $NRI = .021$, $Q(2) = 1499.374$, $p < .001$. The residuals of effect size estimates were significantly heterogeneous for calibration, $Q(19) = 150.597$, $p < .001$, under-/overconfidence, $Q(18) = 92.290$, $p < .001$, and resolution, $Q(12) = 45.913$, $p < .001$.

To examine publication bias, funnel plot asymmetry was tested for statistical significance using Begg's rank correlation test (Begg & Mazumdar, 1994) and Egger's linear regression test (Egger, Davey Smith, Schneider, & Minder, 1997). Egger's linear regression test has been shown to have higher statistical power than Begg's rank correlation test to detect funnel plot asymmetry that can be interpreted as an indicator for publication bias (Macaskill, Walter, & Irwig, 2001; Sterne, Gavaghan, & Egger, 2000). Both tests revealed significant

funnel plot asymmetry and, thus, publication biases for calibration, $\tau = .413$, $p = .009$ and $z = 7.641$, $p < .001$, and resolution, $\tau = .486$, $p < .001$ and $z = 8.300$, $p < .001$. No publication bias was found for under-/overconfidence, $\tau = .137$, $p = .422$ and $z = 1.217$, $p = .224$.

The trim-and-fill method (Duval & Tweedie, 2000a, 2000b) was used to compute corrected effect size estimates. Nine effect sizes were imputed for calibration resulting in a corrected mean effect size estimate of $C = .044$ ($SE = .001$), 95% CI [.042, .046], that remained significantly different from 0, $z = 45.344$, $p < .001$ (for $k = 30$). For resolution, six effect sizes were imputed and the test suggested a corrected mean effect size estimate of $NRI = .021$ ($SE = .001$), 95% CI [.020, .023]. The corrected mean effect size estimate for resolution differed significantly from 0, $z = 35.716$, $p < .001$ (for $k = 20$).

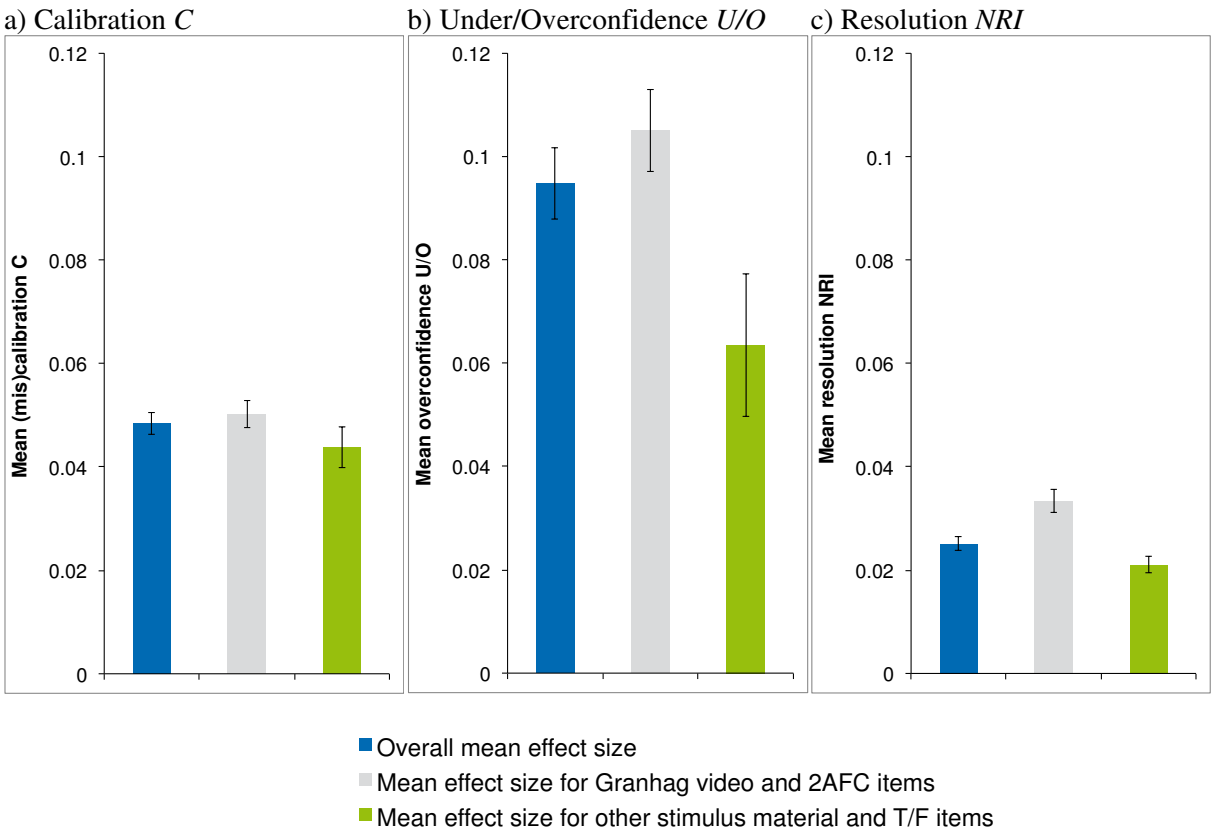


Figure 1. Mean effect size estimates and their 95% confidence intervals for a) calibration, b) under-/overconfidence, and c) resolution across all studies (blue bars), across studies that employed the video first used by Granhag (1997) and that used 2AFC items (grey bars), and across studies that employed other stimulus materials and that used T/F items (green bars).

In summary, witnesses' confidence for events was found to be rather well-calibrated and only slightly, but significantly deviated from the perfect calibration curve. Witnesses tended to overestimate their own performance. Comparing the results from Study 1 to a meta-analysis of the realism of confidence in eyewitness identification decisions suggests that mean overconfidence was however much lower for event memory than for eyewitness identification decisions (cf. Olsson & Juslin, 2002). In contrast to calibration, resolution was very poor. According to the classification proposed by Cohen (1988), the mean effect size of resolution (.021-.025) was small. This pattern of results is at odds with findings from between-subjects and within-subjects correlation analyses. First, between-subjects correlation coefficients are commonly low indicating that witnesses' mean confidence and accuracy do not covary substantially (e.g., Perfect, 2002; Perfect & Hollins, 1996). According to the present meta-analysis, witnesses' confidence ratings are however well-calibrated and approximate accuracy. Second, within-subjects correlations, that have previously been interpreted as indicators of resolution (Higham, Luna, & Bloomfield, 2011), have been found to be higher (Bulevich & Thomas, 2012; Kebbell et al., 2010; Kebbell & Giles, 2000) and also more stable than between-subjects correlations (Robinson & Johnson, 1996). The mean effect size estimate for resolution obtained in the present meta-analysis was however very small indicating that witnesses are not able to discriminate between their correct and incorrect responses. The result pattern observed in this meta-analysis (good calibration, poor resolution) is thus the opposite of the pattern that correlation analyses have commonly found (low and unstable between-subjects correlations, medium and stable within-subjects correlations).

The conclusions that can be drawn from the results of this meta-analysis are limited because almost all of the studies included in this meta-analysis employed the same stimulus material and item type. Moreover, a moderator analysis revealed that study method significantly affected calibration, under-/overconfidence, and resolution. It is unknown

whether the findings obtained in the meta-analysis would replicate in studies using other materials. The results of this meta-analysis can therefore not readily be generalized to eyewitness event memory in general. However, result patterns were similar in the studies using the video first employed by Granhag (1997) and in those using other stimulus materials. Therefore, the overall pattern observed (good calibration, overconfidence, and low resolution) might in fact generalize to other stimulus materials, but no predictions can be made about the magnitude of the effects.

The present meta-analysis was the first quantitative summary of research findings on the realism of confidence in eyewitnesses' memory for events. The aim of this analysis was to understand whether and how confidence and accuracy are related. It however remains unknown when and why witnesses are miscalibrated, what causes witnesses to overestimate their performance, and how resolution might be improved. Future studies should investigate potential moderators that influence the magnitude of calibration, under-/overconfidence, and resolution and that provide insight into the cognitive processes underlying the realism of confidence.

In the following, overconfidence was chosen to be investigated in more detail, because it has been found to be a ubiquitous phenomenon in human judgment under uncertainty (Moore & Healy, 2008). Moreover, as detailed in Chapter 2.4, theories explaining its underlying cognitive processes have been proposed in other research domains. This is appealing, because, as discussed in the next section (2.3), existing research on eyewitness memory has been criticized for lacking theories. Then, a study proposing and testing a theory of the cognitive processes underlying overconfidence in witnesses' memory for events is presented (2.4).

2.3. A Theoretical Analysis of the Confidence-Accuracy Relationship

The meta-analysis presented as Study 1 revealed that witnesses are commonly overconfident in their recollections of crime events. However, due to the restricted diversity of study method employed in the samples included in the meta-analysis, the results could not provide information on the factors that lead to overconfidence. Understanding cognitive processes is important for research findings to be generalized. Nonetheless, research on forensic psychology in general and on eyewitness memory in particular has largely been effect-driven (Clark, 2008; Lane & Meissner, 2008; Ogloff, 2000; Turtle et al., 2008). Studies have commonly focused on outcomes instead of processes and technical aspects, for example when and how confidence ratings should be collected, have been in the focus of most empirical studies (Brewer et al., 2007). Less effort has been spent on researching cognitive processes underlying, for example, the relationship between confidence and accuracy. While considerable advances regarding policy issues have been made based on findings from existing studies (Brewer, 2006; Brewer et al., 2007), a gap has grown between theoretical research on the one hand and the application of psychological findings on the other hand (Lane & Meissner, 2008).

Developing and testing theories is essential to answering applied research questions and to conciliating critics who claim that findings from forensic psychological research are not relevant to the assessment of individual witnesses. Such critics usually argue that findings from psychological experiments cannot be generalized to individual witnesses who are unique. Moreover, critics hold that forensic psychology is unable to investigate all potential factors that influence witnesses' memory (the criticm and counter-arguments are outlined for exampled in Clark, 2008 and Turtle et al., 2008). While this criticism is at least in part justified, theoretical approaches may rectify matters and may accommodate the concerns of applied researchers and practitioners. This is because extrapolation from studied to unstudied people and from studied to unstudied environments is facilitated by modeling cognitive processes underlying witness memory (Bjork, 1973; Clark, 2008; Hintzman, 1991). With

regard to the realism of confidence, this means that understanding the cognitive processes involved is essential for making predictions about how various (non-studied) factors influence witnesses' reports, witnesses' respective subjective certainty, and the relation between the two variables. Developing theories would, thus, also inform the assessment of individual witnesses' testimony accuracy (Brewer, 2006).

As Study 1 revealed, witnesses are overconfident in their recollections of crime event details. Overconfidence is a common phenomenon in cognitive psychology. It has been observed for example in lawyers' predictions of case outcomes (Goodman-Delahunty, Granhag, Hartwig, & Loftus, 2010), in betters' predictions of sports results (Towfigh & Glöckner, 2011), and in weather forecasts (Tyszka & Zielonka, 2002). Several theories have been proposed to explain why and when overconfidence occurs in judgment and decision-making. Little effort has been spent on applying these theories to explain overconfidence in witnesses' memory for events. The aim of Study 2 was to fill this gap.

2.4. Study 2: How Scripts Influence the Overconfidence in Eyewitness Event Memory: A Model-Based Analysis

In Study 2 (Appendix B), *MINERVA-Decision-Making* (MDM; Dougherty, 2001; Dougherty, Gettys, & Ogden, 1999), a variant of the exemplar-based MINERVA-2 memory model (Hintzman, 1984, 1988), was proposed as a model of overconfidence in eyewitnesses' memory for events. MDM seemed particularly appropriate to model witnesses' recognition memory for events for two reasons. First, memory for observed crimes is likely to be very complex. Several studies have shown that when details of an event are linked non-linearly and multiplicatively, exemplar-based models provide more appropriate descriptions of memory representations than cue-based models (Bonham & Gonzalez-Vallejo, 2009; Juslin, Karlsson, & Olsson, 2008; Juslin, Olsson, & Olsson, 2003; Karlsson, Juslin, & Olsson, 2008). The second reason why MDM appeared particularly suitable to describe the confidence-accuracy

relationship in eyewitnesses' event memory was because MDM is capable of modeling scripts and schemata (Hintzman, 1986). Scripts have been found to influence memory in general (Abelson, 1981; Hudson, Fivush, & Kuebli, 1992), and witnesses' memory for events in particular (García-Bajos & Migueles, 2003; García-Bajos, Migueles, & Aizpurua, 2012; Greenberg, Westcott, & Bailey, 1998; Holst & Pezdek, 1992; Tuckey & Brewer, 2003a). More precisely, scripts have been found to be used to interpret ambiguous information and to fill gaps in memory representations, for example when processing load during encoding was high (Hashtroudi, Mutter, Cole, & Green, 1984; Kleider, Pezdek, Goldinger, & Kirk, 2008; Macrae, Hewstone, & Griffiths, 1993). Moreover, confidence in responses to highly typical details has been found to exceed confidence in details of low typicality (García-Bajos et al., 2012).

MDM assumes that every observation is stored as a degraded copy in a memory trace. Highly degraded memory traces contain many details that were not correctly encoded as present or absent, but instead are undefined. The traces stored in memory of a particular type of event, for example a crime, form people's scripts of this type of event (Hintzman, 1986). When requested to recognize a specific detail of an observed crime (e.g., "And the President [John F. Kennedy] fell into Mrs. Kennedy's arms?"), a witness may retrieve the response directly from memory (e.g., the witness remembers seeing John F. Kennedy slump down in his seat). Responses retrieved directly from memory are associated with maximum confidence. If crime details can however not be accessed directly, retrieval has to be based on the perceived familiarity of the crime detail with details in all memory traces that are identified as being similar to the crime under investigation. In this case, confidence ratings regarding the accuracy of recognition judgments are assumed to be proportional to perceived familiarities and, thus, to the frequency of their occurrence in the existing memory traces. When the presence and the absence of a detail are equally familiar, responses are determined

randomly and confidence is at the level of chance. Figure 2 depicts a schematic of both direct and indirect, familiarity-based retrieval.

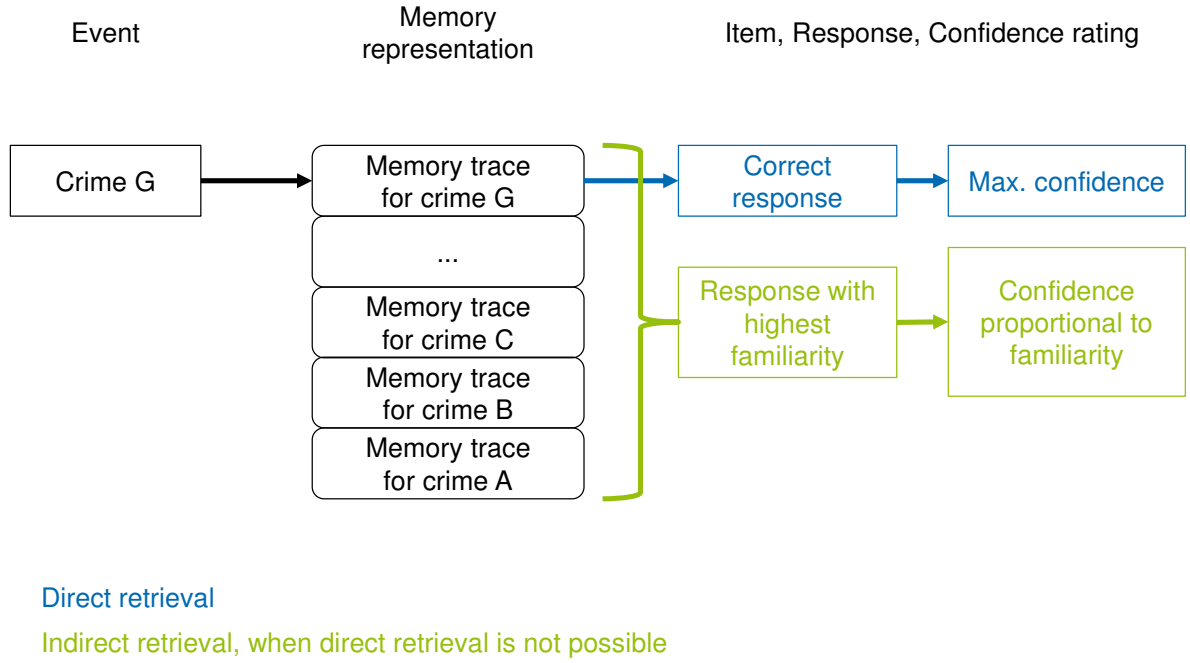


Figure 2. Schematic of the MDM applied to eyewitness event memory; depiction of direct retrieval (blue) and indirect, familiarity-based retrieval (green) regarding a specific crime observed (Crime G) when the witness has previously observed and encoded six crimes (A, B, C, D, E, F) that are taken into account when determining the familiarity of a recognition item.

According to MDM, overconfidence occurs either as a result of misleading familiarity or due to error variance in the memory traces. Misleading familiarity may for example occur when a particular detail appears familiar based on existing memory traces of similar crimes (i.e., the scripts) even though it was not present in the crime under investigation (or vice versa). Because confidence ratings are proportional to perceived familiarity, a detail is incorrectly rejected or incorrectly accepted with confidence exceeding the level of chance. Error variance in memory traces is assumed to be large when memory traces are highly degraded, for example because people devoted limited attention to the crime observation or because details were obscured due to perspective, which is both likely to occur when people witness crimes. Error variance also occurs when the crime of interest and the crimes that were previously observed (for example in movies or books) and stored were highly variable. If this

is the case, many details of observed crimes are not encoded in the respective memory traces, that is, the memory traces are highly degraded. Given the complexity and, thus, variability of crimes, error variance is expected to be particularly high in memory traces of observed crimes.

From MDM, predictions can be derived about how scripts influence accuracy, confidence, and overconfidence. Study 2 tested these predictions in an eyewitness simulation experiment. Seventy-nine students from the University of Düsseldorf viewed a crime simulation showing a young man being robbed by two members of a gang. The crime contained details that were independent of, in line with, or in conflict with existing scripts of a robbery. Following the video, participants completed 102 T/F items that pertained to details that were in line with, in conflict with, or independent of scripts, and provided confidence ratings for each response.

A 2 x 3 mixed-factorial design was employed. To enhance script-based processing, one half of the sample completed a distractor task while watching the video (working memory load condition), whereas the other half completed no additional task (no working memory load control condition). Under working memory load, participants' encoding was impaired. Participants therefore were assumed to make fewer direct retrievals and base their responses on perceived familiarity more often. Script conformity of the items was manipulated within-subjects. One third of the items were script-conforming (i.e., script-based processing resulted in a correct response), another third were script-nonconforming (i.e., script-based processing resulted in an incorrect response), and another third were script-neutral (i.e., script-based processing was equally likely to result in a correct or an incorrect response). The items were pretested for typicality and item difficulty. Because overconfidence has been shown to increase with item difficulty (Juslin, 1993, 1994), a pretest was conducted to determine script conformity of the items and to remove differences in item difficulty between the item sets when no additional working memory load was applied (for more details, see Appendix B).

Thus, no differences in difficulty between the item sets were to be expected in the no working memory load control condition. Accuracy was computed as proportion of correct responses (with 50% denoting chance level). Overconfidence was computed by subtracting accuracy from mean confidence (on a scale from 50 to 100) separately for each participant.

The predictions derived from MDM about the impact of working memory load and script conformity of the items on accuracy, confidence, and overconfidence are detailed in Appendix B. In the following, core predictions regarding accuracy (a), confidence (b), and overconfidence (c) are displayed, tested, and discussed. All predictions were based on the assumption that when the video was viewed under working memory load, encoding would be impaired. Witnesses would, thus, make fewer direct retrievals and would have to base their retrievals on perceived familiarity more often (but nevertheless may still make at least some direct retrievals). MDM predicted perceived familiarity to always be higher for the correct response for script-conforming items. For script-neutral items perceived familiarity was expected to be non-diagnostic with regard to the correct response. For script-nonconforming items, perceived familiarity was expected to be misleading, that is, familiarity-based responses were expected to be incorrect. Confidence was predicted to be lower when recognition judgments were based on perceived familiarity compared to direct retrievals. Confidence for familiarity-based retrievals was however expected to exceed chance level for script-conforming and script-nonconforming items, regardless of whether the response given was correct, because the existing memory traces were expected to make the presence (or absence) of probed crime details appear more familiar than their absence (or presence). For script-neutral items, confidence was expected to approach the level of chance for familiarity-based retrievals, because based on the existing memory traces, neither the presence or absence of crime details was expected to appear particularly familiar.

(a) Accuracy: In the working memory load condition compared to the no working memory load control condition, MDM predicted accuracy to decrease only for script-neutral

and script-nonconforming items, because for script-conforming items, familiarity-based recognition judgments were expected to be correct. Descriptive statistics can be inspected in Figure 3. A 2 x 3 mixed factorial ANOVA with Greenhouse-Geisser corrections for all tests involving the within-subjects factor script conformity was computed. In the following, only results pertaining to the predictions outlined above are presented. The interaction between the working memory load and the script conformity manipulation was statistically significant, $F(1.85, 142.72) = 9.66$, $p < .001$, $\eta_p^2 = .11$. As predicted, accuracy for script-conforming items was not affected significantly by the working memory load manipulation, $F(1, 77) = 1.95$, $p = .167$, $\eta_p^2 = .02$. However, accuracy for script-neutral, $F(1, 77) = 11.83$, $p = .001$, $\eta_p^2 = .13$, and for script-nonconforming items, $F(1, 77) = 49.41$, $p < .001$, $\eta_p^2 = .39$, decreased under working memory load by 6.39% and 13.00%, respectively. Also as predicted, the decrease in accuracy was much larger for script-nonconforming than for script-neutral items. This result pattern was in line with predictions based on MDM and suggests that when the stimulus material was encoded under working memory load, the proportion of familiarity-based retrievals increased. This resulted in more incorrect responses only for script-neutral and script-nonconforming items, but not for script-conforming items.

(b) Confidence: MDM predicted levels of confidence to be comparable for script-conforming and script-nonconforming items, because these items referred to the same features in the memory traces. As explained above, confidence for script-conforming and script-nonconforming items was expected to exceed the level of chance, because one response was expected to be preferred over the other. In contrast, confidence for familiarity-based responses to script-neutral items was expected to approach the level of chance because script-neutral details generally receive less attention and witnesses therefore commonly fail to encode script-neutral details. Consequently, existing memory traces were expected to provide insufficient information to prefer one answer option of script-neutral items over the other. Consequently, level of confidence for script-neutral items was predicted to be lower than

confidence for script-conforming and script-nonconforming items. Again, descriptive results can be inspected in Figure 3. A 2 x 3 mixed factorial ANOVA was computed on confidence. Both the main effects of working memory load, $F(1, 77) = 22.74$, $p < .001$, $\eta_p^2 = .23$, and of script conformity, $F(2, 154) = 24.87$, $p < .001$, $\eta_p^2 = .24$, were significant. Regarding script conformity, specified contrast effects corroborated the hypotheses: a significant difference emerged only between script-conforming and script-nonconforming items on the one hand, and script-neutral items on the other hand, $F(1, 77) = 53.80$, $p < .001$, $\eta_p^2 = .41$. The difference between script-conforming and script-nonconforming items was not significant, $F(1, 77) < 1$. As predicted by the model, no significant interaction between working memory load and script conformity emerged, $F(2, 154) = 1.15$, $p = .321$, $\eta_p^2 = .02$.

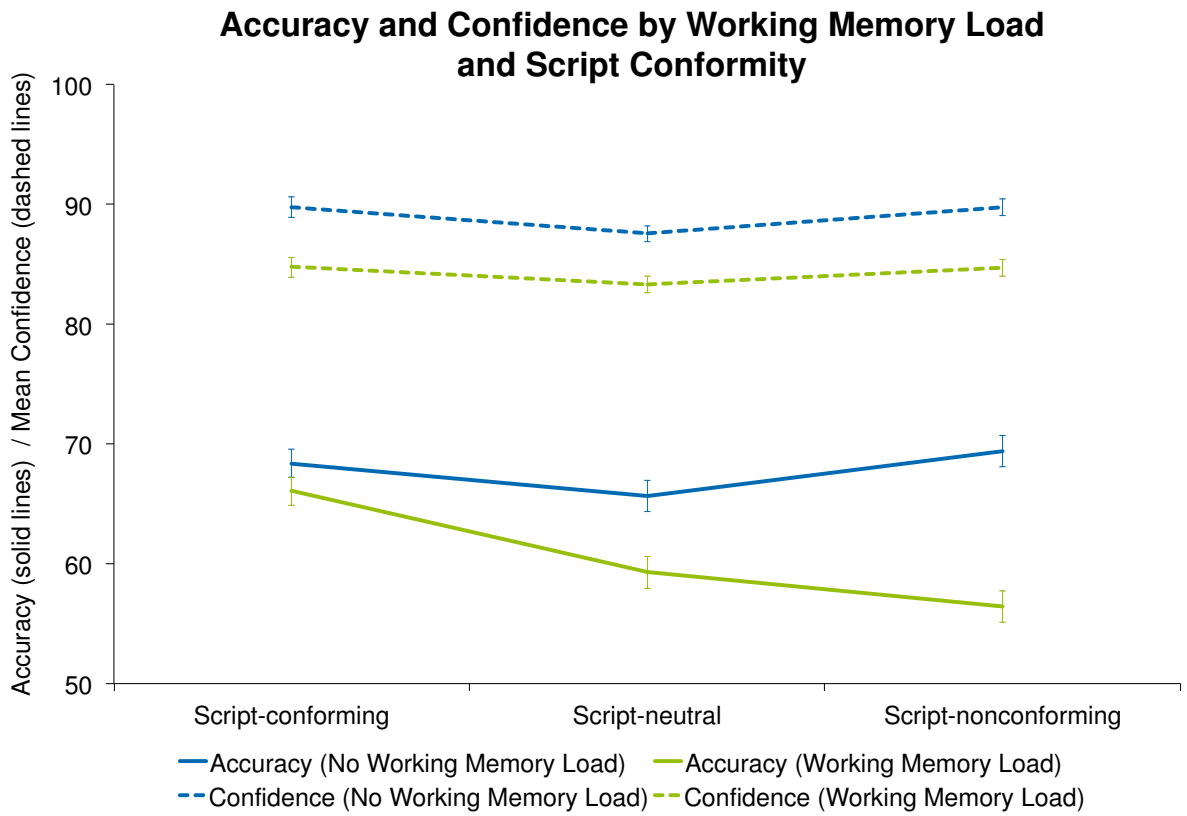


Figure 3. Accuracy and mean confidence (and their standard errors) for script-conforming, script-neutral, and script-nonconforming items (within-subjects) by working memory load (no working memory load vs. working memory load; between-subjects).

(c) Overconfidence: Participants were generally expected to overrate their performance, because according to MDM, error variance in memory traces leads to overconfidence and error variance was to be expected for eyewitnesses' memory traces of observed crime events (for details see Appendix B). This is because crimes were assumed to be highly variable in their event details, and because often witnesses cannot pay full attention to the crimes they observe. In line with this prediction, all witnesses overrated their performance (M = 22.44, SD = 6.19, on a scale from 0 to 50). For script-conforming items, overconfidence was expected to be lower in the working memory load condition compared to the no working memory load condition, because accuracy was predicted to be unaffected by working memory load and confidence was predicted to decrease under working memory load due to an increase in familiarity-based retrievals that are associated with lower confidence levels than direct retrievals. For script-neutral items, overconfidence was predicted to be the same in both conditions, because both accuracy and confidence were predicted to be reduced under working memory load. For script-nonconforming items, overconfidence was expected to be higher under working memory load, because accuracy was predicted to decrease relatively more than confidence under working memory because the proportion of familiarity-based retrievals was expected to increase and accuracy for familiarity-based retrievals was expected to be below the level of chance for these items, whereas confidence for items that relate to scripts was predicted to be always above chance. Descriptive results can be inspected in Figure 4. A 2 x 3 mixed factorial ANOVA was computed on overconfidence. Again, only the results pertaining to the predictions outlined above are presented, the full analysis can be found in Appendix B. In line with predictions, the interaction between the working memory load and the script conformity manipulations was statistically significant, $F(2, 154) = 9.77$, $p < .001$, $\eta_p^2 = .11$. A small decrease in overconfidence under working memory load was observed for script-conforming items, which is in line with predictions. The effect did however not reach statistical significance, $F(1, 77) = 1.86$, $p = .176$, $\eta_p^2 = .02$. MDM did not

predict working memory load to impact overconfidence in script-neutral items and indeed no significant difference in overconfidence occurred for script-neutral items, $F(1, 77) = 1.17$, $p = .282$, $\eta_p^2 = .02$. Also in line with predictions, a large increase in overconfidence was observed for script-nonconforming items under working memory load, $F(1, 77) = 16.91$, $p < .001$, $\eta_p^2 = .18$. As MDM predicted, when participants made more familiarity-based retrievals, the decrease in accuracy was larger than the decrease in confidence only for script-nonconforming items.

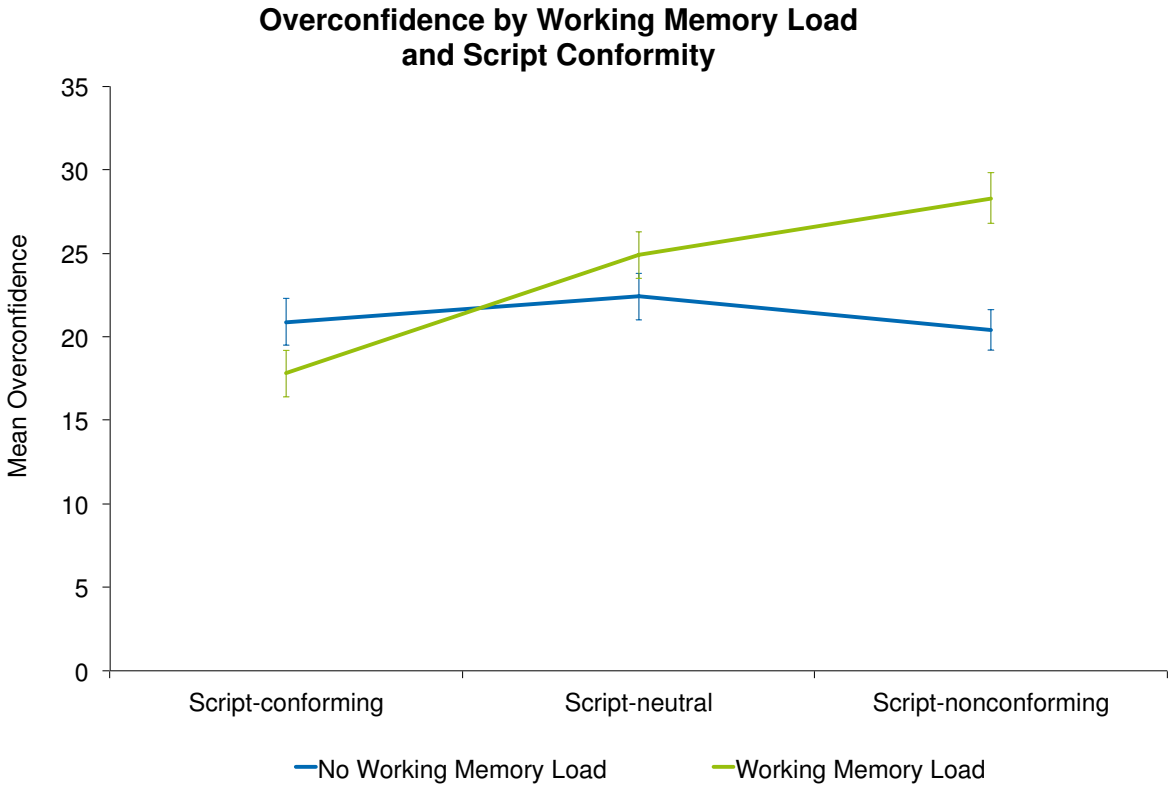


Figure 4. Mean overconfidence (and its standard error) for script-conforming, script-neutral, and script-nonconforming items (within-subjects) by working memory load (no working memory load vs. working memory load; between-subjects).

As predicted, overconfidence for script-conforming items decreased under working memory load, the effect was however small and did not reach statistical significance. The small size of this effect may have been caused by the fact that accuracy for script-conforming

items did not remain completely stable under working memory load, but rather was slightly reduced. For this reason, the predicted decrease in confidence for script-conforming items under working memory load was not sufficient to produce a significant decrease in overconfidence. Overall, the observed result patterns were almost entirely in line with MDM's predictions about the impact of working memory load and script conformity on accuracy, confidence, and overconfidence.

Overconfidence in Study 2 exceeded the mean effect of overconfidence found in the meta-analysis in Study 1. There are two potential explanations for this finding. First, as Study 1 showed, stimulus material and item type impacted overconfidence. The crime simulation and the items employed in Study 2 had never been used in a study investigating overconfidence before. It therefore cannot be precluded that the stimulus material used in Study 2 generally leads to higher overconfidence for event memory than other stimulus materials. A second explanation might be that the proportion of script-nonconforming items was artificially increased in this study to be able to investigate the effect of script conformity systematically. Because script-nonconforming items can generally be expected to be less frequent than script-conforming items and because they are associated with particularly high levels of overconfidence, overconfidence can be expected to be lower in natural settings. But also for script-conforming items, for which familiarity-based responding was predicted to result in correct responses, substantial levels of overconfidence were obtained in the present study. MDM's error variance account for overconfidence (Dougherty, 2001; Erev, Wallsten, & Budescu, 1994) can explain this finding as the result of a large amount of variability in previously observed crimes or as the result of a high working memory load during previous crime observations. More implications and limitations of this study and the application of MDM to witnesses' memory for events are discussed in Chapter 4 and Appendix B.

Study 2 provided support for the applicability of the exemplar-based MDM to eyewitnesses' memory for events. Employing a theory-based approach to investigate

cognitive processes underlying the confidence-accuracy relationship in eyewitnesses' memory for events, Study 2 complied with requests put forward by multiple researchers to shift attention from *what* factors influence witnesses' memory to *how* these factors work (Ogloff, 2000). Moreover, applying a model to the confidence-accuracy relationship allows generalizing findings to other samples and stimulus materials (Clark, 2008). Using MDM as a model of witness memory and confidence allows deriving predictions about factors that may influence the proportion of direct retrievals or the perceived familiarity associated with response options. Thus, predictions about the possible impact of other, yet unstudied system variables or estimator variables on witness memory that may be of interest to researchers or practitioners (cf. Chapter 1) can be derived. Suggestions for applications of MDM to other phenomena observed in research on eyewitnesses' memory for events are also made in Appendix B.

2.5. Discussion of the Confidence-Accuracy Relationship

In this chapter, the confidence-accuracy relationship, in particular, the realism of confidence in eyewitnesses' testimony of events was discussed and two studies were presented. A meta-analysis (Study 1, Chapter 2.2) revealed that witnesses are rather well-calibrated, but tend to be overconfident when making recognition judgments about the crimes they observed. Resolution of confidence judgments was found to be very poor. These findings however need to be treated with caution, since the meta-analysis revealed a rather limited diversity in study method (stimulus material and item format). Lack of diversity in stimulus materials in eyewitness studies has previously been criticized (Memon et al., 2008). The small number of studies that qualified for inclusion in the meta-analysis suggests that either eyewitnesses' memory for events is a neglected area of research in forensic psychology (Sporer, 1996), or that researchers are reluctant to employ calibration analysis in studies on eyewitness event memory despite its obvious advantages over correlation analysis (Brewer,

2006; Juslin et al., 1996), or both. The meta-analysis, thus, points to a need for more studies investigating the realism of confidence in eyewitness event memory.

Because the first quantitative synthesis of research findings on the realism of confidence in witnesses' memory for events found that witnesses are commonly overconfident in the accuracy of their testimonies (cf. also Olsson & Juslin, 2002), a theory-based approach was adopted in Study 2 to shed light on the cognitive processes contributing to overconfidence in eyewitness event memory. The exemplar-based memory model MDM was proposed to derive predictions about how scripts influence accuracy, confidence, and overconfidence in recognition judgments. In a crime simulation study, MDM proved to be a viable model of cognitive processes underlying the confidence-accuracy relationship – in particular, the overconfidence in eyewitnesses' memory for events. The results provide an explanation for the common overconfidence in eyewitnesses' memory for events (reliance on indirect, familiarity-based retrieval) and showed specifically that overconfidence is maximized for atypical crimes when cognitive capacities are limited.

These two studies fill major gaps in the literature on eyewitness event memory. First, rather few studies have examined the realism of confidence in witnesses' testimony about events (as opposed to identifying suspects from line-ups) and the existing studies employ a large variety of experimental manipulations, making it difficult to draw firm conclusions about the actual magnitude of confidence calibration, under-/overconfidence, and resolution as measures of the realism of confidence. The present meta-analysis presents estimates of this magnitude and calls for more studies with diverse study methods (stimulus materials and response formats). Second, suggesting MDM as a model of eyewitnesses' overconfidence enables generalization of research findings via theoretical assumptions. Furthermore, predictions can be derived about the impact of yet unstudied factors on the realism of confidence in eyewitnesses' memory for events. Proposing MDM as a model of overconfidence in witnesses' memory for events is attractive because MDM can easily be

formalized. Formal models are particularly useful research tools because they are highly precise and are able to capture complex relationships (Bjork, 1973). A formal model requires every step to be explicated and to be formally defined. Any vagueness in concepts that may occur in verbally formulated models is therefore avoided (Hintzman, 1991).

However, using confidence to determine the accuracy of witnesses' testimonies has a major disadvantage: measures of the realism of confidence can only be computed if confidence ratings were collected. Moreover, confidence ratings can be biased by situational circumstances, for example, when other witnesses are present and confidence ratings have to be given in public (Shaw, Appio, Zerr, & Pontoski, 2007). Most crimes indeed feature several witnesses. As an alternative to using confidence ratings, the accuracy of witnesses' testimonies may therefore be assessed and crimes may be reconstructed by aggregating multiple witness reports. This approach is detailed in Chapter 3.

3. Assessing Witness Memory by Aggregating Multiple Testimonies

In this chapter, reasons for aggregating multiple witness testimonies are discussed and two aggregation methods are introduced. An empirical study investigating the performance of these aggregation methods in reconstructing crimes, and their dependence on the heterogeneity in witnesses' competence levels (i.e., in the probabilities that witnesses provide correct responses) is presented (3.1). The chapter concludes with a discussion of the benefits and limitations of aggregating witness testimonies (3.2).

Rarely are crimes observed by a single witness. Of 773 students at an Australian university, 75% had previously witnessed a serious criminal event (e.g., physical assault, property vandalism, homicide; Paterson & Kemp, 2006). Of these witnesses, 86% reported that at least one other witness was present at the crime scene. The average number of co-witnesses was 6.77, and the median was 3. Over one third of the witnesses reported more than five co-witnesses, and 14% reported having over ten and up to 100 co-witnesses. Similarly, of 60 people who had witnessed a crime (e.g., violence against a person, robberies, or burglaries) and who were interviewed by the police, 87% reported that at least one other witness was present at the crime scene (Skagerberg & Wright, 2008). In this study, the mean number of co-witnesses was 4.02, and the median was 2.5.

When multiple witnesses observe the same crime, discrepancies between their testimonies are to be expected (Loftus, 1996). However, processes to identify correct from incorrect responses are largely lacking (Bernstein & Loftus, 2009; Sarwar, Sikström, M., & Innes-Ker, 2015). One promising approach to reconstructing a crime from witness reports might be to aggregate the individual testimonies. Aggregating multiple responses has been found to increase the validity of responses in other research domains and has therefore been termed the *wisdom of the crowd* (e.g., Armstrong, 2004; Clemen, 1989; Galton, 1907). Research on the benefits and limitations of aggregation in eyewitness memory has however been scarce. Only three studies have systematically investigated the validity of aggregated

witness reports. Two of these studies have found that aggregated eyewitness identifications were superior to individual eyewitness identifications. This was true even for small groups of only 3-4 witnesses and when applying the intuitive and simple aggregation rule of choosing the modal response (Clark & Wells, 2008; Sanders & Warnick, 1982). The third study investigated aggregation of witnesses' recollections of crime events and found aggregated reconstructions to be more valid when an aggregation rule was employed that could account for individual differences between witnesses and items compared to when an aggregation rule was employed that ignored such differences (Waubert de Puiseau et al., 2012). This study however did not compare aggregated with single responses and, thus, did not directly assess wisdom of the crowd.

The cited research findings suggest that aggregation may be a viable alternative to assessing confidence judgments when crimes have to be reconstructed from witness testimonies. However, there are different ways to aggregate witness reports (cf. Waubert de Puiseau et al., 2012) and it is unknown, which of these aggregation rules is best suited for crime reconstructions. Two of the existing studies that aggregated witness testimonies (Clark & Wells, 2008; Sanders & Warnick, 1982) employed the simple Majority Rule. The Majority Rule provides an unweighted aggregation of testimonies across witnesses. For binary events (e.g., true/false items), the majority is defined in terms of the modal response across all witnesses (the outcome of aggregation based on the Majority Rule is henceforth referred to as *majority reconstruction*). Using the Majority Rule to aggregate testimonies has been found to produce more accurate eyewitness identifications than individual decisions, regardless of the size of the group of witnesses whose testimonies were aggregated (≥ 2; Clark & Wells, 2008; Sanders & Warnick, 1982). The Majority Rule is simple and robust, but it has three important limitations: (a) the Majority Rule ignores whether a majority was strong (close to 100%) or weak (close to 50%), (b) in the case of a tie, the majority is not defined, and (c) the Majority Rule does not weight responses by the competences of the individuals. Witness competence is

defined as a witness's probability to provide an accurate report about the observed crime. Thus, when the majority is incompetent, the majority reconstruction can be incorrect.

Assuming that more competent witnesses enable more accurate crime reconstructions, it would be desirable to be able to weight individual responses by the witnesses' competences when aggregating their testimonies. Competences are however difficult to determine when the truth is unknown. Cultural Consensus Theory (CCT; Romney, Weller, & Batchelder, 1986) provides a method for objectively computing competences when knowledge about the correct responses is not available. CCT was originally suggested to define unknown cultures based on reports of members of these cultures. As Waubert de Puiseau et al. (2012) proposed, there are a number of similarities between an anthropologist trying to understand unknown cultures and a legal expert trying to reconstruct a crime: (a) commonly, neither members of the culture under investigation nor witnesses are in perfect agreement with each other; (b) both members of the culture under investigation and witnesses differ in competences; (c) it is unknown, which responses are correct.

For true/false items, CCT can be formalized as a General Condorcet Model (GCM; Batchelder & Romney, 1986; Karabatsos & Batchelder, 2003; Romney et al., 1986). The model is outlined in detail in Appendix C. The GCM in its most complex form (Karabatsos & Batchelder, 2003; Oravecz, Vandekerckhove, & Batchelder, 2014) provides estimates of witnesses' competences, their tendencies to guess "true" when competence is lacking (henceforth referred to as 'guessing bias'; cf. Two-High Threshold model, 2-HTM; Snodgrass & Corwin, 1988), item difficulties, and the answer key (i.e., the correct responses to the items). The GMC can be seen as extending the 2-HTM by adding item difficulty and the answer key as latent parameters. The applicability of the GCM in its most complex form is limited by two assumptions: (a) there is a common truth underlying all witness reports and the answer key is therefore constant across all witnesses (note that the Majority Rule also makes this assumption), and (b) responses are locally independent, that is, the responses of

individual witnesses are independent of each other, both across items and across witnesses (Romney, 1999). In contrast to the Majority Rule, the GCM is rather difficult to implement and may require extensive computing facilities. Moreover, the GCM is restricted to true/false items (models for other item types exist, it is however not possible to mix response formats). Paralleling the terminology used for the Majority Rule, the crime reconstruction based on the GCM is hereafter referred to as *consensus reconstruction*. Waubert de Puiseau et al. (2012) showed that consensus reconstructions were more accurate than majority reconstructions, regardless of the size of the group of witnesses, whose testimonies were aggregated.

When investigating the performance of aggregation methods, it is important to consider the heterogeneity of knowledge in individuals, whose statements are aggregated, because the superiority of aggregation over individual responses has been found to increase with variability in knowledge (e.g., Davis-Stober et al., 2014). Heterogeneity in competences seems to be particularly important when researching eyewitness memory because heterogeneity in witnesses' competences is expected to be high in real-world samples of witnesses. Most eyewitness studies are however conducted in laboratories under highly standardized conditions and employ student samples. Such studies are therefore suspected to underestimate true heterogeneity in witnesses' competences (Lindsay et al., 2000; Lindsay et al., 1998; Wells et al., 2006). If the superiority of aggregated over individual testimonies increases with heterogeneity in witnesses' competence levels, standard laboratory studies may underestimate the benefit of aggregation. Moreover, it seems likely that competence heterogeneity may benefit some aggregation rules more than other aggregation rules. More precisely, aggregation rules that can account for differences in competences may benefit more from competence heterogeneity than simple rules that employ unweighted aggregation. Empirical studies investigating the impact of competence heterogeneity on the superiority of different aggregation rules over individual witness statements are however lacking.

3.1. Study 3: On the Importance of Considering Heterogeneity in Witnesses' Competence Levels When Reconstructing Crimes from Multiple Witness Testimonies

The aim of Study 3 (Appendix C) was to investigate the impact of heterogeneity in witnesses' competence levels on the validity of crime reconstructions based on aggregation using either the simple Majority Rule or the GCM that takes differences in witnesses' competences into account. Existing studies commonly define heterogeneity in competences in terms of inequality of a group's members' competences. Heterogeneity in competences can therefore be measured in terms of the variance in competences. The validity of an aggregated crime reconstruction is determined by comparing the outcome of the aggregation to an a-priori known answer key. The existing literature provides different answers to the question how heterogeneity in witnesses' competence levels impacts the outcome of simple aggregations based on the Majority Rule. As detailed in the following, studies predict competence heterogeneity to either have no impact on, or to improve the validity of majority reconstructions. Grofman, Owen, and Feld (1983) postulated that the majority reconstruction is not affected by competence heterogeneity if three conditions are met: (a) mean competence is above the level of chance (i.e., .5 given two answer options), (b) heterogeneity does not affect mean competence, and (c) competences are normally distributed around the mean (cf. also Kazmann, 1973). Kanazawa (1998) however formally showed that heterogeneous (compared to homogeneous) groups are more likely to select the correct response to a binary question under the condition that mean individual competences are larger than $(1/2) + (1/2n)$, where n is the number of individuals in the group (cf. also Boland, 1989).

Only one study has investigated the impact of competence heterogeneity on the performance of the GCM. Using a computer simulation, Weller (1987) found that consensus reconstructions in the heterogeneous group were equally accurate as consensus reconstructions in the homogeneous group. This finding has however not yet been replicated

with human participants. Moreover, the study employed a restricted variant of the GCM that only accounted for variance in competence and assumed homogeneous guessing biases and item difficulties. It seems likely that employing a variant of the GCM that is more flexible, consensus reconstructions will outperform majority reconstructions when levels of competence are heterogeneous.

Drawing upon the models and the existing literature, two predictions were derived. First, consensus reconstructions were predicted to be more accurate than majority reconstructions when competences were heterogeneous. Second, and by contrast, in groups of witnesses with equal competences, weighting responses by the witnesses' competence should not affect the quality of crime reconstructions and the validity of majority and consensus reconstructions should therefore not differ.

To test these predictions, heterogeneity in witnesses' levels of competence was experimentally manipulated. One hundred twenty-seven participants (of whom 6 were excluded because they failed to follow the instructions) viewed a crime simulation showing a bank robbery and subsequently completed 128 true/false items about the crime event. Participants were randomly assigned to either the homogeneous competences or the heterogeneous competences condition. To induce heterogeneity, three experimental factors (reduction of information extractable from video; increase of working memory load; distortion of retrieval) with three levels each were manipulated to impair witnesses' competence resulting in 27 experimental cells (Design and Procedure are outlined in detail in Appendix C). Participants in the heterogeneous condition were assigned randomly to one of the 27 cells. By contrast, participants in the homogeneous condition were all assigned to the same cell that constituted the middle level on all three factors that were manipulated in the heterogeneous condition. This was done to generate two groups with different variances while having the same mean.

The analyses were performed using the R statistics software (R Development Core Team, 2015). Details on the configuration of the estimation process are presented in Appendix C. Competences of witnesses measured in terms of proportions of correctly answered items were significantly more heterogeneous in the heterogeneous condition (SD = 7%) compared to the homogeneous condition (SD = 5%), $F(53, 66) = 1.68$, $p = .026$. The mean proportion of correct responses was significantly lower in the homogeneous condition (M = 61%) than in the heterogeneous condition (M = 64%), $t(118.80) = 2.38$, $p = .019$, $d = 0.49$. Because mean competence has been found to influence aggregation outcomes (Grofman et al., 1983; Weller, 1987), a random sampling approach was performed. The procedure employed for the random sampling approach is detailed in Appendix C. A brief summary of how the samples were drawn and of the core results of computations based on the randomly drawn samples are presented below.

Sixty-seven sample pairs of either 10, 20, of 40 witnesses in each sample were drawn by randomly choosing participants from the homogeneous and heterogeneous conditions, respectively. To avoid competence variance and mean competence being confounded and to guarantee a fair comparison, the sampling algorithm was forced to draw sample pairs that were matched in competence (means were allowed to differ only in the third decimal place when competence was measured on a scale ranging from 0, indicating no competence, to 1, indicating maximum competence). When doing so, the sampling algorithm assured that in each pair, the variance of competences in the heterogeneous sample significantly exceeded the variance of competences in the corresponding homogeneous sample according to an F-test. T-tests across samples confirmed that this manipulation was successful for all sample sizes (all ts $\geq$ 10.72, all ps $<$ .001). The sampling algorithm ensured that no sample pair was drawn more than once. By comparing the generated sampling distributions, differences in the validity between the majority reconstructions and the consensus reconstructions could statistically be tested. Moreover, the link between differences in validities of crime

reconstructions and heterogeneity of witnesses' competence levels could be investigated. Moreover, the impact of sample size on aggregation performance could be inspected because samples of different sizes were drawn. The validity of aggregation outcomes was predicted to increase with sample size for both the majority and the consensus reconstructions (Batchelder & Romney, 1988; Grofman et al., 1983; Kazmann, 1973; Romney et al., 1986; Waubert de Puiseau et al., 2012; Weller, 1987, 2007).

The majority and consensus reconstructions were compared to the a-priori known answer key and proportions of correctly estimated responses were computed to assess the validity of estimated answer keys. To determine individual crime reconstructions, the mean proportion of correct responses was computed. As can be seen in Figure 5, both aggregate reconstructions always outperformed individual reconstructions. The consensus reconstruction clearly benefited from the heterogeneity in competence levels; for all sample sizes, the validity of consensus reconstruction increased by about five percent when competences were heterogeneous compared to homogeneous. For majority reconstructions, a very small increase in validity could be observed when competences were heterogeneous compared to when they were homogeneous. This increase was much smaller and rather negligible.

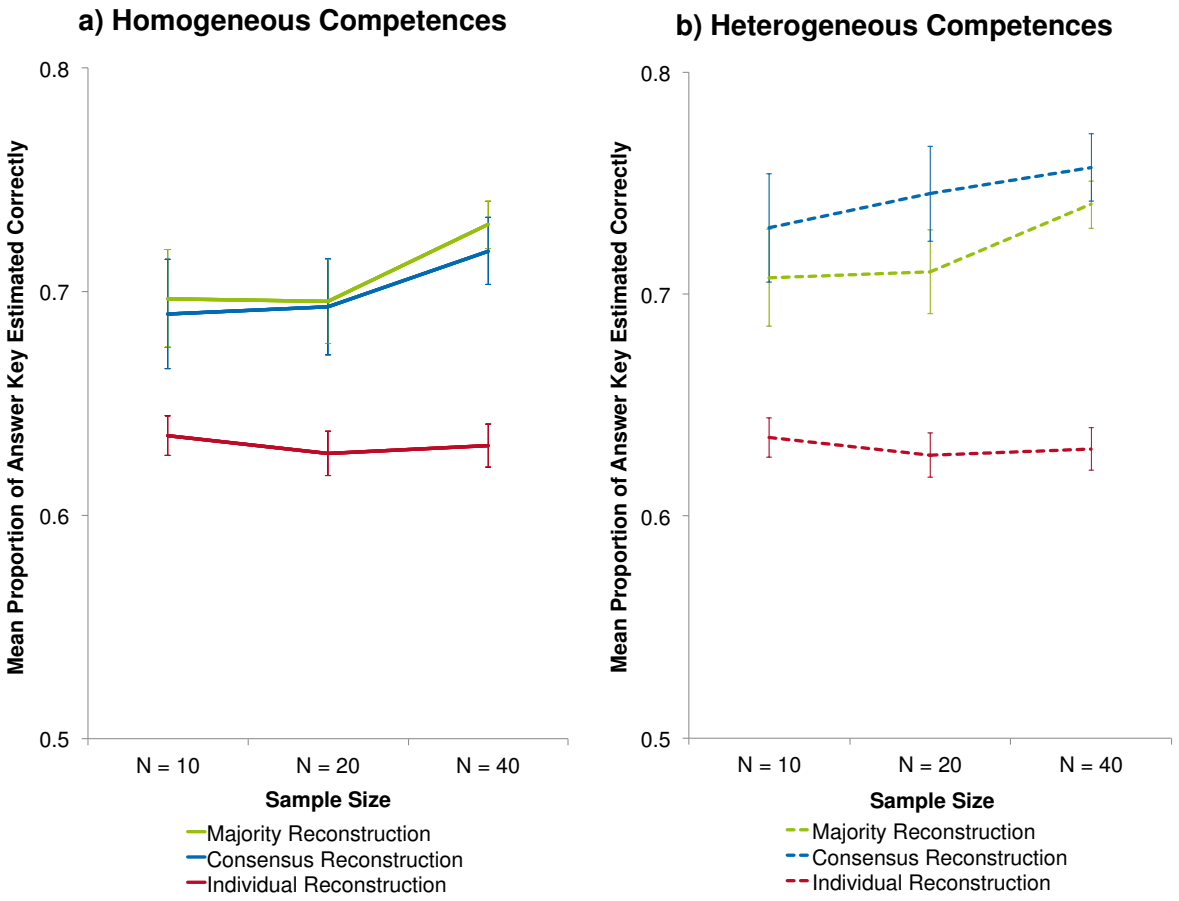


Figure 5. Mean proportions (and their standard errors) of agreement between the true answer key on the one hand and the answer key estimates that were based on the Majority Rule (majority reconstruction, green lines), the GCM (consensus reconstruction, blue lines), and the individual responses (individual reconstruction, red lines) on the other hand as a function of competence heterogeneity (a: homogeneous competences; b: heterogeneous competences) and the number of witnesses (n = 10, 20, and 40).

To assess whether the superiority of consensus over majority reconstructions was significantly associated with a sample being homogeneous or heterogeneous in competences, majority reconstructions and consensus reconstructions were compared within each pair of homogeneous and heterogeneous samples. Three McNemar tests were computed separately for each defined sample size (n = 10, 20, 40). Significant associations could be observed for all sample sizes (Tables 1a-c). The odds ratio that compared the sample pairs in which the consensus reconstruction was more accurate in only the homogeneous or in only the heterogeneous sample (i.e., in Tables 1 a-c, the lower cells in the left column and the upper cells in the right column), increased with sample size. This was due to the number of sample

pairs, in which the GCM outperformed the Majority Rule only in the heterogeneous, but not in the homogeneous condition.

Table 1
2x2 Contingency Tables and the Results of the McNemar Tests Comparing the Majority with the Consensus Reconstructions for the Homogeneous versus Heterogeneous Samples Separately for Different Sample Sizes (a: 10, b: 20, and c: 40)

a) $n = 10$		Homogeneous samples		Total
		GCM > MR	GCM ≤ MR	
Heterogeneous samples	GCM > MR	17	32	49
	GCM ≤ MR	4	14	28
Total		21	46	67
Odds ratio			8	
Proportion of discordant pairs			54%	
McNemar test			$\chi^2(1, N = 67) = 46.00, p < .001$	

b) $n = 20$		Homogeneous samples		Total
		GCM > MR	GCM ≤ MR	
Heterogeneous samples	GCM > MR	24	39	63
	GCM ≤ MR	2	2	4
Total		26	41	67
Odds ratio			19.5	
Proportion of discordant pairs			61%	
McNemar test			$\chi^2(1, N = 67) = 31.61, p < .001$	

c) $n = 40$		Homogeneous samples		Total
		GCM > MR	GCM ≤ MR	
Heterogeneous samples	GCM > MR	9	48	57
	GCM ≤ MR	0	10	10
Total		9	58	67
Odds ratio			NA[a)]	
Proportion of discordant pairs			72%	
McNemar test			$\chi^2(1, N = 67) = 46.02, p < .001$	

Note. MR = Majority Rule, GCM = General Condorcet Model; [a)]no odds ratio could be computed for $n = 40$ because division by 0 is not defined.

In sum, aggregated crime reconstructions were always superior to individual crime reconstructions. Competence heterogeneity benefited the consensus reconstruction more than the majority reconstruction. Moreover, when samples were homogeneous in competences, a trend was observed for consensus reconstructions to be poorer than majority reconstructions (see Figure 5), however, this trend was not substantial in the present study. In line with predictions, the pattern of results suggests that using the GCM to reconstruct crimes was

justified when individual witnesses' varied in competences. In this case, consensus reconstructions were more accurate than majority reconstructions due to oversimplification of the Majority Rule and because the GCM can take individual differences between witnesses and items into account. The most accurate crime reconstructions were observed when competences were heterogeneous and witness reports were aggregated using the GCM.

The validity of aggregated crime reconstructions increased with sample size. However, the smallest sample size investigated in Study 3 (n = 10) exceeded the median number of four witnesses commonly present at a crime scene (Paterson & Kemp, 2006; Skagerberg & Wright, 2008). This was because it was not possible to draw samples of less than ten participants that complied with the restrictions implemented in the algorithm drawing the random samples (equal means and significantly different standard deviations) in the present study. Other studies have previously shown that also for small samples of three or four witnesses, aggregated responses were more accurate that individual responses (Clark & Wells, 2008; Sanders & Warnick, 1982), and that consensus reconstructions were more accurate than majority reconstructions (Waubert de Puiseau et al., 2012). It thus seems likely that the pattern of results obtained for samples of 10, 20, or 40 witnesses would generalize to smaller sample sizes.

3.2. Discussion of the Aggregation Approach

Study 3 demonstrated the validity of aggregating statements of multiple witnesses and, thus, confirmed findings of previous studies (Clark & Wells, 2008; Sanders & Warnick, 1982; Waubert de Puiseau et al., 2012). Aggregation is beneficial because, in contrast to other approaches to assessing the accuracy of witness testimony (e.g., through confidence ratings), witnesses do not need to provide additional information. Moreover, because most crimes feature multiple witnesses (Paterson & Kemp, 2006; Skagerberg & Wright, 2008), aggregation can almost universally be employed. Another benefit of aggregation is that it

makes differences in individual witnesses' competences and discrepancies between their reports more visible. This is important and may help reduce judicial errors because it may help legal actors to develop a more realistic perception of witnesses' competences and may thus lead to witness testimony being given less weight in legal procedures (cf. Chapter 1).

However, several limitations of the aggregation approach need to be mentioned. First, if all individuals answer randomly, aggregating their answers does not result in more reliable crime reconstructions (Clark & Wells, 2008; Sanders & Warnick, 1982). It should be noted that this limitation is not restricted to reconstructions based on aggregated reports, but also applies to crime reconstructions based on single testimonies. As Study 3 suggests, even when witnesses' mean performance is only slightly better than the level of chance, aggregation may benefit the legal fact finding process. However, when inter-witness agreement does not result solely from shared knowledge about the crime under investigation, outcomes of aggregation may be seriously distorted (Clark & Wells, 2008; Waubert de Puiseau et al., 2012; Wells et al., 2006). Examples of such influences include scripts that individuals hold of crimes (Greenberg et al., 1998; Holst & Pezdek, 1992), wrongful information obtained through discussions with fellow witnesses (Gabbert, Memon, & Allan, 2003; Meade & Roediger, 2002; Shaw, Garven, & Wood, 1997), and leading questions posed by interviewers (Loftus, 1975a; Sharman & Powell, 2012). Aggregation methods ignoring witness and item characteristics such as the Majority Rule have been proven to be fairly robust against the interdependence of individual responses (Davis-Stober et al., 2014; Estlund, 1994; Ladha, 1992). Because the GCM is, however, based on the assumption that responses are locally independent, consensus reconstructions may be more vulnerable to such systematic distortions (Waubert de Puiseau et al., 2012).

4. General Discussion

The studies presented in this doctoral thesis investigated two different approaches to assessing the accuracy of witnesses' testimony. On the one hand, the realism of confidence ratings in eyewitnesses' recognition memory for events was investigated in a meta-analysis (Study 1) and in a model-based analysis of the link between overconfidence in witnesses' memory for events and scripts (Study 2). On the other hand, the superiority of two aggregation rules over individual testimonies was investigated when witnesses' levels of competences were either homogeneous or heterogeneous (Study 3).

Study 1 provided the first meta-analysis of the realism of confidence in eyewitnesses' memory for events. Confidence ratings were found to be a reliable indicator of witnesses' testimony's accuracy (good calibration), but a general tendency was observed for witnesses to be overconfident in the accuracy of their recollections. Confidence resolution was rather low, indicating that witnesses were less capable of distinguishing between their correct and their incorrect responses. The pattern of results was thus at odds with results from correlation analyses that have found low between-subjects correlations and higher and more stable within-subjects correlations (e.g., Robinson & Johnson, 1996). The divergence between calibration and between-subjects correlations parallels findings for eyewitness identification decisions (Olsson & Juslin, 2002). More research is needed to understand the divergence of resolution and within-subjects correlation coefficients. The findings of the meta-analysis highlight the importance of considering calibration analysis when investigating the relation between confidence and accuracy.

The meta-analysis further points to the necessity to consider diversity in study method: six out of eight studies included in the meta-analysis employed the same stimulus material and item type. A moderator analysis revealed that study method significantly affected confidence calibration, under-/overconfidence, and resolution. The results of the meta-analysis, thus, need to be treated with caution and cannot readily be generalized to other

stimulus materials and item types. Particularly in applied research settings like forensic psychology, stimulus materials need sufficient ecological validity to enable generalization across persons and conditions (Memon et al., 2008).

To generalize findings from experimental studies, a sound theoretical basis is helpful. For this reason, the memory model MINERVA-Decision-Making (MDM; Dougherty, 2001; Dougherty et al., 1999) was employed in Study 2 to investigate overconfidence in witnesses' memory for events. Three main conclusions could be drawn from the results: (a) Responses and confidence ratings are based on the same cognitive processes; (b) overconfidence should be treated as a ubiquitous phenomenon in eyewitnesses' memories of events; and (c) overconfidence is expected to increase for details of observed crimes that are in conflict with scripts when witnesses cannot pay sufficient attention to the crime. The pattern of results was in line with predictions based on MDM. Study 2 therefore contributes to a general understanding of cognitive processes underlying witness reports and confidence ratings. Thereby, the model-based approach chosen in Study 2 allows deriving additional testable predictions, for example about how misinformation or interviewing delay impacts accuracy, confidence, and overconfidence (see Appendix B). The model-based approach employed in Study 2 is therefore particularly fruitful (Clark, 2008).

Witnesses do not always provide confidence ratings about the assumed correctness of their recollections of observed crime events. As an alternative, when multiple witnesses observe the same crime, aggregating their reports may inform crime reconstructions. The validity of aggregation outcomes has been hypothesized to vary as a function of the heterogeneity of competence levels. Competence levels are expected to be highly heterogeneous in real-world witnesses and are therefore expected to be commonly underestimated in laboratory studies employing student samples under standardized conditions (Lindsay et al., 2000; Lindsay et al., 1998). If high competence heterogeneity improves the validity of aggregated crime reconstructions, laboratory studies likely

underestimate the benefits of the aggregation approach. For this reason, Study 3 aimed to investigate the impact of competence heterogeneity on the validity of crime reconstructions based on two different aggregation rules, the simple Majority Rule and the General Condorcet Model (GCM) that is the formalization of Cultural Consensus Theory (CCT). In contrast to the Majority Rule, the GCM can take individual differences in witnesses' competences and guessing biases and in item difficulties into account. Study 3 showed that crime reconstructions based on either aggregation rule were more accurate than crime reconstructions based on individual testimonies. The study thus found empirical support for crowd wisdom to be prevalent also in witnesses' testimonies. When competence levels of witnesses were heterogeneous, crime reconstructions based on the GCM were more accurate than crime reconstructions based on the Majority Rule. In contrast, when competence levels were homogeneous, the benefit of the GCM disappeared and a trend occurred for the Majority Rule to produce more accurate crime reconstructions than the GCM. Competence levels of real-world witnesses are however unlikely to be homogeneous and the superiority of the Majority Rule over the GCM was not reliable. Future studies should investigate other conditions, under which the performance of the GCM may be improved or impaired. Nevertheless, Study 3 confirmed that aggregating witnesses' reports is a viable approach to reconstructing crimes and further showed that given heterogeneity in witnesses' competence levels the GCM should be preferred over the Majority Rule.

The GCM that was employed in Study 3 is an extension of the Two-High-Threshold Model (2-HTM; Snodgrass & Corwin, 1988). The 2-HTM is a memory model, which postulates that knowledge and guessing processes determine recognition judgments (the GCM extends the 2-HTM by adding item difficulty and the answer key as latent parameters; see Chapter 3.1 and Appendix C for more details). Study 2 and Study 3 thus employed different memory models to explain cognitive processes underlying witness memory. MDM and the 2-HTM can however be treated as complementary (Figure 6). More precisely, MDM specifies

the cognitive processes underlying the knowledge process postulated by the 2-HTM: the knowledge process involves either direct retrieval or familiarity-based retrieval, when one response option is perceived to be more familiar than the other. Moreover, the 2-HTM makes an explicit assumption about individuals' responses when responding has to be based on perceived familiarity and when the answer options appear equally familiar, that is, witnesses have to guess.

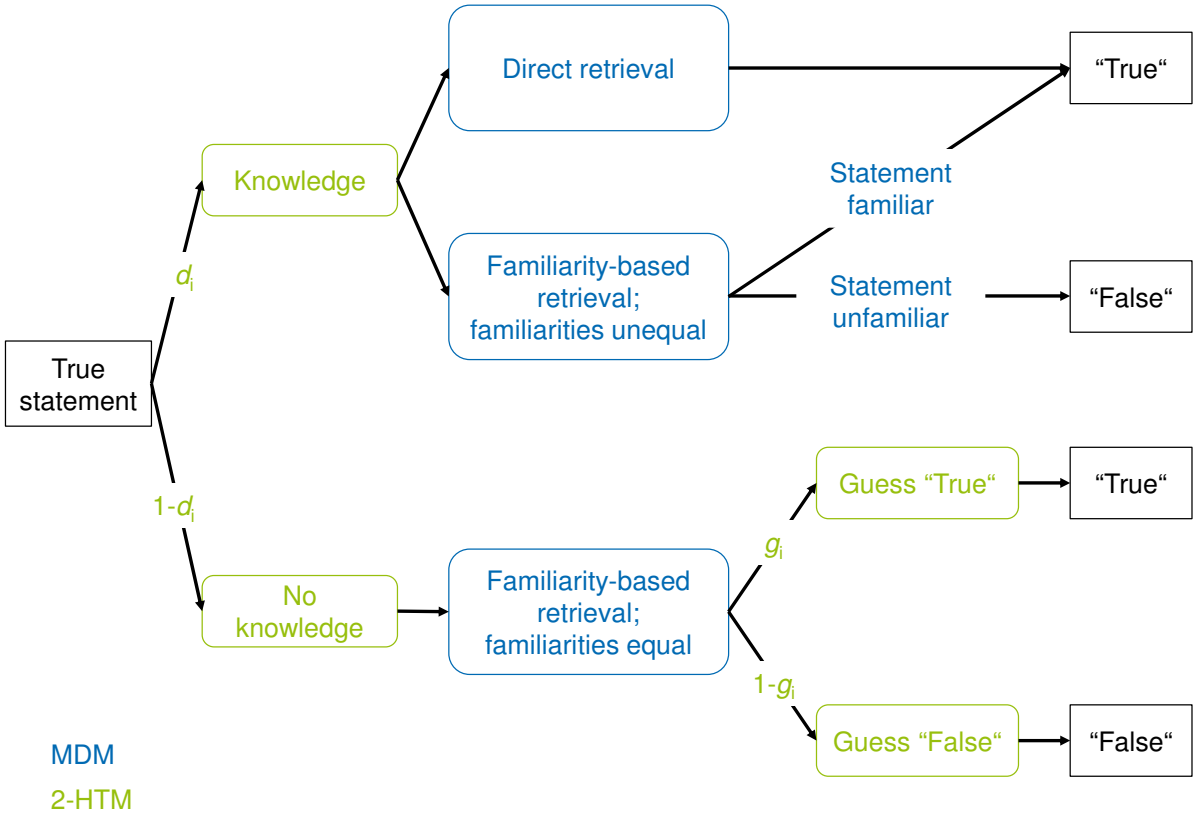


Figure 6. Schematic of MDM (blue) and 2-HTM (green) combined into one memory model for a true statement (for false statements, direct retrieval would result in a "False" response; familiarity-based retrieval with unequal familiarities would result in a "True" response if the statement appeared familiar, and in a "False" response if the statement appeared unfamiliar); d_i denotes a witness's competence, g_i denotes a witness's tendency to guess "true" when the witness does not know the correct response.

4.1. Future Directions

As the meta-analysis in Study 1 revealed, research on eyewitnesses' memory for events suffers from a *lack of methodological diversity*. It is therefore unknown whether the

findings from the meta-analysis would replicate with different stimulus materials and different item types. Moreover, due to the restricted diversity in study method, virtually all factors manipulated in the studies included in the meta-analysis (e.g., timing of interview, presentation of misinformation) were confounded with stimulus materials and item types employed. Future studies should therefore, on the one hand, aim to replicate the findings from the meta-analysis with other crime simulations and question formats, and on the other hand test which factors improve or impair the realism of confidence in witnesses' memory for events.

Mean overconfidence observed in Study 2 (22.44 measured on a scale from 50 to 100) exceeded mean overconfidence observed in the meta-analysis in Study 1 (.09 measured on a scale from 0 to 1). Because in Study 2, mean proportion of correct responses was around 70%, whereas the meta-analysis in Study 1 found mean accuracy to be .65 (on a scale from 0 to 1), this difference cannot be explained in terms of the hard-easy effect (i.e., overconfidence increases with item difficulty; cf. Juslin, 1993; Juslin, 1994). Rather, it seems likely that study material and item selection affected overconfidence. This supports the conclusion drawn from Study 1 that more empirical studies with more diverse stimulus materials are necessary to enable firm and generalizable conclusions about the realism of confidence in general and overconfidence in particular in eyewitnesses' memory for events.

The model-based approach presented in Study 2 provides means to generate research questions and derive *predictions about factors influencing the realism of confidence*. Further empirical support for the applicability of MDM to confidence and accuracy in witnesses' memory for events comes from research on misinformation. The presentation of misinformation subsequent to observing a crime should lead to an increase in overconfidence. According to MDM, this is because witnesses generate a memory trace each time they observe information regarding a crime. Witnesses are therefore expected to generate an additional memory trace for each piece of misinformation they receive. This results in an

increase in the number of relevant memory traces containing features that are incorrect with respect to the crime under investigation. When more similar and relevant memory traces contain details that are in conflict with details of the crime under investigation, the probability of giving an incorrect response due to a false feeling of familiarity is increased. In line with this reasoning, presenting misinformation has been found to decrease accuracy and to increase overconfidence (e.g., Bonham & Gonzalez-Vallejo, 2009; Jack, Zydervelt, & Zajac, 2014). Moreover, Jack et al. (2014) found that decreases in accuracy were largest when the same piece of misinformation was received twice (from both a co-witness and the interviewer) instead of only once. Future studies should systematically test whether similar to the effect of misinformation MDM can account for the effects of factors that have been proposed to moderate the relation between confidence and accuracy.

Future studies should also pay attention to *confidence resolution*. The meta-analysis revealed that, in contrast to confidence calibration, witnesses' confidence resolution in memory for events was very poor. To reiterate, while witnesses' mean confidence was well aligned with their overall testimony accuracy, confidence ratings only poorly differentiated between single correct and incorrect responses. This divergence parallels previous research findings that conditions under which confidence calibration is maximized may impair resolution (Keren, 1991; Yates, 1982). Intriguingly, poor confidence resolution may particularly impair the perceived credibility of witnesses with high overall accuracy, because confidence resolution has been suggested to be perceived as particularly informative regarding the abilities of individual decision-makers (Yates, Price, Lee, & Ramirez, 1996). In this vein, errors that witnesses made were found to damage their credibility more if witnesses were confident in their reports (Tenney, MacCoun, Spellman, & Hastie, 2007).

Given the influence of confidence resolution on the perception of the accuracy of witnesses' testimonies, it is highly relevant to understand what factors lead to good or poor resolution. As outlined in Chapters 2.3 and shown empirically for overconfidence (Chapter

2.4, Appendix B), adopting a theoretical approach will be fruitful for future research. Again, the MDM might be helpful in understanding why resolution of witnesses' confidence ratings about the accuracy of their accounts of observed events is low (next to the resolution-calibration divergence). According to MDM, confidence should be maximal for responses based on direct retrievals – this should be true for correct responses and also for incorrect responses. This lack of discrimination may contribute to low resolution, but cannot fully account for this finding observed in the meta analysis because only few responses based on direct retrievals are expected to be incorrect. Therefore, familiarity-based retrievals need to be inspected. Confidence for familiarity-based retrievals for questions with two answer options is predicted to increase with the difference in perceived familiarity between the two answer options. Resolution would therefore be expected to be high for familiarity-based retrievals if only for correct responses one response option is much more familiar than the other response option. For incorrect responses, both response options would have to be perceived to be almost equally familiar resulting in lower confidence levels. This would be the case if, for example, (a) all script-relevant items (for which response options should differ largely in perceived familiarity resulting in higher confidence levels) would be answered correctly, whereas (b) script-neutral items (for which response options should be similar in perceived familiarity, cf. Chapter 2.4 and Appendix B, resulting in lower confidence levels) would be answered incorrectly. This is a rather unrealistic scenario. Regarding (a), this should be true for script-conforming items, if the probed details were in line with the script, and for script non-conforming items, if the probed details were in conflict with the script. If the crime under investigation is however not perfectly in line with existing scripts, correct and incorrect responses are similarly likely to be associated with high and low confidence ratings. Regarding (b), the probability of familiarity-based responses to script-neutral items being correct is at the level of chance. Because confidence ratings are also expected to approach the level of chance, regardless of whether the response is correct or not, no systematic differences

in confidence between correct and incorrect responses are to be expected. Applying MDM to witness memory is, thus, able to account for low resolution. In addition, this reasoning can explain the co-occurrence of good calibration and poor resolution. Mean confidence should – on average – be aligned with the relatively high accuracy for script conforming items (that are expected to be more frequent than script-neutral and script-nonconforming items, cf. Chapter 2.4) and the relatively low accuracy for script-neutral items.

Study 3 confirmed previous findings that aggregation is a viable approach to assessing the accuracy of witnesses' testimonies and to reconstructing crimes. As outlined in Chapter 3.2, the validity of aggregated crime reconstructions may however suffer when responses of witnesses are not independent of each other. If this is the case, witnesses' responses are informed not only by the observed crime, but also by other knowledge structures that witnesses share and that influence their responding, such as scripts, co-witness talk, or leading questions (cf. also Chapter 3.2 and Appendix C). The influence of scripts on witnesses' reports has been found to increase with time elapsed after the crime observation (Tuckey & Brewer, 2003a, 2003b). Because witnesses may provide testimony after some time has elapsed, it is important to understand the influence of scripts on aggregated crime reconstructions. Similarly, co-witness talk seems particularly problematic, because empirical studies have found that most witnesses discuss their observations with co-witnesses (86% in Paterson & Kemp, 2006; 58% in Skagerberg & Wright, 2008). It is unknown how *robust* the models underlying the aggregation rules are and to what extent individual testimonies suffer from potential distortions. Future research should focus on examining the conditions, under which aggregation works in favor or against the validity of crime reconstructions. In particular, these future studies should investigate whether given script-based responding or co-witness talk, consensus reconstructions are impaired more than for example majority reconstructions and individual testimonies.

As in Study 2, a memory model was employed in Study 3. The GCM was found to provide a valid model of eyewitnesses' recognition memory for events. The 2-HTM, that can be seen as a simplified version of the GCM assuming homogeneous item difficulties and a known answer key, has recently been applied to separate knowledge from guessing processes in eyewitness identification decisions. Traditionally, studies using ratio-based measures to assess identification performance have found sequential lineups to be superior to simultaneous lineups (cf. Steblay, Dysart, Fulero, & Lindsay, 2001). However, ratio-based measures of performance are confounded with witnesses' guessing biases (i.e., whether their response criterion for choosing a suspect is rather liberal or conservative). When employing the 2-HTM to control for guessing biases, simultaneous lineups were found to outperform sequential lineups with regard to the diagnosticity of a witness's identification (Gronlund, Wixted, & Mickes, 2014; Mickes, Flowe, & Wixted, 2012; Mickes, Moreland, Clark, & Wixted, 2014; Wixted, Gronlund, & Mickes, 2014; Wixted & Mickes, 2014). The debate has not yet been resolved (see Wells, Smalarz, & Smith, 2015 for a critique of applying Signal Detection Theory to lineup data), but, nevertheless, the studies by Mickes, Wixted, and colleagues have shown that *theory-driven research may provide new and valuable insights* and may even lead to widely accepted conclusions being abandoned. Memory models such as MDM, the 2-HTM or the GCM should therefore be considered more often when investigating witnesses' memory for events.

In Study 2 and Study 3, confidence and aggregation were investigated separately. It may however be the case that several witnesses observe the same crime and that they all provide confidence ratings regarding the correctness of their reports. In this case, witnesses' testimonies may be aggregated and confidence ratings may be used at the same time to assess the accuracy of witnesses' testimonies. It seems likely that *combining confidence calibration and the aggregation approach* may improve the assessment of witnesses' memory for events even more than employing only one of these approaches. Future research should therefore aim

to integrate confidence ratings into aggregation models and to test whether such models outperform simple aggregation models ignoring confidence ratings.

The studies presented in this doctoral thesis investigated cognitive psychological research questions and neglected *individual differences*. However, several studies have suggested differences between individual witnesses in memory accuracy, confidence, and the realism of confidence (Keren, 1991; Loftus, Levidow, & Duensing, 1992; Morgan et al., 2007; Searcy, Bartlett, & Memon, 2000). It seems desirable to find out whether some witnesses are more accurate, more confident, better calibrated, less overconfident, or show better resolution than others. Linking these measures to other interindividual differences (for example regarding personality or cognitive abilities; cf. Morgan et al., 2007) may directly inform the assessment of an individual's testimony.

This doctoral thesis focused on recognition memory. The items employed in Study 2 and Study 3 were true/false items. In the meta-analysis (Study 1), studies using two-alternative forced-choice (2AFC) items were further included. Recognition items, in particular yes/no or true/false items have been found to be central in police interrogations (Fisher et al., 1987; George & Clifford, 1992; Peterson & Grant, 2001; Wright & Alison, 2004; cf. Chapter 1). However, *other question types* are also relevant. For example, the Cognitive Interview (Geiselman, Fisher, Mackinnon, & Holland, 1986) that is often used in police interrogations employs free and cued recall questions. The Majority Rule can easily be adapted to other response formats. Moreover, other versions of the GCM have been developed to account for different response formats (e.g., three-way network data, Batchelder, Kumbasar, Boyd, 1997; continuous data, Batchelder, Strashny, & Romney, 2010). Future research should investigate the generalizability of and aim to extend the present findings on the realism of confidence, MDM, the GCM and the Majority Rule to other types of memory and other items formats.

4.2. Practical Implications

The present doctoral thesis investigated how the accuracy of witnesses' memory for events can be determined. Witness testimony is often central to the reconstruction of crimes and actors of the legal system perceive witness memory to be highly reliable (e.g., Simons & Chabris, 2011). However, witness testimony has also been identified as a major source of error in legal procedures (see Chapter 1). Yet, particularly in recent years, little effort has been spent on investigating witnesses' memory for events. Rather, attention has been paid to eyewitness identification decisions. This doctoral thesis therefore investigates a core issue of forensic psychological research and the studies presented in this thesis may inform judicial decision-making.

In particular, Study 1 and Study 2 provide information to improve the validity of assessing witnesses' testimony by using confidence ratings. This is important because high confidence ratings, confidence calibration, and confidence resolution are perceived as valid predictors to the accuracy of witnesses' testimonies (McClure et al., 2013; Potter & Brewer, 1999; Simons & Chabris, 2011; Tenney et al., 2007; Tenney, Small, Kondrad, Jaswal, & Spellman, 2011). According to the meta-analysis, witnesses' confidence ratings can be treated as informative with respect to the overall accuracy of the witnesses' reports, but in general, witnesses can be expected to be somewhat more confident than accurate. At the same time, a witness's individual confidence rating is unlikely to be diagnostic with respect to the accuracy of a particular response or recognition judgment. In conclusion, the results from Study 1 suggest that confidence ratings may be used only at a rather general level.

Study 2 may help understand the factors that influence accuracy, confidence, and the realism of confidence and is therefore directly applicable to legal decision-making. In particular, results from Study 2 suggest that when details of a crime are inconsistent with an existing script and witnesses' cognitive capacities were limited when observing the crime (which is likely to be the case in real-life crime observations), confidence ratings may be misleading and details recognized with high certainty may be false. Study 2 also enables

making predictions regarding other potential influences on witnesses' memory for events. For example, an intuitive approach to reducing overconfidence may be to provide witnesses with a warning to not overestimate their own accuracy. MDM as introduced in Study 2 however predicts warnings to be ineffective because the model posits that accuracy and confidence result from largely automatic processes that occur during encoding (Dougherty, 2001) and that cannot be consciously controlled. In line with this prediction, García-Bajos and Migueles (2003) found that cautioning witnesses to report only facts and to not make inferences decreased both the number of correctly and incorrectly recalled details but did not decrease overconfidence. In line with García-Bajos and Migueles (2003), MDM predicts that warning should not be employed to reduce overconfidence, as they are unlikely to be effective and may confuse witnesses or legal decision-makers.

As Study 3 revealed, aggregation rules considering individual differences between witnesses and items may provide more accurate reconstructions of crimes than simple models, on which for example the Majority Rule is based. This was found to be true when competence levels of witnesses were heterogeneous, which is likely to be the case in real-world groups of witnesses. However, the Majority Rule is a rather intuitive rule and can easily be applied. Therefore, actors of the legal system are likely to employ the Majority Rule when aggregating multiple witness testimonies. For example, following the assassination of US-American President John F. Kennedy during a parade in Dallas, Texas, in 1963, over 500 witnesses were interviewed. When questioned about the presumed hiding place of Kennedy's assassinator, 46.5% reported that the shots had come from a nearby schoolbook depository, whereas 20.2% indicated that the shots had been fired from a grassy knoll next to the street. One third of the witnesses reported that the shots had originated from another or from multiple places (President's Commission on the Assassination of President Kennedy, 1964). In line with the Majority Rule, the official report concluded that the assassinator had been hiding in the schoolbook depository. The report did however not consider whether the

witnesses who reported that the shots had come from the grassy knoll had paid more attention to the parade, had been closer to the crime scene, or maybe were generally better at localisating origins of sound. These differences between the witnesses may however have influenced their competences and may even have led to the majority response being false. The results from Study 3 thus strongly recommend that when aggregating multiple witness reports, potential individual differences should be modeled and taken into account, whereas simple aggregation rules that may seem intuitively appealing should be used with caution.

Both Study 2 and Study 3 suggest that any information acquired by multiple witnesses following a crime observation (for example through co-witness talk; cf. Paterson & Kemp, 2006) might impair the assessment of testimony accuracy and, thus, crime reconstructions. Actors of the legal system should therefore strongly discourage witnesses to discuss their observations. Means to protect observations from such influences have recently been suggested by Gabbert, Hope, and Fisher (2009), who introduced the Self-Adminsitered Interview. Because it should be largely free from bias that may, for example, result from co-witness talk or from reading about an observed crime in the media (Hope, Gabbert, & Fisher, 2011), data from the Self-Administered Interview might be particularly suitable for aggregation.

Much of the research conducted in forensic psychology is informed by the legal system of the United States of America. For example, all cases investigated by the Innocence Project were tried in U.S. American courts. Much less is known about the frequency of wrongful convictions and to what extent wrongful eyewitness testimony or unreliable assessments of witness reports contribute to legal errors in Germany, for example. Given the obvious fallibility of eyewitness testimony, it seems likely that legal errors also occur in decisions made by German judges. In particular, Studies 2 and 3 corroborate this claim. The findings from the studies presented in this doctoral thesis may therefore inform researchers and actors of the legal system regardless of the jurisdiction they live or work in. Given the

central importance of witness testimony to legal decision-making, more research should focus on the cognitive processes underlying witness memory for events. This doctoral thesis provided some insights into the cognitive mechanisms underlying witness memory that may inform both forensic psychological researchers and legal decision-makings when investigating or assessing the accuracy of witnesses' reports about observed crime events.

References

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist, 36*, 715-729. doi: 10.1037/0003-066x.36.7.715

Allwood, C. M., Granhag, P. A., & Johansson, M. (2003). Increased realism in eyewitness confidence judgements: The effect of dyadic collaboration. *Applied Cognitive Psychology, 17*, 545-561. doi: 10.1002/acp.888

Allwood, C. M., Innes-Ker, A. H., Homgren, J., & Fredin, G. (2008). Children's and adults' realism in their event-recall confidence in responses to free recall and focused questions. *Psychology, Crime & Law, 14*, 529-547. doi: 10.1080/10683160801961231

Allwood, C. M., Knutsson, J., & Granhag, P. A. (2006). Eyewitnesses under influence: How feedback affects the realism in confidence judgements. *Psychology, Crime & Law, 12*, 25-38. doi: 10.1080/10683160512331316316

Armstrong, J. S. (2004). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting. A handbook for researchers and practitioners* (pp. 417-439). Boston: Kluwer.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55*, 412-428. doi: 10.3758/bf03205299

Batchelder, W. H., Kumbasar, E., & Boyd, J. P. (1997). Consensus analysis of three-way social network data. *Journal of Mathematical Sociology, 22*, 29-58. doi: 10.1080/0022250x.1997.9990193

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (pp. 103-112). Greenwich, CT: JAI Press.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika, 53*, 71-92. doi: 10.1007/BF02294195

Batchelder, W. H., Strashny, A., & Romney, A. K. (2010). Cultural consensus theory: Aggregating continuous responses in a finite interval. *Advances in Social Computing, Lecture Notes in Computer Science, 6007*, 98-107. doi: 10.1007/978-3-642-12079-4_15

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088-1101. doi: 10.2307/2533446

Bernstein, D. M., & Loftus, E. F. (2009). How to tell if a particular memory is true or false. *Perspectives on Psychological Science, 4*, 370-374. doi: 10.1111/j.1745-6924.2009.01140.x

Bjork, R. A. (1973). Why mathematical models? *American Psychologist, 28*, 426-433. doi: 10.1037/h0034623

Boland, P. J. (1989). Majority systems and the Condorcet Jury Theorem. *The Statistician, 38*, 181-189. doi: 10.2307/2348873

Bonham, A. J., & Gonzalez-Vallejo, C. (2009). Assessment of calibration for reconstructed eyewitness memories. *Acta Psychologica, 131*, 34-52. doi: 10.1016/j.actpsy.2009.02.008

Bothwell, R. K., Brigham, J. C., & Deffenbacher, K. A. (1987). Correlational of eyewitness accuracy and confidence. Optimality hypothesis revisited. *Journal of Applied Psychology, 72*, 691-695. doi: 10.1037/0021-9010.72.4.691

Brewer, N. (2006). Uses and abuses of eyewitness identification confidence. *Legal and Criminological Psychology, 11*, 3-23. doi: 10.1348/135532505x79672

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*, 353-364. doi: 10.1023/A:1015380522722

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied, 8*, 44-56. doi: 10.1037//1076-898x.8.1.44

Brewer, N., Weber, N., & Semmler, C. (2007). A role for theory in eyewitness identification research. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology. Volume II. Memory for people* (pp. 201-218). Mahwah, NJ: Lawrence Erlbaum Associates.

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*, 11-30. doi: 10.1037/1076-898X.12.1.11

Buckhout, R. (1974). Eyewitness testimony. *Scientific American, 231*(6), 23-31.

Bulevich, J. B., & Thomas, A. K. (2012). Retrieval effort improves memory and metamemory in the face of misinformation. *Journal of Memory and Language, 67*, 45-58. doi: 10.1016/J.Jml.2011.12.012

Buratti, S., & Allwood, C. M. (2012). The accuracy of meta-metacognitive judgments: Regulating the realism of confidence. *Cognitive Processes, 13*, 243-253. doi: 10.1007/s10339-012-0440-5

Buratti, S., Allwood, C. M., & Johansson, M. (2014). Stability in the metamemory realism of eyewitness confidence judgments. *Cognitive Processes, 15*, 39-53. doi: 10.1007/s10339-013-0576-y

Clark, S. E. (2008). The importance (necessity) of computational modelling for eyewitness identification research. *Applied Cognitive Psychology, 22*, 803-813. doi: 10.1002/acp.1484

Clark, S. E., & Wells, G. L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior, 32*, 406-422. doi: 10.1007/s10979-007-9115-7

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559-583. doi: 10.1016/0169-2070(89)90012-5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cutler, B. L., & Penrod, S. (1989). Forensically relevant moderators of the relation between eyewitness identification accuracy and confidence. *Journal of Applied Psychology, 74*, 650-652. doi: 10.1037/0021-9010.74.4.650

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision-making in eyewitness identification cases. *Law and Human Behavior, 12*, 41-55. doi: 10.1007/bf01064273

Dahl, M., Allwood, C. M., Scimone, B., & Rennemark, M. (2015). Old and very old adults as witnesses: Event memory and metamemory. *Psychology, Crime & Law, 21*, 764-775. doi: 10.1080/1068316X.2015.1038266

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision, 1*, 1-4. doi: 10.1037/dec0000004

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

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General, 130*, 579-599. doi: 10.1037//0096-3445.130.4.579

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*, 180-209. doi: 10.1037/0033-295x.106.1.180

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89-98. doi: 10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias. *Biometrics, 56*, 455-463. doi: 10.1111/j.0006-341X.2000.00455.x

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634. doi: 10.1136/bmj.315.7109.629

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous overconfidence and underconfidence: The role of error in judgment processes. *Psychological Review, 101*, 519-527. doi: 10.1037/0033-295x.101.3.519

Estlund, D. M. (1994). Opinion leaders, independence, and Condorcet's Jury Theorem. *Theory and Decision, 36*, 131-162. doi: 10.1007/BF01079210

Fawcett, J. M., Russell, E. J., Peace, K. A., & Christie, J. (2013). Of guns and geese: A meta-analytic review of the 'weapon focus' literature. *Psychology, Crime & Law, 19*, 35-66. doi: 10.1080/1068316x.2011.599325

Fisher, R. P., Geiselman, R. E., & Raymond, D. S. (1987). Critical analysis of police interview techniques. *Journal of Police Science and Administration, 15*, 177-185.

Gabbert, F., Hope, L., & Fisher, R. P. (2009). Protecting eyewitness evidence: Examining the efficacy of a self-administered interview tool. *Law and Human Behavior, 33*, 298-307. doi: 10.1007/s10979-008-9146-8

Gabbert, F., Memon, A., & Allan, K. (2003). Memory conformity: Can eyewitnesses influence each other's memories for an event? *Applied Cognitive Psychology, 17*, 533-543. doi: 10.1002/acp.885

Galton, F. (1907). Vox populi. *Nature, 75*, 450-451. doi: 10.1038/075450a0

García-Bajos, E., & Migueles, M. (2003). False memories for script actions in a mugging account. *European Journal of Cognitive Psychology, 15*, 195-208. doi: 10.1080/09541440244000102

García-Bajos, E., Migueles, M., & Aizpurua, A. (2012). Bias of script-driven processing on eyewitness memory in young and older adults. *Applied Cognitive Psychology, 26*, 737-745. doi: 10.1002/Acp.2854

Geiselman, R. E., Fisher, R. P., Mackinnon, D. P., & Holland, H. L. (1986). Enhancement of eyewitness memory with the Cognitive Interview. *American Journal of Psychology, 99*, 385-401. doi: 10.2307/1422492

George, R., & Clifford, B. (1992). Making the most of witnesses. *Policing, 8*, 185-198.

Goodman-Delahunty, J., Granhag, P. A., Hartwig, M., & Loftus, E. F. (2010). Insightful or wishful: Lawyers' ability to predict case outcomes. *Psychology, Public Policy, and Law, 16*, 133-157. doi: 10.1037/a0019060

Granhag, P. A. (1997). Realism in eyewitness confidence as a function of type of event witnessed and repeated recall. *Journal of Applied Psychology, 82*, 599-613. doi: 10.1037/0021-9010.82.4.599

Granhag, P. A., Jonsson, A. C., & Allwood, C. M. (2004). The Cognitive Interview and its effect on witnesses' confidence. *Psychology, Crime & Law, 10*, 37-52. doi: 10.1080/1068316021000030577

Green, B. F., & Hall, J. A. (1984). Quantitative methods for literature reviews. *Annual Review of Psychology, 35*, 37-53. doi: 10.1146/annurev.ps.35.020184.000345

Greenberg, M. S., Westcott, D. R., & Bailey, S. E. (1998). When believing is seeing: The effect of scripts on eyewitness memory. *Law and Human Behavior, 22*, 685-694. doi: 10.1023/a:1025758807624

Grofman, B., Owen, G., & Feld, S. (1983). Thirteen theorems in search of the truth. *Theory and Decision, 15*, 261-278. doi: 10.1007/BF00125672

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science, 23*, 3-10. doi: 10.1177/0963721413498891

Hashtroudi, S., Mutter, S. A., Cole, E. A., & Green, S. K. (1984). Schema-consistent and schema-inconsistent information: Processing demands. *Personality and Social Psychology Bulletin, 10*, 269-278. doi: 10.1177/0146167284102013

Higham, P. A., Luna, K., & Bloomfield, J. (2011). Trace-strength and source-monitoring accounts of accuracy and metacognitive resolution in the misinformation paradigm. *Applied Cognitive Psychology, 25*, 324-335. doi: 10.1002/Acp.1694

Hintzman, D. L. (1984). MINERVA 2: A simultation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*, 96-101. doi: 10.3758/BF03202365

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411-428. doi: 10.1037/0033-295X.93.4.411

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528-551. doi: 10.1037/0033-295x.95.4.528

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdoch* (pp. 39-56). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holst, V. F., & Pezdek, K. (1992). Scripts for typical crimes and their effects on memory for eyewitness testimony. *Applied Cognitive Psychology, 6*, 573-587. doi: 10.1002/acp.2350060702

Hope, L., Gabbert, F., & Fisher, R. P. (2011). From laboratory to the street: Capturing witness memory using the Self-Adminstered Interview. *Legal and Criminological Psychology, 16*, 211-226. doi: 10.1111/j.2044-8333.2011.02015.x

Hudson, J. A., Fivush, R., & Kuebli, J. (1992). Scripts and episodes. The development of event memory. *Applied Cognitive Psychology, 6*, 483-505. doi: 10.1002/acp.2350060604

Jack, F., Zydervelt, S., & Zajac, R. (2014). Are co-witnesses special? Comparing the influence of co-witness and interviewer misinformation on eyewitness reports. *Memory, 22*, 243-255. doi: 10.1080/09658211.2013.778291

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology, 5*, 55-71. doi: 10.1080/09541449308406514

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*, 226-246. doi: 10.1006/obhd.1994.1013

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition, 106*, 259-298. doi: 10.1016/j.cognition.2007.02.003

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132*, 133-156. doi: 10.1037/0096-3445.132.1.133

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

Kanazawa, S. (1998). A brief note on a further refinement of the Condorcet Jury Theorem for heterogeneous groups. *Mathematical Social Sciences, 35*, 69-73. doi: 10.1016/S0165-4896(97)00028-0

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika, 68*, 373-389. doi: 10.1007/BF02294733

Karlsson, L., Juslin, P., & Olsson, H. (2008). Exemplar-based inference in multi-attribute decision making: Contingent, not automatic, strategy shifts? *Judgment and Decision Making, 3*, 244-260. Retrieved from: http://www.sjdm.org/~baron/journal/bn5.pdf

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research. A new survey of the experts. *American Psychologist, 56*, 405-416. doi: 10.1037//0003-066x.56.5.405

Kazmann, R. G. (1973). Democratic organization: A preliminary mathematical model. *Public Choice, 16*, 17-26. doi: 10.1007/BF01718803

Kebbell, M. R., Evans, L., & Johnson, S. D. (2010). The influence of lawyers' questions on witness accuracy, confidence, and reaction times and on mock jurors' interpretation of witness accuracy. *Journal of Investigative Psychology and Offender Profiling, 7*, 261-271. doi: 10.1002/jip.125

Kebbell, M. R., & Giles, D. C. (2000). Some experimental influences of lawyers' complicated questions on eyewitness confidence and accuracy. *Journal of Psychology, 134*, 129-139. doi: 10.1080/00223980009600855

Keren, G. (1991). Calibration and probability judgments. Conceptual and methodological issues. *Acta Psychologica, 77*, 217-273. doi: 10.1016/0001-6918(91)90036-y

Kleider, H. M., Pezdek, K., Goldinger, S. D., & Kirk, A. (2008). Schema-driven source misattribution errors: Remembering the expected from a witnessed event. *Applied Cognitive Psychology, 22*, 1-20. doi: 10.1002/acp.1361

Köhnken, G., Milne, R., Memon, A., & Bull, R. (1999). The Cognitive Interview: A meta-analysis. *Psychology, Crime & Law, 5*, 3-27. doi: 10.1080/10683169908414991

Ladha, K. K. (1992). The Condorcet Jury Theorem, free speech, and correlated votes. *American Journal of Political Science, 36*, 617-634. doi: 10.2307/2111584

Lane, S. M., & Meissner, C. A. (2008). A 'middle road' approach to bridging the basic-applied divide in eyewitness identification research. *Applied Cognitive Psychology, 22*, 779-787. doi: 10.1002/acp.1482

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159-183. doi: 10.1016/0030-5073(77)90001-0

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty* (pp. 275-324). Cambridge: Cambridge University Press.

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior, 24*, 685-697. doi: 10.1023/A:1005504320565

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science, 9*, 215-218. doi: 10.1111/1467-9280.00041

Loftus, E. F. (1975a). Leading questions and the eyewitness report. *Cognitive Psychology, 7*, 560-572. doi: 10.1016/0010-0285(75)90023-7

Loftus, E. F. (1975b). Reconstructing memory: The incredible eyewitness. *Jurimetrics Journal, 15*, 188-193. Retrieved from: http://www.jstor.org/stable/29761487

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

Loftus, E. F., Levidow, B., & Duensing, S. (1992). Who remembers best? Individual differences in memory for events that ccurred in a science museum. *Applied Cognitive Psychology, 6*, 93-107. doi: 10.1002/acp.2350060202

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine, 20*, 641-654. doi: 10.1002/sim.698

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23*, 77-87. doi: 10.1002/ejsp.2420230107

McClure, K. A., Myers, J. J., & Keefauver, K. M. (2013). Witness vetting: What determines detectives' perceptions of witness credibility? *Journal of Investigative Psychology and Offender Profiling, 10*, 250-267. doi: 10.1002/jip.1391

Meade, M. L., & Roediger, H. L. (2002). Explorations in the social contagion of memory. *Memory & Cognition, 30*, 995-1009. doi: 10.3758/BF03194318

Memon, A., Mastroberardino, S., & Fraser, J. (2008). Münsterberg's legacy: What does eyewitness research tell us about the reliability of eyewitness testimony? *Applied Cognitive Psychology, 22*, 841-851. doi: 10.1002/acp.1487

Memon, A., Meissner, C. A., & Fraser, J. (2010). The Cognitive Interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law, 16*, 340-372. doi: 10.1037/a0020518

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361-376. doi: 10.1037/a0030609

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute d', not the diagnosticity ratio. *Journal of Applied Research in Memory and Cognition, 3*, 58-62. doi: 10.1016/j.jarmac.2014.04.007

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*, 502-517. doi: 10.1037/0033-295X.115.2.502

Morgan, C. A., Hazlett, G., Baranoski, M., Doran, A., Southwick, S., & Loftus, E. (2007). Accuracy of eyewitness identification is significantly associated with performance on

a standardized test of face recognition. *International Journal of Law and Psychiatry, 30*, 213-223. doi: 10.1016/j.ijlp.2007.03.005

Münsterberg, H. (1908). *On the witness stand. Essays on psychology and crime*. New York: Doubleday, Page & Co.

O'Toole, T. P., & Shay, G. (2006). Manson v. Brathwaite revisited: Towards a new rule of decision for due process challenges to eyewitness identification procedures. *Valparaiso University Law Review, 41*, 109-148. Retrieved from: http://scholar.valpo.edu/vulr/vol41/iss1/2

Ogloff, J. R. P. (2000). Two steps forward and one step backward: The law and psychology movement(s) in the 20th century. *Law and Human Behavior, 24*, 457-483. doi: 10.1023/A:1005596414203

Olsson, N. (2000). A comparison of correlation, calibration, and diagnosticity as measures of the confidence-accuracy relationship in witness identification. *Journal of Applied Psychology, 85*, 504-511. doi: 10.1037//0021-9010.85.4.504

Olsson, N., & Juslin, P. (2002). Calibration of confidence among eyewitnesses and earwitnesses. In P. Chambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition. Process, function and use* (pp. 203-218). New York: Springer.

Oravecz, Z., Vandekerckhove, J., & Batchelder, W. H. (2014). Bayesian Cultural Consensus theory. *Field Methods, 26*, 207-222. doi: 10.1177/1525822X13520280

Palmer, M. A., 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. doi: 10.1037/a0031602

Paterson, H. M., & Kemp, R. I. (2006). Co-witness talk: A survey of eyewitness discussion. *Psychology, Crime & Law, 12*, 181-191. doi: 10.1080/10683160512331316334

Perfect, T. J. (2002). When does eyewitness confidence predict performance? In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 95-120). Cambridge: Cambridge University Press.

Perfect, T. J., & Hollins, T. S. (1996). Predictive feeling of knowing judgements and postdictive confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology, 10*, 371-382. doi: 10.1002/(sici)1099-0720(199610)10:5<371::aid-acp389>3.0.co;2-o

Peterson, C., & Grant, M. (2001). Forced-choice: Are forensic interviewers asking the right questions? *Canadian Journal of Behavioural Science (Revue Canadienne Des Sciences Du Comportement), 33*, 118-127. doi: 10.1037/h0087134

Potter, R., & Brewer, N. (1999). Perceptions of witness behaviour-accuracy relationships held by police, lawyers and mock-jurors. *Psychiatry, Psychology and Law, 6*, 97-103. doi: 10.1080/13218719909524952

President's Commission on the Assassination of President Kennedy. (1964). Report of the President's Commission on the Assassination of President Kennedy. Washington, DC: U.S. Government Printing Office. Retrieved from http://www.archives.gov/research/jfk/warrencommission-report/letter.html

R Development Core Team. (2015). The R-project for statistical computing. Retrieved from http://www.r-project.org/

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence-accuracy correlation. *Journal of Applied Psychology, 81*, 587-594. doi: 10.1037/0021-9010.81.5.587

Romney, A. K. (1999). Consensus as a statistical model. *Current Anthropology, 40*, 103-115. doi: 10.1086/200062

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist, 88*, 313-338. doi: 10.1525/aa.1986.88.2.02a00020

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82. doi: 10.1146/annurev.psych.52.1.59

Sanders, G. S., & Warnick, D. H. (1982). Evaluating identification evidence from multiple eyewitnesses. *Journal of Applied Social Psychology, 12*, 182-192. doi: 10.1111/j.1559-1816.1982.tb00858.x

Sarwar, F., Sikström, S., M., A. C., & Innes-Ker, A. H. (2015). Predicting correctness of eyewitness statements using the semantic evaluation method (SEM). *Quality & Quantity, 49*, 1735-1745. doi: 10.1077/s11135-014-9997-7

Searcy, J., Bartlett, J. C., & Memon, A. (2000). Influence of post-event narratives, line-up conditions and individual differences on false identification by young and older witnesses. *Legal and Criminological Psychology, 5*, 219-235. doi: 10.1348/135532500168100

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

Sharman, S. J., & Powell, M. B. (2012). A comparison of adult witnesses' suggestibility across various types of leading questions. *Applied Cognitive Psychology, 26*, 48-53. doi: 10.1002/acp.1793

Shaw, J. S., Appio, L. M., Zerr, T. K., & Pontoski, K. E. (2007). Public eyewitness confidence can be influenced by the presence of other witnesses. *Law and Human Behavior, 31*, 629-652. doi: 10.1007/s10979-006-9080-6

Shaw, J. S., Garven, S., & Wood, J. M. (1997). Co-witness information can have immediate effects on eyewitness memory reports. *Law and Human Behavior, 21*, 503-523. doi: 10.1023/a:1024875723399

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the U.S. population. *Plos One, 6*, 1-7. doi: 10.1371/journal.pone.0022757

Skagerberg, E. M., & Wright, D. B. (2008). The prevalence of co-witnesses and co-witness discussions in real eyewitnesses. *Psychology, Crime & Law, 14*, 513-521. doi: 10.1080/10683160801948980

Smith, V. L., Ellsworth, P. C., & Kassin, S. M. (1989). Eyewitness accuracy and confidence: Within- versus between-subjects correlations. *Journal of Applied Psychology, 74*, 356-359. doi: 10.1037/0021-9010.74.2.356

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34-50. doi: 10.1037/0096-3445.117.1.34

Sporer, S. L. (1996). Psychological aspects of person descriptions. In S. L. Sporer, R. S. Malpass, & G. Köhnken (Eds.), *Psychological issues in eyewitness identification* (pp. 53-86). Mahwah, NJ: Lawrence Erlbaum Associates.

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. doi: 10.1037//0033-2909.118.3.315

Steblay, N. (1992). A meta-analytic review of the weapon focus effect. *Law and Human Behavior, 16*, 413-424. doi: 10.1007/bf02352267

Steblay, N., Dysart, J. E., Fulero, J., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25*, 459-473. doi: 10.1023/A:1012888715007

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53*, 1119-1129. doi: 10.1016/S0895-4356(00)00242-0

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science, 18*, 46-50. doi: 10.1111/j.1467-9280.2007.01847.x

Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. *Developmental Psychology, 47*, 1065-1077. doi: 10.1037/a0023273

Towfigh, E. V., & Glöckner, A. (2011). GAME OVER: Empirical support for soccer bets regulation. *Psychology, Public Policy, and Law, 17*, 475-506. doi: 10.1037/a0023402

Tuckey, M. R., & Brewer, N. (2003a). How schemas affect eyewitness memory over repeated retrieval attempts. *Applied Cognitive Psychology, 17*, 785-800. doi: 10.1002/acp.906

Tuckey, M. R., & Brewer, N. (2003b). The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied, 9*, 101-118. doi: 10.1037/1076-898x.9.2.101

Turtle, J., Read, J. D., Lindsay, D. S., & Brimacombe, C. A. E. (2008). Toward a more informative psychological science of eyewitness evidence. *Applied Cognitive Psychology, 22*, 769-778. doi: 10.1002/acp.1481

Tyszka, T., & Zielonka, P. (2002). Expert judgments: Financial analysts versus weather forecasters. *The Journal of Psychology and Financial Markets, 3*, 152-160. doi: 10.1207/S15327760JPFM0303_3

Wagenaar, W. A. (1988). Calibration and the effects of knowledge and reconstruction in retrieval from memory. *Cognition, 28*, 277-296. doi: 10.1016/0010-0277(88)90016-9

Waubert de Puiseau, B., Aßfalg, A., Erdfelder, E., & Bernstein, D. M. (2012). Extracting the truth from conflicting eyewitness reports: A formal modeling approach. *Journal of Experimental Psychology: Applied, 18*, 390-403. doi: 10.1037/a0029801

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*, 490-499. doi: 10.1037/0021-9010.88.3.490

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied, 10*, 156-172. doi: 10.1037/1076-898X.10.3.156

Weber, N., & Varga, M. (2012). Can a modified lineup procedure improve the usefulness of confidence? *Journal of Applied Research in Memory and Cognition, 1*, 152-157. doi: 10.1016/j.jarmac.2012.06.007

Weller, S. C. (1987). Shared knowledge, intracultural variation, and knowledge aggregation. *American Behavioral Scientist, 31*, 178-193. doi: 10.1177/000276487031002004

Weller, S. C. (2007). Cultural Consensus Theory: Applications and frequently asked questions. *Field Methods, 19*, 339-368. doi: 10.1177/1525822X07303502

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

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence. Improving its probative value. *Psychological Science in the Public Interest, 7*, 45-75. doi: 10.1111/j.1529-1006.2006.00027.x

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

Wells, G. L., & Murray, D. M. (1987). Eyewitness confidence. In G. L. Wells (Ed.), *Eyewitness testimony*. Cambridge: Cambridge University Press.

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54*, 277-295. doi: 10.1146/annurev.psych.54.101601.145028

Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition, 4*, 313-317. doi: 10.1016/j.jarmac.2015.08.008

Wheatcroft, J. M., Kebbell, M. R., & Wagstaff, G. F. (2001). Courtroom questioning. *Forensic Update, 65*, 20-25.

Wise, R. A., Pawlenko, N. B., Safer, M. A., & Meyer, D. (2009). What US prosecutors and defence attorneys know and believe about eyewitness testimony. *Applied Cognitive Psychology, 23*, 1266-1281. doi: 10.1002/acp.1530

Wise, R. A., & Safer, M. A. (2004). What US judges know and believe about eyewitness testimony. *Applied Cognitive Psychology, 18*, 427-443. doi: 10.1002/acp.993

Wixted, J. T., Gronlund, S. D., & Mickes, L. (2014). Policy regarding the sequential lineup is not informed by probative value but is informed by receiver operating characteristic analysis. *Current Directions in Psychological Science, 23*, 17-28. doi: 10.1177/0963721413510934

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*, 262-276. doi: 10.1037/a0035940

Wright, A. M., & Alison, L. (2004). Questioning sequences in Canadian police interviews: Constructing and confirming the course of events? *Psychology, Crime & Law, 10*, 137-154. doi: 10.1080/1068316031000099120

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*, 611-617. doi: 10.1037/0033-2909.110.3.611

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30*, 132-156. doi: 10.1016/0030-5073(82)90237-9

Yates, J. F., Price, P. C., Lee, J.-W., & Ramirez, J. (1996). Good probabilistic forecasters: The 'consumer's' perspective. *International Journal of Forecasting, 12*, 41-56. doi: 10.1016/0169-2070(95)00636-2

Appendix

List of Appendices

Appendix A Study 1: Waubert de Puiseau, B. & Musch, J. (2016). A meta-analysis of the realism of confidence in eyewitness event memory. *Manuscript submitted for publication.* 78

Appendix B Study 2: Waubert de Puiseau, B., Weide, A. C., & Musch, J. (2016). How scripts influence overconfidence in eyewitness event memory: A model-based analysis. *Manuscript submitted for publication.* 117

Appendix C Study 3: Waubert de Puiseau, B., Platzek, S., Aßfalg, A., & Musch, J. (2016). On the importance of considering heterogeneity in witnesses' competence levels when reconstructing crimes from multiple witness testimonies. *Manuscript submitted for publication.* 164

Appendix A

Study 1:

Waubert de Puiseau, B. & Musch, J. (2016). A meta-analysis of the realism of confidence in eyewitness event memory. *Manuscript submitted for publication.*

A Meta-Analysis of the Realism of Confidence in Eyewitness Event Memory

Berenike Waubert de Puiseau and Jochen Musch

University of Duesseldorf, Germany

Word count (excluding Abstract, Tables, Figures, and References): 6,390

Author Note

Berenike Waubert de Puiseau, Institute of Experimental Psychology, University of Duesseldorf, Germany; Jochen Musch, Institute of Experimental Psychology, University of Duesseldorf, Germany.

The authors thank Svenja Jessica Löffler and Julia Meisters for their support in conducting this research.

Correspondence concerning this article should be addressed to Berenike Waubert de Puiseau, Department of Experimental Psychology, Institute of Experimental Psychology, University of Duesseldorf, Universitaetsstraße 1, 40225 Duesseldorf, Germany, phone: +49 – (0)211 – 8112063, fax: +49 – (0)211 – 121753, e-mail: bwdp@hhu.de

Abstract

We present the first meta-analysis of the realism of confidence in eyewitness event recognition memory measured in terms of confidence calibration, under-/overconfidence, and resolution. Across eight studies reporting 24 independent assessments of the confidence-accuracy relationship, we found that witnesses were slightly but significantly miscalibrated (*C* = .048), significantly overconfident (*U/O* = .095), and poor in resolution (*NRI* = .025). Almost all of the evaluated studies had employed the same crime video (from Granhag, 1997) and two-alternative forced-choice items. A moderator analysis revealed systematic influences of stimulus material and item type on the realism of confidence. Correcting effect size estimates for publication bias resulted in slightly reduced effect size estimates for calibration (*C* = .044) and resolution (*NRI* = .021). We recommend more methodological diversity in future studies to enable stronger conclusions about the generalizability and replicability of the findings.

Word count (Abstract): 139

Keywords: confidence-accuracy relationship; eyewitness memory; calibration analysis; overconfidence; meta-analysis

A Meta-Analysis of the Realism of Confidence in Eyewitness Event Memory

Witnesses play an important role in legal fact-finding processes and have substantial influences on trial outcomes. However, empirical studies have shown that witness memory is often flawed (for reviews, see Wells, Memon, & Penrod, 2006; Wells & Olson, 2003). When faulty witness testimony influences legal decision-making, justice is at stake; and mistaken eyewitness identification evidence has been identified as the single most important contributor to wrongful convictions (Innocence Project, 2015).[1] An accurate assessment of the validity of eyewitness testimony is therefore important. Confidence ratings expressing the witness's subjective certainty are often employed as an index of the accuracy of a testimony. The use of confidence ratings is based on the assumption that witnesses can provide accurate meta-cognitive evaluations of their memory performance. Supporting this assumption, confidence ratings have been found to be linked to actual accuracy in several research domains including general knowledge questions (e.g., Perfect & Hollins, 1996).

Witnesses' ratings of their confidence in the accuracy of their testimony have been found to influence both lay people's and legal experts' perceptions of the accuracy of witness testimonies (Brewer, 2006; Kassin, Tubb, Hosch, & Memon, 2001; Simons & Chabris, 2011; Wise, Pawlenko, Safer, & Meyer, 2009; Wise & Safer, 2004). However, this reliance on witness confidence is at odds with research findings that suggest that the relationship between confidence and accuracy may be less stable for eyewitness than for semantic memory (Hollins & Perfect, 1997; Luna & Martín-Luengo, 2012; Perfect, 2002; Perfect & Hollins, 1996).

There are two limitations to the existing literature on the confidence-accuracy relationship in eyewitness memory. First, most existing studies have investigated the confidence-accuracy relationship for eyewitness identification testimony. It is unknown whether findings from studies in which participants were required to identify faces or individuals from groups of people can be generalized to witnesses' memories of crime events. Eyewitness event memory also includes memory for actions, conversations, and surroundings.

Even though eyewitness event memory plays an important role in legal processes, little is known about the validity of confidence ratings with regard to the accuracy of eyewitness event memory (Allwood, Knutsson, & Granhag, 2006; Hollins & Perfect, 1997; Luna & Martín-Luengo, 2012; Sporer, 1996). A second limitation is that most studies that have investigated the confidence-accuracy relationship employed correlation analyses, which are of limited use for the assessment of the realism of an individual witness's confidence ratings (Juslin, Olsson, & Winman, 1996; Kebbell, Wagstaff, & Covey, 1996). Instead, an analysis of the calibration of confidence is necessary to assess the correspondence rather than the covariation between accuracy and confidence (e.g., Lichtenstein & Fischhoff, 1977; Wagenaar, 1988).

In this article, we present—to our knowledge—the first meta-analysis of studies that investigated the realism of confidence in terms of the calibration of confidence in eyewitness event recognition memory. We limited our analysis to studies that examined recognition memory because – contrary to the recommendation to employ recall questions – witnesses are typically required to make recognition judgments in legal interviews (Fisher, Geiselman, & Raymond, 1987; George & Clifford, 1992; Peterson & Grant, 2001). For example, an investigation of 19 Canadian police interviews revealed that throughout all interviews, police officers primarily asked closed questions; open questions were mostly posed in the beginning or in the end of an interview (Wright & Alison, 2004). This failure to employ more sophisticated interviewing techniques based on free or cued recall may be caused by a lack of time available for interviewing witnesses during crime investigations (Kebbell, Milne, & Wagstaff, 1999). Empirical findings suggest that cognitive processes influencing confidence and accuracy differ between recognition and recall questions (cf. Perfect, 2002; Robinson & Johnson, 1996). Studies that investigated cued recall or free recall were, therefore, not part of the present investigation.

In the following, we first present research on the confidence-accuracy relationship in eyewitness memory. Next, we introduce how calibration analyses can be performed and discuss their merits. Last, we describe the results of a meta-analysis of the realism of confidence in eyewitness event recognition memory.

Correlation Analyses of the Confidence-Accuracy Relationship in Eyewitness Memory

For the domain of eyewitness identification, several meta-analyses and reviews have found low or at best moderate correlations between confidence ratings and the accuracy of identification decisions (Bothwell, Brigham, & Deffenbacher, 1987; Cutler & Penrod, 1989; Deffenbacher, 1980; Sporer, Penrod, Read, & Cutler, 1995; Wells & Murray, 1984). Much less research has been conducted on the confidence-accuracy relationship in eyewitness event memory. It is therefore unclear whether the confidence-accuracy relationship is comparably weak for the memory of events (Allwood et al., 2006).

In a typical study investigating confidence in eyewitness event memory, participants observe a simulated crime and subsequently answer questions about the observed event either in an interview or in writing. Directly after responding or after some delay, witnesses are asked to assign a confidence rating to each answer. To assess the relationship between confidence and accuracy, correlation coefficients can be computed across individual witnesses' mean confidence and accuracy (between-subjects correlations), resulting in one correlation coefficient per sample. Alternatively, correlation coefficients can be computed across the confidence ratings that witnesses assign to each answer and the correctness of their respective responses (within-subjects correlations; Nelson, 1984; Smith, Ellsworth, & Kassin, 1989). These analyses result in one correlation coefficient for each witness.

Between-subjects correlations assessing the confidence-accuracy relationship for eyewitness event memory have often been found to be weak. A summary of seven articles containing 12 studies that assessed the confidence-accuracy relationship for general knowledge and eyewitness event memory using recognition and recall items reported a mean

between-subjects correlation of $r = .21$ for eyewitness memory and $r = .51$ for general knowledge (Perfect, 2002). In a direct comparison, within-subjects correlations were found to be higher and more stable than between-subjects correlations (Robinson & Johnson, 1996). However, research in recent years has produced somewhat mixed results with within-subjects correlation coefficients ranging from .10 (Wheatcroft, Kebbell, & Wagstaff, 2001) to over .60 (Bulevich & Thomas, 2012; Kebbell, Evans, & Johnson, 2010; Kebbell & Giles, 2000) for recognition questions. Analyses of the correlation between confidence and accuracy have therefore led to the conclusion that this relationship is modest in size at best and that confidence ratings cannot be used as reliable predictors of the accuracy of testimony for eyewitness event recognition memory (Odinot, Wolters, & van Koppen, 2009).

However, assessing the relationship between confidence and accuracy in terms of correlation coefficients bears two problems. The size of a correlation coefficient in a sample depends on the variance in accuracy and confidence and also on the shapes of their distributions. A low correlation might indicate a weak association between accuracy and confidence, a small amount of variance in accuracy, a small amount of variance in confidence, or any combination thereof (Juslin et al., 1996; Kebbell et al., 1996). Small amounts of variance in accuracy or confidence might occur when homogeneous student samples are assessed under highly standardized conditions (Brewer, 2006; Lindsay, Nilsen, & Read, 2000; Lindsay, Read, & Sharma, 1998). Correlation coefficients obtained in standard eyewitness studies might thereby underestimate the validity of confidence ratings as predictors of testimony accuracy.

Another limitation of correlation analyses is that between-subjects correlations can be interpreted at only the group level. Therefore, within-subjects correlations have been argued to be more informative for the assessment of the accuracy of eyewitness testimony (Juslin et al., 1996). However, neither between-subjects nor within-subjects correlations are diagnostic with respect to a single response. As Brewer (2006) pointed out, the “interpretation of the

point biserial correlation is not straightforward in the forensic context. For example, it is not clear how knowing that the [confidence-accuracy] correlation is .23 or .37 should contribute to a juror's interpretation of the likelihood that a witness may be accurate when that witness has reported 90% confidence in their identification" (p. 11).

Calibration Analysis of the Confidence-Accuracy Relationship in Eyewitness Memory

To overcome the limitations of correlation analyses, calibration analysis has been proposed (Lichtenstein & Fischhoff, 1977; Lichtenstein, Fischhoff, & Phillips, 1982; Olsson, 2000; Wagenaar, 1988). Calibration analysis can reveal meaningful relationships between confidence and accuracy even when correlation analyses suggest that there is only a weak association. Juslin et al. (1996) showed that participants might be either poorly, rather well, or even perfectly calibrated under the minimum requirement that the confidence-accuracy correlation is not negative. Calibration analysis is also less vulnerable to a skewed distribution for confidence or accuracy and thereby circumvents the major shortcomings of correlation analyses (Juslin et al., 1996).

The analysis of confidence calibration has been employed extensively in research on judgment and decision-making (cf. Lichtenstein et al., 1982). The relationship between confidence and accuracy can be assessed visually in terms of calibration curves or by means of three statistics that together can be used as complementary indices of the realism of confidence: the calibration coefficient *C*, the under-/overconfidence coefficient *U/O*, and the normalized resolution index *NRI* (e.g., Baranski & Petrusic, 1994; Brewer & Wells, 2006; Lichtenstein & Fischhoff, 1977; Yaniv, Yates, & Smith, 1991). The calibration coefficient *C* describes the extent to which the observed calibration curve deviates from a perfect calibration curve. This coefficient ranges from 0 (no deviation and, thus, a perfect calibration) to 1 (maximum deviation and, thus, very poor calibration). To compute *C*, the confidence scale is divided into *J* class intervals, and each class interval is inspected separately. The following formula is used to compute *C*:

$$C = \frac{1}{n}\sum_{j=1}^{J} n_j \left(c_j - a_j\right)^2, \quad (1)$$

where n_j is the number of observations in class interval *j*, c_j is the mean confidence level in class *j*, and a_j is the proportion of correct responses in class interval *j*.

The under-/overconfidence statistic *U/O* is computed in a similar manner except that differences between mean confidence levels c_j and the respective proportions of correct responses a_j are not squared. Consequently, *U/O* ranges from -1 to +1 with figures below 0 indicating underconfidence and figures above 0 indicating overconfidence. When witnesses are overconfident, their confidence exceeds their accuracy; that is, they overestimate their performance. The opposite is true when witnesses are underconfident.

As a third measure of the realism of confidence, the resolution of witnesses' confidence ratings can be computed. The resolution of witnesses' confidence ratings refers to whether they can discriminate between correct and incorrect recognition judgments. As the measure of resolution, the *NRI* ranges from 0 (no discrimination between correct and incorrect recognition judgments, that is, no resolution) to 1 (perfect discrimination, that is, perfect resolution). To compute the *NRI*, a normalized sum of the squared differences between the proportions of correct responses a_j in each class interval *j* and the grand proportion of correct responses *a* is computed:

$$NRI = \frac{\frac{1}{n}\sum_{j=1}^{J} n_j \left(a_j - a\right)^2}{a(1-a)} \quad (2)$$

The interpretation of the effect size of *NRI* is analogous to η^2 (Baranski & Petrusic, 1994); that is, values below .06 may be considered small, and values exceeding .13 should be considered large (Cohen, 1988).

For identification decisions, a meta-analysis of seven studies and 52 independent assessments of the confidence-accuracy relationship (24 visual face identification tasks and 28 auditory voice identification tasks) reporting both calibration and correlation analyses revealed a mean correlation between confidence and accuracy of r_{pb} = .30, with a range from r_{pb} = .00 to r_{pb} = .63 (Olsson & Juslin, 2002). Approximately one third of the studies showed virtually no miscalibration and reported calibration coefficients that were not significantly different from zero. Some earwitness studies reported calibrations of up to .4, whereas the maximum calibration reported for face identifications was below .2. Witnesses were more overconfident when making an auditory voice identification and when a full-range scale (from 0 to 1) instead of a half-range scale (from .5 to 1) was used. Overall, eyewitnesses showed little overconfidence. For face identifications, overconfidence ranged from .2 to .3; for voice identifications, the maximum level of overconfidence was approximately .6. It is important to note that correlation coefficients were found to be independent of the calibration coefficient *C* (r = -.11, *ns*). This finding supports the notion that the results of correlation analyses cannot readily be generalized to measures of the realism of confidence (cf. Juslin et al., 1996). Altogether, the meta-analysis by Olsson and Juslin (2002) suggested the relatively high validity of confidence as a predictor of testimony accuracy.

In recent years, a growing number of empirical studies have investigated the realism of confidence in eyewitness event recognition memory. A multitude of experimental factors were manipulated in these studies. Witnesses responded either individually or in pairs (Allwood, Granhag, & Johansson, 2003), received correct or incorrect feedback after responding (Allwood et al., 2006), perceived correct or incorrect information following a crime stimulus (Bonham & Gonzalez-Vallejo, 2009), were administered the same questions

multiple times (Granhag, Jonsson, & Allwood, 2004), and discussed their observation of the crime before testifying (Sarwar, Allwood, & Innes-Ker, 2014). Most of the studies employed rather small samples. The small sizes of the samples and the diversity of the experimental manipulations make it difficult to draw firm conclusions about the realism of confidence in eyewitness event recognition memory.

The Present Study

To the best of our knowledge, no systematic review has been conducted to summarize and synthesize the results of previous studies on the realism of confidence judgments in eyewitness event recognition memory. We therefore conducted what we believe to be the first meta-analysis of this issue. We restricted our analyses to studies that employed recognition questions and that used crime simulations as the stimulus material. Unlike real case studies, crime simulations provide an objective criterion that can be used to compute the accuracy of witnesses' reports.

We identified eight articles that investigated the confidence-accuracy relationship in terms of the realism of confidence in witnesses' recognition judgments of crime events. We restricted our meta-analysis to published studies to ensure that our conclusions would match the ones that a potential practitioner or expert witness would draw if asked to make a statement in court about the realism of confidence in eyewitness event recognition memory using the results obtainable from the published literature.

Because we investigated measures of relationships and not differences, we treated experimental conditions as study units. In the eight articles included in the meta-analysis, we identified 24 independent study units as relevant. Wherever possible, we computed effect size measures of calibration, under-/overconfidence, and/or resolution for each study unit. As control variables, we also computed effect size estimates of accuracy and confidence. Begg's rank correlation test (Begg & Mazumdar, 1994), Egger's linear regression test (Egger, Davey Smith, Schneider, & Minder, 1997), and the trim-and-fill procedure (Duval & Tweedie,

2000a, 2000b) were employed to detect publication bias for all measures of the realism of confidence.

Method

In conducting this meta-analysis, we followed the guidelines proposed by Field and Gillett (2010), Rosenthal and DiMatteo (2001), and Viechtbauer (2010).

Selection of Studies

The studies had to meet four criteria to be included in the meta-analysis. (a) An objective criterion had to be available to compute the accuracy of witnesses' reports. We therefore included only studies that used videos, slides, or stagings depicting criminal behavior and excluded analyses of real-world cases. (b) Studies had to employ recognition questions (e.g., two-alternative forced-choice questions, yes/no or true/false questions, multiple-choice questions with more than two answer alternatives, etc.) about the crime event. Studies using free or cued recall questions were excluded because different cognitive processes are assumed to determine performance in recall and recognition memory (cf. Allwood, Innes-Ker, Homgren, & Fredin, 2008). (c) Studies had to employ adult samples because there are special problems associated with child witnesses (cf. Allwood et al., 2008; Buratti, Allwood, & Johansson, 2014). Because of potential confounds with decreasing or impaired memory, studies using samples of elderly adults were also excluded (Dahl, Allwood, Scimone, & Rennemark, 2015). (d) We included only studies that computed at least one of the previously mentioned three measures of the realism of confidence (calibration, under-/overconfidence, or resolution). If information necessary to interpret the coefficients or to perform the analyses was missing, we contacted the authors of the respective articles. Studies for which the required information was unobtainable were excluded from the meta-analysis.

The literature search was performed in several steps. We first conducted keyword searches in the Web of Science Core Collection, PsychINFO, Dissertation Abstracts, and GoogleScholar databases using the terms "eyewitness memory," "confidence," "certainty,"

"confidence-accuracy relationship," "realism," and "calibration." We then conducted forward and backward searches for all articles identified as suitable for the meta-analysis. Moreover, we conducted forward searches for three core articles on calibration analyses that together have been cited over 3,000 times according to the GoogleScholar database (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982; Yates, 1994). After applying all selection criteria, we identified as relevant for our meta-analysis eight studies with 24 independent study units and 803 participants and published in the English language in scientific journals or books.

Computation of Effect Size Estimates

We did not aim to meta-analyze tests of differences between conditions. Rather, we wanted to examine the relationship between two target variables (confidence and accuracy) and potential moderators of this relationship. Therefore, we treated experimental conditions as separate study units (cf. Borenstein, Hedges, Higgins, & Rothstein, 2010; Olsson & Juslin, 2002). To guarantee the independence of effect sizes, we collapsed the cells in within-subjects designs and computed mean effect sizes across the within-subjects conditions. This resulted in 24 study units that could be used in the meta-analysis. Table 1 summarizes these studies and shows all study units included in the meta-analysis and their respective core study characteristics.

Because measures of the realism of confidence are usually computed on a sufficiently large number of observations in eyewitness event memory studies (see Table 1, column 9 "Items"), we treated them as interval scaled variables (Brewer & Wells, 2006) and standardized the raw means by their standard deviations and sample sizes to use them as effect size measures for *C*, *U/O*, and *NRI* (Shadish & Haddock, 1994; Viechtbauer, 2010). We also standardized the control variables accuracy and confidence by their standard deviations and sample sizes to obtain effect size measures.

- Place Table 1 about here -

Results

We computed fixed-effect models to determine the mean effect size estimates. We calculated 95% confidence intervals for all dependent variables and used *z*-tests to establish whether the mean effect size estimates differed significantly from 0. *Q*-tests were computed to test whether the effect sizes were heterogeneous. Because the results of weighted (using 1/variance; Borenstein et al., 2010; Viechtbauer, 2010) and unweighted mean effect size analyses did not differ, we report only the results of the analyses that employed the weighted effect sizes. As fixed-effect models have been found to produce false-positive errors more often than random-effects models (e.g., Hunter & Schmidt, 2000; Ioannidis, 2008), we applied random-effects models for control purposes. The pattern of results did not change as a result of choosing either a fixed-effect or random-effects model, however. In particular, all effect sizes that were significant when we used a fixed-effect model were also significant when we used a random-effects model. Therefore, we report only the results from the fixed-effect model analyses. All analyses were performed using the package metafor (Viechtbauer, 2010) in the R statistics environment (R Development Core Team, 2015).

Overall Mean Effect Sizes

To place all assessments of accuracy and confidence on the same scale, we linearly rescaled all measures to a common scale ranging from 0 to 1 when they were not already scaled in this manner. All study units employed questions with two answer options. Respondents answering randomly could therefore be expected to achieve an accuracy of .5. Accuracy and confidence were reported for 23 of the 24 study units included in the meta-analysis. The weighted mean accuracy was .647 (*SE* = .003), 95% *CI* [.641, .652]. The 95% *CI* did not include .5; the accuracy was thus shown to significantly surpass the level of chance. The weighted mean confidence was .744 (*SE* = .002), 95% *CI* [.739, .748].

Calibration coefficients were available for 21 of the 24 study units. The meta-analysis of calibration coefficients resulted in an effect size estimate larger than zero, thereby revealing a significant miscalibration of witnesses, $C = .048$ ($SE = .001$), 95% CI [.046, .051], $z = 44.689$, $p < .001$. The calibration effect sizes showed significant heterogeneity, $Q(20) = 157.812$, $p < .001$, suggesting that moderator variables may have influenced the calibration coefficients.

Effect sizes for under-/overconfidence were provided for 20 of the 24 study units. Their meta-analysis revealed that witnesses significantly overestimated their accuracy, $U/O = .095$ (SE = .004), 95% CI [.088, .102], $z = 27.068$, $p < .001$. The test for heterogeneity suggested the potential existence of moderator variables that influenced mean overconfidence, $Q(19) = 118.605$, $p < .001$.

Measures of resolution were available for 14 of the 24 study units. The mean resolution significantly exceeded zero, $NRI = .025$ ($SE = .001$), 95% CI [.024, .026], $z = 37.733$, $p < .001$, indicating that witnesses' confidence ratings could be used to distinguish between correct and incorrect responses with a probability that exceeded chance. However, the effect was small in magnitude, and again, we found significant heterogeneity, which implied that the resolution may have been influenced by moderator variables, $Q(13) = 121.512$, $p < .001$.

Moderator Analyses

Our analysis revealed that in six of the eight articles, a video first used by Granhag (1997) was presented as the stimulus material. Results for 17 of the 24 study units were therefore based on the same simulated crime event. Moreover, in all study units employing the video by Granhag (1997), two-alternative forced-choice (2AFC) items were used to assess witness testimony, whereas in all other study units, true/false (T/F) items were employed. We computed fixed-effect models with moderators to test whether the materials and procedures that had been employed (Granhag video and 2AFC items vs. other stimuli and T/F items) had

induced systematic variance into the measures of the realism of confidence. To test the influence of the moderator variables for significance, we computed *Q*-tests.

Meta-analytical effect size estimates were also determined for the two control variables accuracy and confidence. Employing a fixed-effect model for accuracy ($k = 23$) resulted in an effect size estimate that was significantly lower for the Granhag video with 2AFC items (accuracy = .631, $SE = .003$; 95% *CI* [.626, .637]) than for the other stimulus materials with T/F items (accuracy = .718, $SE = .006$; 95% *CI* [.706, .731]), $Q(2) = 60492.599$, $p < .001$. A fixed-effect model with moderators for confidence ($k = 23$) revealed that confidence was also significantly lower for the Granhag video with the 2AFC items (confidence = .728, $SE = .003$; 95% *CI* [.723, .733]) compared with the other stimulus materials with the T/F items (confidence = .792, $SE = .005$; 95% *CI* [.783, .801]), $Q(2) = 104314.592$, $p < .001$. Thus, we found that for confidence and accuracy, there were systematic differences between the responses to the 2AFC items after viewing the Granhag video and the responses to the T/F items after viewing any other stimulus materials. Participants who viewed the Granhag video and completed the 2AFC items performed significantly worse and were significantly less confident in their responses than participants who were presented with other stimulus material and completed T/F items.

A similar result pattern emerged for all three measures of the realism of confidence. A fixed-effect model with moderators revealed that the people who participated in study units that employed the video by Granhag with 2AFC items were significantly more miscalibrated, $C = .050$ ($SE = .001$), 95% *CI* [.048, .053], $z = 39.234$, $p < .001$, than people who participated in study units that used other stimulus materials with T/F items, $C = .044$ ($SE = .002$), 95% *CI* [.040, .048], $z = 21.564$, $p < .001$, $Q(2) = 2004.304$, $p < .001$. However, differences in stimulus material and item type accounted for only some of the variance in the calibration data as the residual variance was still significantly heterogeneous, $Q(19) = 150.597$, $p < .001$.

Similar to calibration, the fixed-effect model with moderators for overconfidence (k = 20) indicated that participants were significantly more overconfident in their responses to the 2AFC items about the Granhag video, *U/O* = .105 (*SE* = .004), 95% *CI* [.097, .113], z = 26.016, $p < .001$, than participants who responded to the T/F items about the other stimulus materials, *U/O* = .064 (*SE* = .007), 95% *CI* [.050, .077], $z = 9.064$, $p < .001$, $Q(2) = 758.999$, $p < .001$. Overconfidence for the 2AFC items about the Granhag video was almost twice as large as overconfidence for T/F items about the other stimulus materials. Due to the complete confounding of these two variables, we could not determine whether this effect resulted from differences in stimulus material or item type. Moreover, the residual variance was still significantly heterogeneous, $Q(18) = 92.290$, $p < .001$, implying that other moderators also potentially influenced the magnitude of overconfidence in the data.

A fixed-effect model with moderators for resolution (k = 14) revealed that the people who participated in the study units that presented the Granhag video and 2AFC items were significantly more accurate in distinguishing between correct and incorrect responses, *NRI* = .033 (*SE* = .001), 95% *CI* [.031, .036], $z = 28.665$, $p < .001$, than the participants of studies that employed other stimulus materials and T/F items, *NRI* = .021 (*SE* = .001), 95% *CI* [.020, .023], $z = 26.032$, $p < .001$, $Q(2) = 1499.374$, $p < .001$. Regardless of stimulus material and item type, the resolution was significantly different from zero but small in effect size, suggesting that witnesses generally had trouble distinguishing between correct and incorrect responses. The residual heterogeneity was large and significant, $Q(12) = 45.913$, $p < .001$. Thus, unidentified moderator variables also could have influenced the magnitude of the resolution effects.

To summarize, the moderator analyses revealed that the study method in terms of stimulus material and item type significantly influenced the accuracy, confidence, calibration, under-/overconfidence, and resolution. To disentangle the pervasive confounding of stimulus

material and item type that makes the present moderator analyses difficult to interpret, future studies should manipulate stimulus material and item type independently.

Publication Bias Analyses

For all measures of the realism of confidence, we created funnel plots (Figures 1 to 3). In line with recommendations, we present the effect size estimates on the x-axis and their standard errors on the y-axis (Sterne & Egger, 2001). To test the significance of the funnel plot asymmetry, we computed Begg's rank-correlation test and Egger's regression method. When significant funnel plot asymmetry occurred, we computed corrected effect size estimates using the trim-and-fill method.

Visual inspection of the funnel plot for calibration effect sizes indicated an asymmetry and, thus, an overestimation of the calibration index (Figure 1). This asymmetry was significant according to the rank-correlation test, $\tau = .413$, $p = .009$, and Egger's regression test, $z = 7.641$, $p < .001$. The trim-and-fill method suggested the imputation of nine effects to the left side and a corrected effect size estimate of $C = .044$ ($SE = .001$), 95% *CI* [.042, .046], that, however, was still significantly different from 0, $z = 45.344$, $p < .001$, for $k = 30$, as compared with a mean effect size estimate of $C = .048$ with no effect size imputation.

- Place Figure 1 about here -

The funnel plot for under-/overconfidence showed a rather symmetrical distribution of effect sizes (Figure 2). Accordingly, neither the rank-correlation test, $\tau = .137$, $p = .422$, nor the more powerful regression test, $z = 1.217$, $p = .224$, identified a significant funnel plot asymmetry, suggesting that no publication bias occurred for under-/overconfidence.

- Place Figure 2 about here -

The funnel plot for resolution showed the strongest asymmetry in the effect sizes of the three measures of the realism of confidence (Figure 3). As was to be expected on the basis of the visual inspection, both the rank-correlation test, $\tau = .486$, $p < .001$, and the regression test, $z = 8.300$, $p < .001$, indicated significant asymmetry and, thus, a strong publication bias in the data for resolution. The trim-and-fill method suggested the addition of six effect sizes to the left side, resulting in a slightly decreased mean effect size for resolution ($k = 20$), *NRI* = .021 (*SE* = .001), 95% *CI* [.020, .023], that nevertheless still significantly differed from 0, $z = 35.716$, $p < .001$, as compared with a mean effect size estimate for the resolution of *NRI* = .025 with no effect size imputation.

- Place Figure 3 about here -

Summing up the results for the publication bias analyses, all tests conducted to detect a bias in the data suggested significant asymmetry in the effect sizes for calibration and resolution. No statistical evidence for a publication bias in the under-/overconfidence data was found. Applying the trim-and-fill method resulted in slightly reduced effect size estimates for calibration (.044 instead of .048) and resolution (.021 instead of .025). Even though a publication bias was found, its impact on the mean effect size estimates was rather small.

Discussion

We conducted what we believe to be the first meta-analysis of studies investigating the realism of confidence in eyewitness recognition memory for events. Eight studies providing 24 independent assessments of the realism of confidence in 803 witnesses were included. We found that witnesses' confidence deviated only slightly but significantly from a perfect calibration curve. Witnesses tended to be overconfident; that is, they overestimated the accuracy of their recognition judgments. The resolution of witnesses' confidence ratings was found to be low. This means that in the studies included in the meta-analysis, witnesses did

not give consistently higher confidence ratings for their correct responses than they did for their incorrect responses.

The results of this meta-analysis suggest that for the accuracy of their accounts of observed events, eyewitnesses' confidence might be a more valid predictor than for the accuracy of their identification decisions. In a meta-analysis of the realism of confidence in eye- and earwitness identifications (Olsson & Juslin, 2002), some calibration coefficients approached .40 (0 implies perfect calibration and 1 implies maximum miscalibration). By contrast, in the present study, the highest observed calibration coefficient was .11 (Sarwar et al., 2014). The overall mean effect size estimate for calibration in the present meta-analysis was .048. The deviation between the observed and the perfect calibration curve was rather small but was statistically significant. Correcting for publication bias resulted in an even smaller mean effect size estimate for calibration (.044). The mean effect size estimate for overconfidence in the meta-analysis on eyewitness and earwitness identifications (Olsson & Juslin, 2002) was located between 8 and 43% (depending on the scale format that was used). A mean level of overconfidence of around 9% for event recognition memory observed in the present meta-analysis was thus close to the lower margin of overconfidence for eyewitness and earwitness identifications. Nevertheless, both Olsson and Juslin's (2002) meta-analysis on identification performance and the present meta-analysis on event memory found that witnesses overestimated their accuracy, thus supporting the notion that overconfidence is a rather ubiquitous phenomenon when people make decisions under uncertainty. In this vein, Goodman-Delahunty, Granhag, Hartwig, and Loftus (2010) also found overconfidence in lawyers' predictions of case outcomes, and Towfigh and Glöckner (2011) found people betting on sports to overestimate their likelihood of winning. Attempts to reduce or even eliminate overconfidence (e.g., through warnings; García-Bajos & Migueles, 2003) have often been found to be ineffective. It therefore seems reasonable to expect at least some

overconfidence when witnesses are asked to express their confidence in the accuracy of their testimonies.

The pattern of results regarding calibration and resolution was at odds with previous findings from studies that computed between-subjects and within-subjects correlation coefficients. In the present meta-analysis, the witnesses were rather well-calibrated, and their confidence ratings were low in resolution according to the classification of effect sizes for η^2 (Cohen, 1988), to which the interpretation of the Normalized Residual Index (*NRI*) is analogous (cf. Baranski & Petrusic, 1994). Empirical studies have commonly found low and unstable between-subjects correlations (between mean confidence and accuracy computed across witnesses; cf. Perfect, 2002) and medium and more stable within-subjects correlations (between responses and confidence ratings computed across all responses of individual witnesses; e.g., Bulevich & Thomas, 2012; Kebbell et al., 2010; Kebbell & Giles, 2000). Whereas between-subjects correlation coefficients can be interpreted as paralleling calibration coefficients because they compare overall levels of confidence and accuracy, within-subjects correlation coefficients have previously been interpreted as an alternative measure of resolution (cf. Higham, Luna, & Bloomfield, 2011). The magnitude of between-subjects correlation coefficients depends on the variance in mean confidence ratings and accuracies in study samples. Several studies have suggested that variances in student samples that are commonly used in experiments for investigating eyewitness memory might underestimate the actual variance in witnesses' competences. Between-subjects correlations might thereby underestimate the true relationship between confidence and accuracy (cf. Lindsay et al., 1998, 2000). This notion is supported by the present meta-analytical result that the witnesses were only slightly miscalibrated. Because a calibration coefficient is less vulnerable to skewed distributions or to restrictions in the variances of confidence or accuracy (cf. Juslin et al., 1996), this measure should generally be preferred over between-subjects correlation coefficients. Similarly, the magnitude of within-subjects correlation coefficients depends on

the distribution of individual confidence ratings and item difficulties. Because eyewitness studies rarely provide information about item difficulties and individual confidence ratings, it is difficult to interpret within-subjects correlation coefficients. Strong within-subjects correlations might occur, for example, when researchers aim to generate particularly difficult items (cf. Juslin, 1993, 1994) because this is likely to lead to artificially inflated variances in item difficulties and confidence ratings. Like the calibration coefficient, the *NRI* is less vulnerable to skewed distributions or extreme variances in confidence and accuracy and might therefore provide a more realistic measure of resolution. The finding from the present meta-analysis that witnesses are quite capable of estimating their overall accuracy but less able to distinguish between correct and incorrect responses might therefore be more realistic than the opposite pattern observed in studies using correlation analyses. The observed combination of good calibration and poor resolution is in line with previous studies that have suggested that the conditions under which confidence calibration is maximized might not necessarily benefit confidence resolution (cf. Keren 1991; Yates, 1982).

Of the eight studies included in this meta-analysis, we found six (i.e., 17 of the 24 study units) that used the same stimulus material (a video first used by Granhag, 1997). Moreover, the stimulus material was perfectly confounded with item type: All studies using the Granhag (1997) video employed two-alternative forced-choice (2AFC) items, whereas the two studies that presented different material employed true/false (T/F) items. A moderator analysis revealed a significant influence of stimulus material and item type on accuracy, confidence, and the three measures of the realism of confidence. Due to the confounding of stimulus material and item type, it was not possible to determine whether the variation in stimulus material, item type, or both affected accuracy, confidence, and the realism of confidence. For this reason, the mean effect size estimates need to be interpreted with caution; it is impossible to tell whether the present results would be likely to be replicated if other materials are used.

We found a significant publication bias in the effect sizes for calibration and resolution. It is somewhat surprising that miscalibration was overestimated due to publication bias. We propose the following explanation for this finding. In most of the studies included in the present meta-analysis, experimental manipulations were employed to investigate their effects on the realism of confidence. Calibration and resolution effect sizes are measured on a scale from 0 to 1, and mean effect size estimates for the two were at the floor level of .048 and .025, respectively. This made it difficult to detect a statistically significant impact of any experimental manipulation. If studies fail to detect the effect of an experimental manipulation because calibration or resolution approaches zero, studies with larger coefficients for calibration and resolution are more likely to be published because they are more likely to produce significant effects of experimental manipulations. This interpretation receives support from the finding that for under-/overconfidence coefficients ranging between -1 and +1, no publication bias was found.

We investigated publication bias using a methodology that was based on funnel plot asymmetry. However, asymmetry in effect size distributions might also be caused by true heterogeneity in the data (Ioannidis, 2008; Terrin, Schmid, Lau, & Olkin, 2003). In the present study, such heterogeneity might, for example, have resulted from experimental manipulations that impacted measures of the realism of confidence. Whether funnel plot asymmetry was due to publication bias or true heterogeneity cannot be decided on the basis of the present data. Moreover, all of the publication bias analyses examining the symmetry of the effect size distributions were based on the assumption that the mean effect size estimate approximates the true effect size (Sutton, 2009). If the mean effect size measure is itself biased, the publication bias might be under- or overestimated.

Confidence ratings influence legal decision-makers' perceptions of eyewitness testimony (Brewer, 2006; Kassin et al., 2001; Simons & Chabris, 2011; Wise, et al., 2009; Wise & Safer, 2004). Recent empirical findings suggest that witnesses' confidence calibration

might influence their perceived credibility even more than the subjective certainty they express in their confidence ratings (Tenney, MacCoun, Spellman, & Hastie, 2007). Understanding the relation between confidence and accuracy, and in particular, understanding the realism of confidence is therefore of central importance to legal justice. In the present meta-analysis, we summarized findings on the realism of confidence measured in terms of confidence calibration, under-/overconfidence, and confidence resolution in eyewitness event recognition memory. Confidence judgments appeared to be more valid predictors of accuracy than previous correlation analyses have suggested. Eyewitnesses were found to have a rather good grasp of their overall accuracy but tended to overestimate their performance to a small degree. However, due to the low level of diversity in the stimulus materials and the confounding of material with item type, the generalizability of our results must be considered limited. As researchers have previously pointed out, a lack of diversity in the stimulus materials employed in forensic psychological studies sets narrow limits on the generalizability of the research findings (cf. Memon, Mastroberardino, & Fraser, 2008). Future studies on the realism of confidence in eyewitness event recognition memory should employ a wider range of stimuli and methods to provide findings that can be better generalized to other crimes or persons than the ones from the rather small set of studies that qualified for inclusion in the present meta-analysis. Conducting additional research with more diverse stimulus materials is necessary to determine whether the present positive results on confidence calibration in eyewitness memory for events are replicable and whether they are more than an artifact resulting from the repeated use of a potentially unrepresentative stimulus.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Allwood, C. M., Granhag, P. A., & Johansson, M. (2003). Increased realism in eyewitness confidence judgements: The effect of dyadic collaboration. *Applied Cognitive Psychology, 17*, 545-561. doi: 10.1002/acp.888

*Allwood, C. M., Innes-Ker, A., Homgren, J., & Fredin, G. (2008). Children's and adults' realism in their event-recall confidence in responses to free recall and focused questions. *Psychology, Crime & Law, 14*, 529-547. doi: 10.1080/10683160801961231

*Allwood, C. M., Knutsson, J., & Granhag, P. A. (2006). Eyewitnesses under influence: How feedback affects the realism in confidence judgements. *Psychology, Crime & Law, 12*, 25-38. doi: 10.1080/10683160512331316316

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55*, 412-428. doi: 10.3758/bf03205299

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088-1101. doi: 10.2307/2533446

*Bonham, A. J., & Gonzalez-Vallejo, C. (2009). Assessment of calibration for reconstructed eyewitness memories. *Acta Psychologica, 131*, 34-52. doi: 10.1016/j.actpsy.2009.02.008

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97-111. doi: 10.1002/jrsm.12

Bothwell, R. K., Brigham, J. C., & Deffenbacher, K. A. (1987). Correlation of eyewitness accuracy and confidence. Optimality hypothesis revisited. *Journal of Applied Psychology, 72*, 691-695. doi: 10.1037/0021-9010.72.4.691

Brewer, N. (2006). Uses and abuses of eyewitness identification confidence. *Legal and Criminological Psychology, 11*, 3-23. doi: 10.1348/135532505x79672

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*, 11-30. doi: 10.1037/1076-898X.12.1.11

Bulevich, J. B., & Thomas, A. K. (2012). Retrieval effort improves memory and metamemory in the face of misinformation. *Journal of Memory and Language, 67*(1), 45-58. doi: 10.1016/J.Jml.2011.12.012

*Buratti, S., & Allwood, C. M. (2012). The accuracy of meta-metacognitive judgments: Regulating the realism of confidence. *Cognitive Processes, 13*, 243-253. doi: 10.1007/s10339-012-0440-5

Buratti, S., Allwood, C. M., & Johansson, M. (2014). Stability in the metamemory realism of eyewitness confidence judgments. *Cognitive Processes, 15*, 39-53. doi: 10.1007/s10339-013-0576-y

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cutler, B. L., & Penrod, S. (1989). Moderators of the confidence-accuracy correlation in face recognition: The role of information processing ad base-rates. *Applied Cognitive Psychology, 3*, 95-107. doi: 10.1002/acp.2350030202

Dahl, M., Allwood, C. M., Scimone, B., & Rennemark, M. (2015). Old and very old adults as witnesses: Event memory and metamemory. *Psychology, Crime & Law, 21*, 764-775. doi: 10.1080/1068316X.2015.1038266

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

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89-98. doi: 10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias. *Biometrics, 56*, 455-463. doi: 10.1111/j.0006-341X.2000.00455.x

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634. doi: 10.1136/bmj.315.7109.629

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*, 665-694.

Fisher, R. P., Geiselman, R. E., & Raymond, D. S. (1987). Critical analysis of police interview techniques. *Journal of Police Science and Administration, 15*, 177-185.

García-Bajos, E., & Migueles, M. (2003). False memories for script actions in a mugging account. *European Journal of Cognitive Psychology, 15*, 195-208. doi: 10.1080/09541440244000102

George, R., & Clifford, B. (1992). Making the most of witnesses. *Policing*, 185-198.

Goodman-Delahunty, J., Granhag, P. A., Hartwig, M., & Loftus, E. F. (2010). Insightful or wishful: Lawyers' ability to predict case outcomes. *Psychology, Public Policy, and Law, 16*, 133-157. doi: 10.1037/a0019060

Granhag, P. A. (1997). Realism in eyewitness confidence as a function of type of event witnessed and repeated recall. *Journal of Applied Psychology, 82*, 599-613. doi: 10.1037/0021-9010.82.4.599

*Granhag, P. A., Jonsson, A. C., & Allwood, C. M. (2004). The Cognitive Interview and its effect on witnesses' confidence. *Psychology, Crime & Law, 10*, 37-52. doi: 10.1080/1068316021000030577

Hollins, T. S., & Perfect, T. J. (1997). The confidence-accuracy relation in eyewitness event memory: The mixed question type effect. *Legal and Criminological Psychology, 2*, 205-218. doi: 10.1111/j.2044-8333.1997.tb00344.x

Higham, P. A., Luna, K., & Bloomfield, J. (2011). Trace-strength and source-monitoring accounts of accuracy and metacognitive resolution in the misinformation paradigm. *Applied Cognitive Psychology, 25*, 324-335. doi: 10.1002/Acp.1694

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275-292. doi: 10.1111/1468-2389.00156

Ioannidis, J. P. A. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice, 14*, 951-957. doi: 10.1111/j.1365-2753.2008.00986.x

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

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research. A new survey of the experts. *American Psychologist, 56*, 405-416. doi: 10.1037//0003-066x.56.5.405

Kebbell, M. R., Evans, L., & Johnson, S. D. (2010). The influence of lawyers' questions on witness accuracy, confidence, and reaction times and on mock jurors' interpretation of witness accuracy. *Journal of Investigative Psychology and Offender Profiling, 7*, 261-271. doi: 10.1002/jip.125

Kebbell, M. R., & Giles, D. C. (2000). Some experimental influences of lawyers' complicated questions on eyewitness confidence and accuracy. *Journal of Psychology, 134*, 129-139. doi: 10.1080/00223980009600855

Kebbell, M. R., Milne, R., & Wagstaff, G. F. (1999). The cognitive interview: A survey of its forensic effectiveness. *Psychology, Crime and Law, 5*, 101-115. doi: 10.1080/10683169908414996

Kebbell, M. R., Wagstaff, G. F., & Covey, J. A. (1996). The influence of item difficulty on the relationship between eyewitness confidence and accuracy. *British Journal of Psychology, 87*, 653-662. doi: 10.1111/j.2044-8295.1996.tb02614.x

Keren, G. (1991). Calibration and probability judgments. Conceptual and methodological issues. *Acta Psychologica, 77*, 217-273. doi: 10.1016/0001-6918(91)90036-y

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159-183. doi: 10.1016/0030-5073(77)90001-0

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty* (pp. 275-324). Cambridge: Cambridge University Press.

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior, 24*, 685-697. doi: 10.1023/A:1005504320565

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science, 9*, 215-218. doi: 10.1111/1467-9280.00041

*Luna, K., & Martín-Luengo, B. (2010). New advances in the study of the confidence-accuracy relationship in the memory for events. *European Journal of Psychology Applied to Legal Context, 2*, 55-71. Retrieved from: http://dialnet.unirioja.es/servlet/articulo?codigo=3112506&orden=408554&info=link

Luna, K., & Martín-Luengo, B. (2012). Confidence-accuracy calibration with general knowledge and eyewitness memory cued recall questions. *Applied Cognitive Psychology, 26*, 289-295. doi: 10.1002/acp.1822

Memon, A., Mastroberardino, S., & Fraser, J. (2008). Münsterberg's legacy: What does eyewitness research tell us about the reliability of eyewitness testimony? *Applied Cognitive Psychology, 22*, 841-851. doi: 10.1002/acp.1487

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109-133. doi: 10.1037/0033-2909.95.1.109

Odinot, G., Wolters, G., & van Koppen, P. J. (2009). Eyewitness memory of a supermarket robbery: A case study of accuracy and confidence after 3 months. *Law and Human Behavior, 33*, 506-514. doi: 10.1007/s10979-008-9152-x

Olsson, N. (2000). A comparison of correlation, calibration, and diagnosticity as measures of the confidence-accuracy relationship in witness identification. *Journal of Applied Psychology, 85*, 504-511. doi: 10.1037//0021-9010.85.4.504

Olsson, N., & Juslin, P. (2002). Calibration of confidence among eyewitnesses and earwitnesses. In P. Chambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition. Process, function and use* (pp. 203-218). New York: Springer.

Perfect, T. J. (2002). When does eyewitness confidence predict performance? In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 95-120). Cambridge: Cambridge University Press.

Perfect, T. J., & Hollins, T. S. (1996). Predictive feeling of knowing judgements and postdictive confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology, 10*, 371-382. doi: 10.1002/(sici)1099-0720(199610)10:5<371::aid-acp389>3.0.co;2-o

Peterson, C., & Grant, M. (2001). Forced-choice: Are forensic interviewers asking the right questions? *Canadian Journal of Behavioural Science (Revue Canadienne Des Sciences Du Comportement), 33*, 118-127. doi: 10.1037/h0087134

R Development Core Team. (2014). The R-project for statistical computing. Retrieved from http://www.r-project.org/

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence-accuracy correlation. *Journal of Applied Psychology, 81*, 587-594. doi: 10.1037/0021-9010.81.5.587

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82. doi: 10.1146/annurev.psych.52.1.59

*Sarwar, F., Allwood, C. M., & Innes-Ker, A. (2014). Effects of different types of forensic information on eyewitness' memory and confidence accuracy. *European Journal of Psychology Applied to Legal Context, 6*, 17-27. doi: 10.5093/ejpalc2014a3

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis.* (pp. 261-281). New York: Russell Sage Foundation.

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the U.S. population. *Plos One, 6*, 1-7. doi: 10.1371/journal.pone.0022757

Smith, V. L., Ellsworth, P. C., & Kassin, S. M. (1989). Eyewitness accuracy and confidence: Within- versus between-subjects correlations. *Journal of Applied Psychology, 74*, 356-359. doi: 10.1037/0021-9010.74.2.356

Sporer, S. L. (1996). Psychological aspects of person descriptions. In S. L. Sporer, R. S. Malpass, & G. Köhnken (Eds.), *Psychological issues in eyewitness identification* (pp. 53-86). Mahwah, NJ: Lawrence Erlbaum Associates.

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. doi: 10.1037//0033-2909.118.3.315

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54*, 1046-1055. doi: 10.1016/S0895-4356(01)00377-8

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 435-452). New York: Russell Sage Foundation.

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science, 18*, 46-50. doi: 10.1111/j.1467-9280.2007.01847.x

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Staitstics in Medicine, 22*, 2113-2126. doi: 10.1002/sim.1461

Towfigh, E. V., & Glöckner, A. (2011). GAME OVER: Empirical support for soccer bets regulation. *Psychology, Public Policy, and Law, 17*, 475-506. doi: 10.1037/a0023402

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1-48. doi: 10.18637/jss.v036.i03

Wagenaar, W. A. (1988). Calibration and the effects of knowledge and reconstruction in retrieval from memory. *Cognition, 28*, 277-296. doi: 10.1016/0010-0277(88)90016-9

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence. Improving its probative value. *Psychological Science in the Public Interest, 7*, 45-75. doi: 10.1111/j.1529-1006.2006.00027.x

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: Cambridge University Press.

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54*, 277-295. doi: 10.1146/annurev.psych.54.101601.145028

Wheatcroft, J. M., Kebbell, M. R., & Wagstaff, G. F. (2001). Courtroom questioning. *Forensic Update, 65*, 20-25.

Wise, R. A., Pawlenko, N. B., Safer, M. A., & Meyer, D. (2009). What US prosecutors and defence attorneys know and believe about eyewitness testimony. *Applied Cognitive Psychology, 23*, 1266-1281. doi: 10.1002/acp.1530

Wise, R. A., & Safer, M. A. (2004). What US judges know and believe about eyewitness testimony. *Applied Cognitive Psychology, 18*, 427-443. doi: 10.1002/acp.993

Wright, A. M., & Alison, L. (2004). Questioning sequences in Canadian police interviews: Constructing and confirming the course of events? *Psychology, Crime & Law, 10*, 137-154. doi: 10.1080/1068316031000099120

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*, 611-617. doi: 10.1037/0033-2909.110.3.611

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30*, 132-156. doi: 10.1016/0030-5073(82)90237-9

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 381-410). Chichester: Wiley.

Footnotes

1) http://www.innocenceproject.org/causes-wrongful-conviction, retrieved on October 26, 2015

Table 1

List of the Studies Included in the Meta-Analysis Split into 24 Separate Study Units.

No.	Authors	*N*	Exp.	Condition	Manipulation	Design	Stimulus	Items	*Accuracy*	*Confidence*	*Measures of the realism of confidence*		
											Calibration	*Under-/overconfidence*	*Resolution*
1	Allwood, Granhag, & Johansson (2003)	40	1	1	Joint answer and confidence rating	BS	VG	45 2AFC	.66 (.06)	.77 (.05)	.05 (.03)	.11 (.07)	.04 (.02)
2		40	1	2	Pooled across individual responses with individual confidence ratings and joint answers with joint confidence ratings	BS / WS	VG	45 2AFC	.63 (.06)	.73 (.04)	.04 (.02)	.10 (.06)	.04 (.02)
3		22	2	1	Pooled across individual responses with individual confidence ratings and repeated individual responses with repeated individual confidence ratings	WS	VG	44 2AFC	.59 (.06)	.71 (.06)	.05 (.03)	.12 (.07)	.04 (.02)
4	Allwood, Innes-Ker, Homgren, & Fredin (2008)	38	2		No manipulation		VG	44 2AFC	.54 (.07)	.72 (.09)		.18 (.12)	
5	Allwood, Knutsson, & Granhag (2006)	29	1	1	No manipulation	BS	VG	44 2AFC	.63 (.06)	.75 (.07)		.12 (.09)	.03 (.02)
6		31		2	Pooled across confirmatory and disconfirmatory feedback		VG	44 2AFC	.63 (.07)	.77 (.09)		.13 (.12)	.04 (.02)
7	Bonham & González-Vallejo (2009)	35	1	1	No manipulation	BS	Other video	66 T/F	.78 (.08)	.79 (.07)	.03 (.02)	.01 (.08)	.02 (.01)
8		35	1	2	Pooled across types of misinformation	BS	Other video	66 T/F	.70 (.09)	.80 (.07)	.05 (.04)	.10 (.10)	.02 (.01)
9		32	2	1	Control narrative	BS	Other video	66 T/F	.68 (.08)	.76 (.08)	.05 (.03)	.08 (.11)	.03 (.02)
10		32	2	2	Consistent narrative	BS	Other video	66 T/F	.75 (.09)	.81 (.05)	.04 (.03)	.06 (.09)	.03 (.02)
11		32	2	3	Misinformation narrative	BS	Other video	66 T/F	.70 (.10)	.81 (.07)	.06 (.05)	.11 (.13)	.02 (.01)
12		35	2	4	General narrative	BS	Other video	66 T/F	.68 (.10)	.76 (.07)	.05 (.03)	.08 (.11)	.02 (.01)

13		34	1	1	Pooled across control condition and exclusion task	BS / WS	VG	50 2AFC	.76 (.06)	.82 (.08)	.05 (.03)		
14	Buratti & Allwood (2012)	35	1	2	Pooled across control condition with realism and exclusion task	BS / WS	VG	50 2AFC	.75 (.08)	.82 (.09)	.05 (.03)		
15		66	2	1	Pooled across control condition and exclusion task	WS	VG	50 2AFC	.77 (.08)	.88 (.07)	.04 (.03)		
16		26		1	No manipulation	BS	VG	45 2AFC	.55 (.09)	.64 (.06)	.05 (.03)	.09 (.09)	.02 (.02)
17	Granhag, Jonsson, & Allwood (2004)	26		2	First cognitive interview, then recognition judgments; pooled across questions mentioned and questions not mentioned in interview	BS / WS	VG	45 2AFC	.54 (.06)	.69 (.06)	.07 (.04)	.15 (.08)	.02 (.02)
18		27		3	First standard interview, then recognition judgments; pooled across questions mentioned and questions not mentioned in interview	BS / WS	VG	45 2AFC	.55 (.09)	.68 (.07)	.06 (.03)	.13 (.09)	.03 (.03)
19	Luna & Martín-Luengo (2010)	37	1				VG	24 T/F			.10 (.06)		
20		22	1	1	No manipulation	BS	VG	44 2AFC	.54 (.07)	.62 (.06)	.05 (.03)	.08 (.09)	
21		16	1	2	Discussions in lab after presentation of video and before providing recognition judgments	BS	VG	44 2AFC	.53 (.07)	.63 (.04)	.05 (.02)	.09 (.08)	
22	Sarwar, Allwood, & Innes-Ker (2014)	17	1	3	Discussions with family after presentation of video and before providing recognition judgments	BS	VG	44 2AFC	.55 (.07)	.66 (.07)	.05 (.03)	.10 (.08)	
23		19	1	4	Retelling in lab after presentation of video and before providing recognition judgments	BS	VG	44 2AFC	.56 (.06)	.64 (.07)	.05 (.02)	.08 (.07)	
24		77	2	1	Pooled across central and peripheral details		VG	63 2AFC	.62 (.09)	.64 (.11)	.11 (.06)	.03 (.13)	

Note. Column 6 contains descriptions of the manipulations applied to each study unit. Columns 7-9 display the core study characteristics. Columns 10-14 show the means and standard deviations of the control variables (accuracy and confidence) and the measures of the realism of confidence (calibration, under-/overconfidence, resolution) included in the meta-analysis. BS = between-subjects design; WS = within-subjects design; BS / WS = a between-subjects design that contained an additional within-subjects manipulation; 2AFC = two-alternative forced-choice items; T/F = true-false items; *C* = calibration index; *U/O* = under-/overconfidence index; *NRI* = normalized resolution index.

Figure 1. Funnel plot for calibration with observed (black dots) and imputed (white dots) effect sizes plotted against their standard errors.

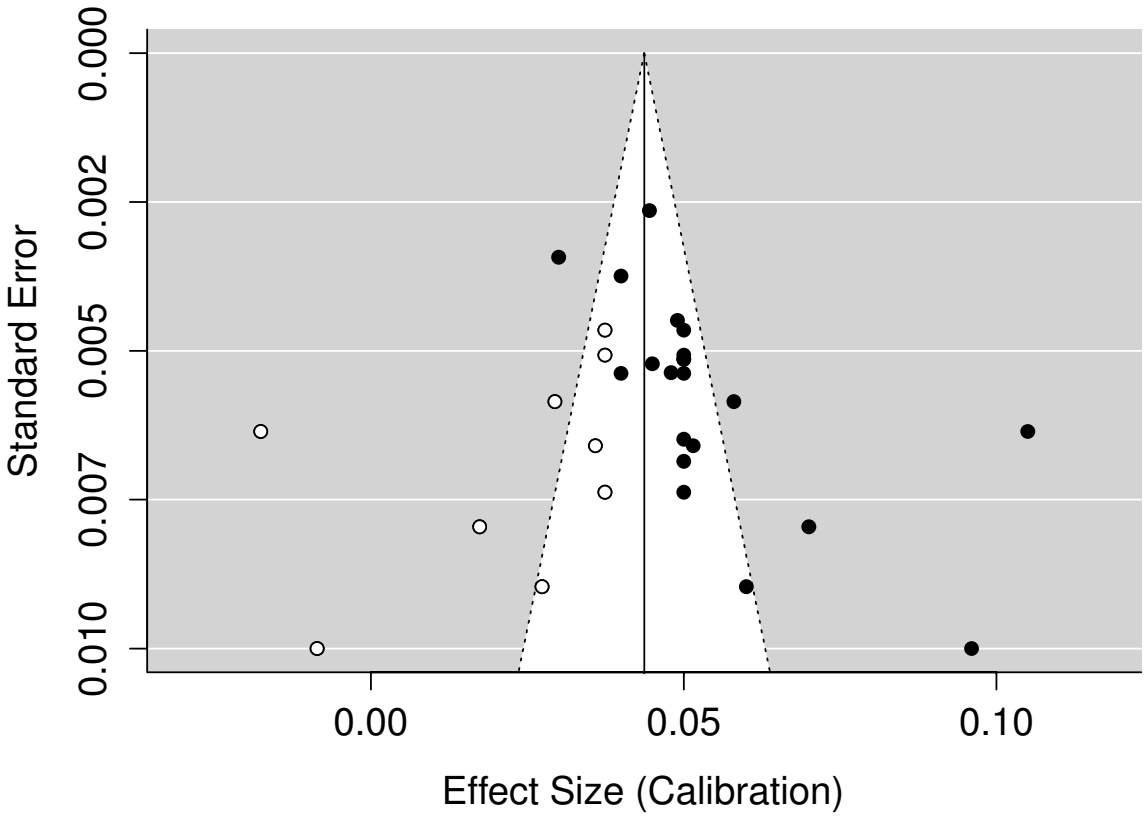


Figure 2. Funnel plot for under-/overconfidence with observed effect sizes plotted against their standard errors.

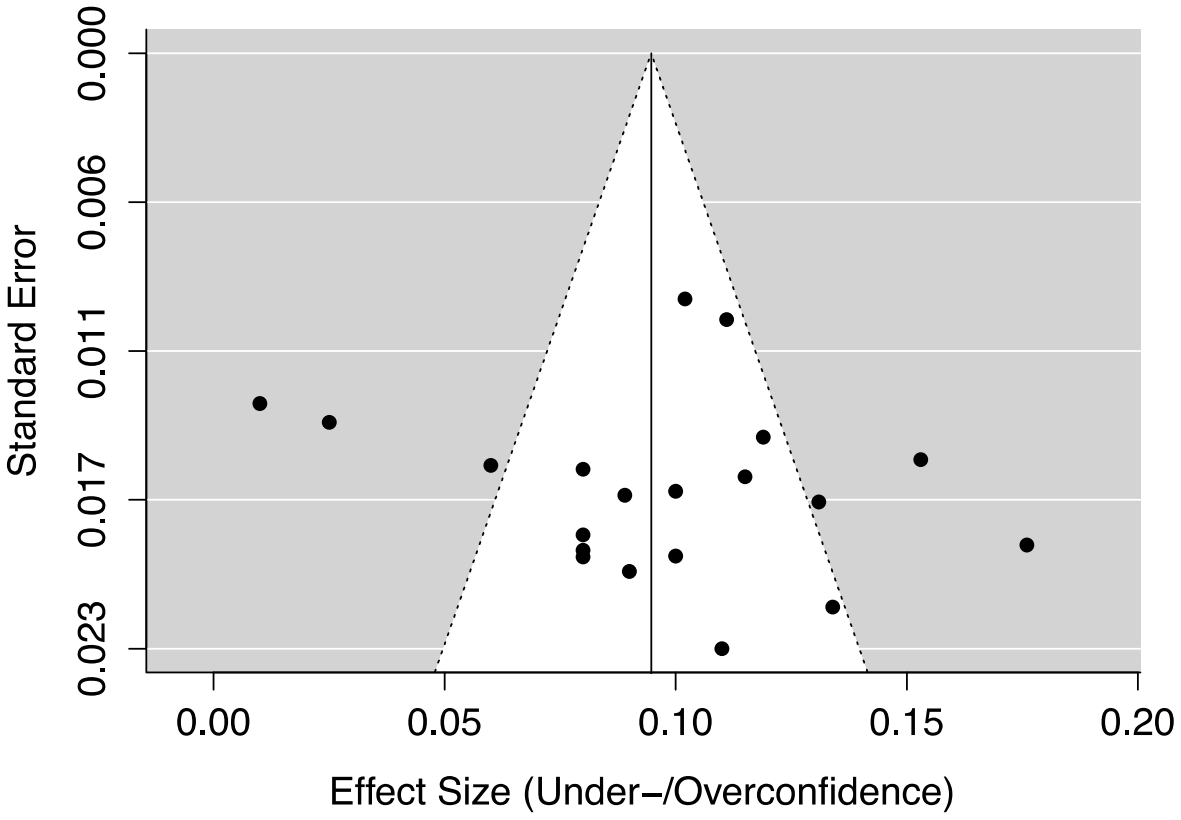
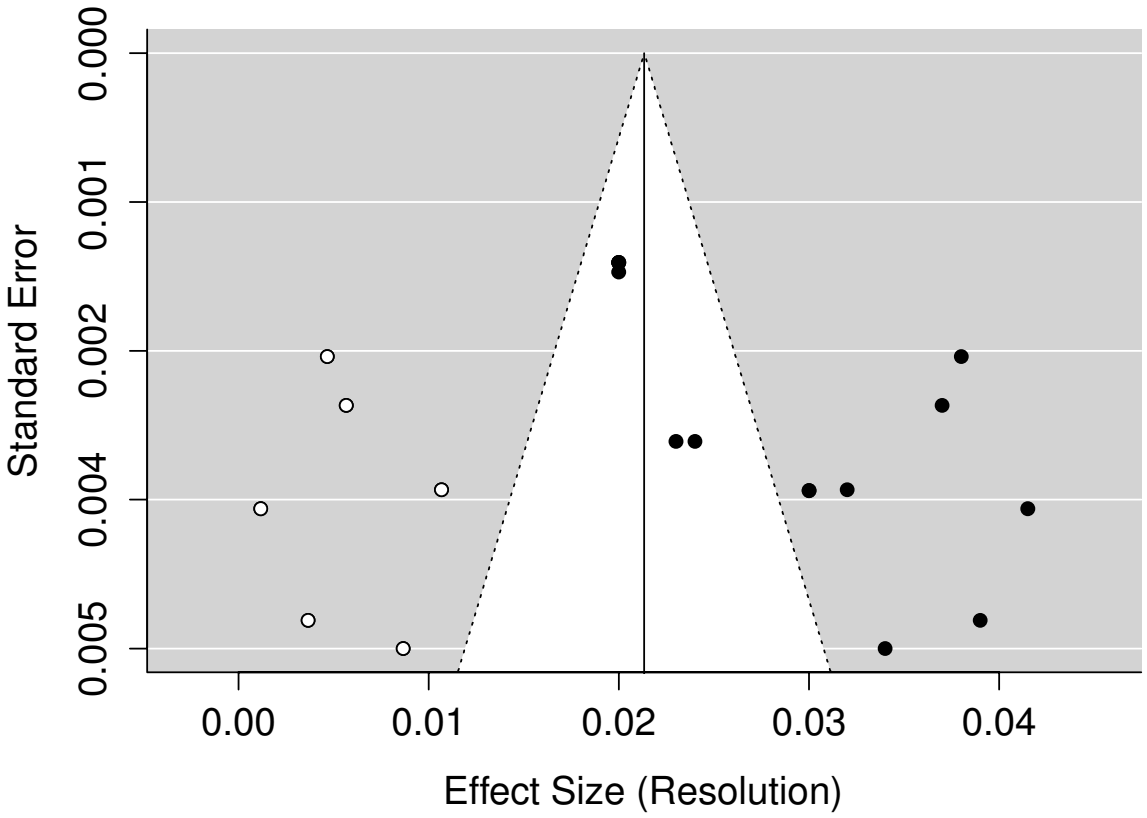


Figure 3. Funnel plot for resolution with observed (black dots) and imputed (white dots) effect sizes plotted against their standard errors.



Appendix B

Study 2:

Waubert de Puiseau, B., Weide, A. C., & Musch, J. (2016). How scripts influence overconfidence in eyewitness event memory: A model-based analysis. *Manuscript submitted for publication.*

How Scripts Influence Overconfidence in Eyewitness Event Memory: A Model-Based Analysis

Berenike Waubert de Puiseau[1], Anneke C. Weide[2], Jochen Musch[1]

[1]University of Duesseldorf, Germany

[2]University of Bonn, Germany

Word count (excluding Abstract, Tables, Figures, and References): 9,405

Author Note

Berenike Waubert de Puiseau and Jochen Musch, Department of Experimental Psychology, University of Duesseldorf, Germany; Anneke C. Weide, Department of Psychology, University of Bonn, Germany.

The authors would like to thank Edgar Erdfelder for helpful feedback on earlier versions of this manuscript. We would further like to thank Arvid Hofmann, Carolin Meschede, Thorsten Meschede, Tim Eichhorn, Jana Einicke, Richard Barker, and Arne Stops for their support in conducting this research.

Correspondence concerning this article should be addressed to Berenike Waubert de Puiseau, Department of Experimental Psychology, University of Duesseldorf, Universitätsstrasse 1, Building 23.03, 40225 Düsseldorf, Germany, phone: +49 – (0)211 – 8112063, fax: +49 – (0)211 – 8111753, e-mail: bwdp@uni-duesseldorf.de

Abstract

Witnesses are often overconfident about the accuracy of their testimonies on observed events. Drawing upon MINERVA-Decision-Making (Dougherty, 2001; Dougherty, Gettys, & Ogden, 1999), an exemplar-based theory capable of explaining overconfidence, we propose a theoretical account of accuracy, confidence, and overconfidence in eyewitness event recognition memory. This account is based on the assumption that actual and previous observations influence recognition and confidence judgments and that memory traces of previous observations determine the scripts witnesses have of crimes. Overconfidence is predicted to occur when the direct retrieval of information about a specific crime fails and familiarity based on memory traces is misleading. To test the model, we conducted an eyewitness simulation study. Seventy-nine students viewed a simulated crime and subsequently completed 102 true/false items about the video that conformed with, did not conform with, or were independent of commonly held scripts of the presented crime. To increase the proportion of familiarity-based retrievals, half of the participants watched the crime video while completing a distractor task. As predicted by the model, overconfidence increased under working memory load only for script-nonconforming items for which familiarity is misleading but not for script-conforming and script-neutral items. Limitations and potential applications of the proposed model are discussed.

Word count (Abstract): 200

Keywords: eyewitness memory; overconfidence; MINERVA-DM; confidence-accuracy relationship; recognition memory

How Scripts Influence Overconfidence in Eyewitness Event Memory: A Model-Based Analysis

Witnesses play a central role in the legal system (Rattner, 1988; Wells & Olson, 2003). With witnesses' help, fact-finders aim to reconstruct crimes and to facilitate the identification of perpetrators. However, eyewitness memory is usually not perfect, and the resulting reports are often flawed (Rattner, 1988; Wells, Memon, & Penrod, 2006). It is therefore important to reliably determine the accuracy of eyewitness testimonies. To this end, witnesses are often asked to provide confidence ratings to indicate their subjective certainty of whether statements about a crime they observed are correct. The assumption underlying the collection of such confidence ratings is that witnesses are able to monitor the accuracy of their reports. Confidence ratings have indeed been shown to reliably predict performance on general knowledge questions (Luna & Martín-Luengo, 2012; Perfect, Watson, & Wagstaff, 1993), but for eyewitness event memory, the correlation between confidence and accuracy seems to be less stable (Bothwell, Brigham, & Deffenbacher, 1987; Brewer & Wells, 2006; Palmer, Brewer, Weber, & Nagesh, 2013; Penrod & Cutler, 1995; Sporer, Penrod, Read, & Cutler, 1995).

However, legal fact-finders are not usually interested in such correlations. Rather, they want to know whether witnesses' mean confidence ratings approximate their accuracy (commonly measured as the proportion of correct responses; Buratti, Allwood, & Johansson, 2014; Juslin, Olsson, & Winman, 1996). Research on eyewitness memory has therefore recently shifted its focus to the calibration of confidence ratings, that is, the degree of correspondence between subjective certainty and accuracy (Allwood, Knutsson, & Granhag, 2006; Brewer, 2006). Studies have found that eyewitnesses are commonly overconfident about their responses to recognition questions; that is, their average subjective certainty about the correctness of the information they provide exceeds their average actual accuracy (Allwood, Granhag, & Johansson, 2003; Allwood, Innes-Ker, Homgren, & Fredin, 2008;

Allwood et al., 2006; Bonham & Gonzalez-Vallejo, 2009; Buratti & Allwood, 2012; Granhag, Jonsson, & Allwood, 2004; Granhag, Strömwall, & Allwood, 2000). This mirrors a general tendency toward overconfidence observed when people make decisions under uncertainty. For example, empirical studies have found evidence for overconfidence in lawyers' predictions of case outcomes (Goodman-Delahunty, Granhag, Hartwig, & Loftus, 2010), people betting on sports (Towfigh & Glöckner, 2011), financial analysts' forecasts of market behavior, and weather forecasts (Tyszka & Zielonka, 2002).

To account for such findings, the MINERVA-Decision-Making model was proposed (Dougherty et al., 1999). MINERVA-Decision-Making provides a comprehensive model of overconfidence (Dougherty, 2001) by specifying in detail the underlying cognitive processes that lead to good or poor calibration in general knowledge tasks. The model is a variant of the exemplar-based MINERVA-2 memory model (Hintzman, 1984), which has been successfully applied to explain negative correlations between accuracy and confidence in eyewitness identification decisions (Clark, 1997). However, little effort has been put toward understanding the cognitive processes underlying eyewitnesses' overconfidence in their testimony of observed events. The purpose of the present paper is to fill this gap by providing a detailed theoretical account of the cognitive processes underlying recognition accuracy and confidence judgments in eyewitness event memory. To this end, we used the MINERVA-Decision-Making model (Dougherty, 2001; Dougherty et al., 1999) as a framework for explaining overconfidence and the potential influence of scripts in eyewitness event recognition memory. This model has previously been proposed to account for the misinformation effect (Loftus, 1975) by Bonham and Gonzalez-Vallejo (2009). We used MINERVA-Decision-Making to take into account both episodic and semantic aspects of eyewitness event memory and to derive hypotheses about the determinants and the influence of scripts on eyewitnesses' accuracy, confidence, and overconfidence. These hypotheses were then tested in an eyewitness event recognition memory study.

In the following, we first review the existing research on the confidence-accuracy relationship in eyewitness memory. Second, we explain the MINERVA-Decision-Making model and introduce the application of this model to true/false items regarding criminal events observed by a witness. Third, we derive specific hypotheses about the influence of scripts and working memory capacity on overconfidence. Finally, we describe how we tested the hypotheses in an empirical crime simulation study.

Eyewitness Memory and the Confidence-Accuracy Relationship

In standard empirical studies assessing the association between confidence and accuracy, participants are presented with a video or a series of pictures showing a crime. They are subsequently asked to answer questions about their observations. Accuracy is usually measured as the proportion of correctly answered items. For each of their responses, the witnesses provide a confidence rating on a numerical scale. A correlation is then computed either between mean confidence and proportion correct across witnesses, or between confidence ratings and the correctness of individual responses across items (Smith, Ellsworth, & Kassin, 1989).

Multiple studies have examined the validity of confidence as a predictor of eyewitness identification accuracy. Several meta-analyses summarizing the results of these empirical studies found only a weak to moderate correlation between confidence and accuracy (Bothwell et al., 1987; Cutler & Penrod, 1989; Sporer et al., 1995). For event recognition memory, however, the confidence-accuracy relationship has been examined in only a few studies, and no definitive answer has been provided about whether the findings from identification research can be generalized to episodic memory (Allwood et al., 2006; Perfect, 2002).

An important caveat is that conclusions about the performance of an individual witness cannot be drawn from correlational analyses across witnesses or across individuals' responses (Juslin et al., 1996). The analysis of confidence calibration appears to be more

useful for judging the accuracy of an individual witness's testimony (Brewer & Wells, 2006; Buratti et al., 2014; Lichtenstein & Fischhoff, 1977, 1980). Unlike correlation coefficients, which indicate only the extent to which confidence and accuracy covary across or within subjects, calibration analyses can reveal whether witnesses' confidence exceeds their accuracy (overconfidence) or their accuracy exceeds their confidence (underconfidence; Juslin et al., 1996).

Following the approach most frequently chosen in previous research, we computed overconfidence as the mean difference between individual confidence and accuracy (for alternative definitions, see Moore & Healy, 2008). Multiple studies investigating the calibration of eyewitnesses' episodic memory have found witnesses to be overly confident in their recognition judgments, and rates of about 8% to more than 15% overconfidence have been observed (Allwood et al., 2003; Allwood et al., 2008; Allwood et al., 2006; Bonham & Gonzalez-Vallejo, 2009; Bornstein & Zickafoose, 1999; Buratti & Allwood, 2012; Granhag et al., 2004; Granhag et al., 2000). Particularly high overconfidence was observed when witnesses testified repeatedly (13-18% overconfidence; Allwood et al., 2008; Granhag et al., 2004).

The MINERVA-Decision-Making Model

There is a dearth of research on accuracy and confidence in eyewitness recognition memory of events (Allwood et al., 2006). Therefore, we applied the MINERVA-Decision-Making (MDM) model, a variant of the MINERVA-2 memory model (Hintzman, 1984, 1988), to shed light on the cognitive processes underlying overconfidence in eyewitness event recognition memory. MDM is an exemplar-based theory that was originally formulated to account for two-option general knowledge questions. Therefore, it can be easily adapted to the true/false questions that have been shown to be typical of police interviews (Fisher, 1995; Fisher, Geiselman, & Raymond, 1987; Peterson & Grant, 2001). We applied MDM to a standard recognition memory paradigm presenting items that referred to aspects of a crime

that happened and could be observed (targets) or did not happen and could not be observed (distractors).

To illustrate our approach, consider the following robbery: A man is attacked by a male criminal who forces the victim to hand over his wallet and cell phone. The witness who observed this crime is requested to help the police reconstruct the event. Details or aspects of the robbery that may be of interest include actions (e.g., whether the perpetrator approached the victim with a weapon), conversations (e.g., whether the perpetrator threatened the victim verbally), characteristics of the people who were involved (e.g., height, weight, ethnic background, accent), and the surroundings (e.g., whether cars were driving by). Let us assume that the police are particularly interested in whether the attacker was armed. Police officers may then ask witnesses to indicate their agreement with the following statement: "The perpetrator had a weapon." Whereas MDM as originally formulated by Dougherty et al. (1999) and Dougherty (2001) deals with two-alternative forced-choice questions, we adapted the model to true/false questions referring to either target or distractor items. According to the model, the above statement would be a target item if the robber did indeed carry a weapon. Conversely, if the robber did not carry a weapon, the statement would be a distractor item.

According to MDM, memory consists of memory traces. Whenever witnesses observe a particular event such as a robbery, they are assumed to store their observations as memory traces. Memory traces are assumed to be degraded copies of the observed events. Specific features of an event are either correctly encoded as present or absent or not encoded at all. For example, the robber may have worn a jacket and grey shoes and had no bag. If the witness noticed the jacket, this feature would be encoded as present in the witness's memory trace. If the witness did not notice the grey shoes, this feature would not be encoded. Finally, if the witness explicitly noticed that the perpetrator carried no bag, the feature "bag" would be encoded as absent.

When probed, witnesses may be able to retrieve their response directly from memory, or they may have to resort to a familiarity-based judgment. Direct responses are assumed to be usually correct and to be given with a high level of confidence. Consequently, according to the model, there is little room for overconfidence when responses are retrieved directly from memory. However, direct retrieval is possible only if the witness had an opportunity to observe a probed event detail and paid sufficient attention to the crime during encoding.

If no direct retrieval is possible, witnesses' responses are assumed to be based on familiarity. Familiarity is determined by the similarity between an event detail probed by the question and the robberies a witness has observed in the past (according to MDM, for a memory trace to be considered, it has to encode a robbery and it has to be identified as encoding a robbery). More precisely, the memory probe is compared with each identified memory trace of a robbery. The average of the resulting similarities determines the perceived familiarity of the probe. Witnesses of a robbery who have never observed—personally or in the media—a robbery without the use of a weapon are therefore expected to experience a strong feeling of familiarity when confronted with a statement about the use of a weapon in the robbery. This effect is expected to increase with the proportion of past robberies in which a weapon was involved and to be strongest for robberies that are highly similar to the ones the witness has stored in memory. Judgments can be based on familiarity even if witnesses have never observed a real crime in the past because virtually all witnesses can be assumed to hold memory traces of crimes due to repeated and long-term media exposure. It has been shown for the MINERVA-2 model, on which MDM is based, that scripts are abstracted from these memory traces (Hintzman, 1986). Witnesses have been found to use such scripts when testifying (Greenberg, Westcott, & Bailey, 1998; Holst & Pezdek, 1992). According to MDM, a person's confidence rating of a response should be proportional to the response's relative perceived familiarity when judgments cannot be based on direct retrieval. More precisely, confidence is assumed to be proportional to the ratio of the perceived familiarity of the more

familiar option to the sum of the perceived familiarities of both answer options. This implies that whenever the perceived familiarity for the true or the false response exceeds the perceived familiarity associated with the opposite answer option, confidence ratings are predicted to exceed the level of chance. If not even perceived familiarity can help a witness to prefer one answer alternative (i.e., true/false) over the other, witnesses are forced to respond randomly according to the model and are therefore expected to provide confidence ratings at a chance level.

Event details are more likely to be perceived and subsequently reported in eyewitness testimony when they are linked to scripts, that is, when they are script-consistent or script-inconsistent (cf. Abelson, 1981; Holst & Pezdek, 1992; Hudson, Fivush, & Kuebli, 1992; Schank & Abelson, 1977). If a detail complies with a script, a familiarity-based judgment is likely to result in high accuracy and high confidence. According to the model, however, overconfidence is predicted to occur if the observed event is atypical and therefore inconsistent with existing scripts. In this case, familiarity is deceiving and leads to inaccurate responses, even though witnesses may nevertheless report high levels of confidence due to perceived familiarity; the result is overconfidence. Familiarity-based retrievals become more likely when encoding of the specific crime under investigation was impaired, for example, because the witness had been distracted and therefore had not paid sufficient attention to the crime. Confirming this reasoning, empirical research has found that under cognitive load, people experience more script-conforming intrusions and exhibit a response tendency to preferentially choose typical answers. Under these conditions, they also make more script-conforming source misattributions (Bower, Black, & Turner, 1979; Kleider, Pezdek, Goldinger, & Kirk, 2008; Macrae, Hewstone, & Griffiths, 1993; Migueles & García-Bajos, 2006; Sherman, Groom, Ehrenberg, & Klauer, 2003; Stangor & Duan, 1991; Stangor & McMillan, 1992; Tuckey & Brewer, 2003a, 2003b).

Another potential cause of overconfidence is error variance in cognitive processes (cf. Erev, Wallsten, & Budescu, 1994). According to MDM, error variance may be caused by high levels of degradation in memory traces (e.g., due to limited attention) or by a high degree of diversity in the memory traces that belong to the same category (e.g., robberies). In either case, features may be encoded as present or absent or may not be encoded at all. For example, a witness may be assumed to have observed 10 robberies. In six of them, the robber may have used a weapon. If the witness was unable to pay sufficient attention while encoding two of the corresponding six memory traces, the witness may have failed to encode the presence of a weapon. In a similar vein, in two of the memory traces for robberies in which no weapon was used, the witness may have failed to encode the absence of a weapon. When probed about the presence of a weapon, the witness's perceived familiarity then has to be based on only six of the 10 memory traces because four of the memory traces provided no information about the presence of a weapon. When the level of perceived familiarity is determined as the mean similarity across all relevant memory traces, the variance of the similarities is larger when fewer memory traces are considered, and the chance that the probe will be incorrectly accepted or rejected thereby increases. The confidence level, however, is assumed to be proportional to the mean similarity and not to be affected by the number of memory traces considered. In other words, error variance decreases only accuracy but not confidence, and this is the essence of the error variance account of overconfidence in the MDM model (cf. Dougherty, 2001). Similarly, when the stored crime events are characterized by high diversity, the chance that specific features of an observed crime will not be encoded in individual memory traces for past crimes increases, and this also reduces the number of memory traces available to determine perceived familiarity. High levels of degradation or diversity in memory traces may also lead to a failure to identify relevant memory traces and, thus, a failure to consider them when determining perceived familiarity. This leads to a

reduced number of traces that can be used to determine the person's responses and confidence ratings.

To summarize, the single-process MDM model provides a comprehensive theoretical framework for overconfidence in eyewitness event recognition memory. It is important to note that it can also account for the effects of two aspects that were independently suggested as sources of overconfidence: misleading familiarity and error variance in cognitive processes (cf. Dougherty, 2001).

The Present Study

The present investigation is the first to employ MDM to model the interplay of episodic and semantic memory processes underlying eyewitness overconfidence in recognizing events details. To examine the validity of the model, we tested the predictions MDM makes about two important moderators of overconfidence: working memory load and the relevance of crime details to a script. MDM makes specific and falsifiable predictions about the impact of these two variables on accuracy, confidence, and overconfidence. We tested these predictions in an experimental study using a simulated crime that entailed script-consistent, script-inconsistent, and script-neutral details. We manipulated the possibility of familiarity-based retrieval by imposing a working memory load on half of the participants. To impair their encoding, these participants had to complete a demanding distractor task while watching the simulated crime video. We thus hoped to reduce the likelihood of complete encoding and a subsequent direct retrieval of crime details and to increase the necessity to make familiarity-based judgments. Participants watching the video without additional memory load served as the control group.

After watching the video, participants judged the correctness of 102 statements about the observed crime. Three kinds of item sets were generated: one third of the items were script-conforming[1]. For these items, script-based processing would result in a correct response. For example, given a script containing a male robber, an item referring to the robber

being male would be answered correctly using the script if the robber was indeed male in the observed crime. Another third of the items were script-nonconforming. For these items, script-based processing would result in an incorrect response. For example, an item referring to the robber being male would be answered incorrectly using the script if the robber was actually female in the observed crime. Script-conforming and script-nonconforming items were considered script-relevant. The final third of the items were script-neutral. For these items, no answer was expected to be preferred on the basis of script-consistency.

Overconfidence has been shown to covary with item difficulty (the hard-easy effect; cf. Juslin, 1993, 1994). In the no-working-memory-load control group, differences in difficulty between item sets were, therefore, intentionally removed by generating item sets of equal average difficulty. Differences in accuracy, confidence, and overconfidence between the experimental and the control group could thus be attributed to the effect of working memory load and, thereby, to the proportions of direct and familiarity-based retrievals. This allowed us to test MDM's predictions about the effects of script relevance on eyewitness event recognition memory.

Hypotheses

Crimes are often observed when working memory capacities are limited. Furthermore, the complexity of crimes likely contributes to high variability in memory traces. Both of these factors lead to high levels of error variance according to the model, which therefore predicts an overall high level of overconfidence for eyewitness event recognition memory. In the following, we explain the additional hypotheses that can be derived from the model regarding the effects of script relevance and working memory load on witnesses' accuracy, confidence, and overconfidence.

Accuracy. According to the model, the proportion of familiarity-based judgments is expected to increase under working memory load because complete encoding and, thus, direct retrievals are made difficult. Therefore, accuracy for script-nonconforming items was

predicted to decrease considerably under working memory load because the feeling of familiarity is misleading for these items. For example, take a script-inconsistent female robber. Because male robbers are much more familiar, the familiarity-based response that would be given if a direct retrieval of the robber's gender was impossible would be "the robber was male." This answer would be wrong if a script-inconsistent female robber had committed the crime. Given sufficient working memory capacity, witnesses can be expected to be able to encode the script-inconsistent gender of a female robber. However, this is not the case under working memory load when witnesses have to base recognition judgments on perceived familiarity, and the expected proportion of correct familiarity-based responses is low. A contrasting prediction can be derived for script-consistent crime details. For these aspects of the crime, familiarity-based judgments made under working memory load can be expected to be correct. Thus, for accuracy, it should make little difference whether responses are retrieved directly or based on perceived familiarity.

Aspects of the crime that are neutral with respect to the underlying script, (e.g., the color of the perpetrator's jacket) generally receive little attention (Bower et al., 1979; Hashtroudi, Mutter, Cole, & Green, 1984), and witnesses therefore often fail to encode such details. Under working memory load, the proportion of direct retrievals is reduced for script-neutral details, and accuracy would therefore be expected to decrease in this situation. As familiarity-based retrievals usually favor neither response option (true/false), the accuracy of familiarity-based responses for script-neutral items was predicted to be at or near the level of chance. Consequently, the decrease in accuracy under working memory load should be smaller for script-neutral than for script-nonconforming items for which the expected accuracy for familiarity-based responses was predicted to be below the level of chance.

In the present study, the items were designed to be equally difficult in all three item sets in the control condition. Therefore, no significant differences between item sets regarding accuracy were expected when no additional working memory load was applied. According to

the model, under working memory load in the experimental condition, the highest accuracy would be expected for items conforming to the script. The lowest accuracy would be expected for items that conflicted with the script. The accuracy for script-neutral items should lie between the accuracies for script-conforming and script-nonconforming item sets.

Confidence. According to the model, direct retrievals are associated with maximum confidence. When witnesses base their responses on familiarity, confidence ratings should be proportional to perceived familiarity and should therefore be lower than for direct retrievals. Because eyewitnesses make fewer direct retrievals with high confidence under working memory load, their mean confidence was predicted to be reduced in the experimental condition regardless of script-relevance.

The model predicts similar levels of confidence for script-conforming and script-nonconforming items because script-relevant items generally attract more attention during encoding than script-irrelevant items, leading to higher proportions of direct retrievals. Furthermore, all script-relevant items refer to the same memory trace features. Script-conforming and script-nonconforming items should therefore produce similar levels of perceived familiarity, on which witnesses base their confidence ratings. Again, this can be illustrated by using a script-inconsistent female robber. Regardless of whether an item referring to a female robber is true (and thus, script-nonconforming) or false (and therefore, script-conforming), the witnesses have to refer to the same features in the same memory traces. A judgment about whether a robber was male or female therefore always has to be based on the same perceived feeling of familiarity. As a result, the model suggests that confidence ratings should not depend on whether script-consistent or script-inconsistent aspects of the crime are actually true or false.

As script-neutral items such as the color of the robber's clothes attract less attention than script-relevant details, the model generally predicts only a small number of direct retrievals with maximum confidence for script-neutral items. Furthermore, the perceived

levels of familiarity and the resulting confidence levels are generally predicted to be near the level of chance for script-neutral items and, thus, to be lower than for script-relevant items because, by definition, script-neutral aspects are less distinct than script-consistent or script-inconsistent aspects of a crime. Consequently, participants' confidence in their judgments of script-neutral items was predicted to be lower than their confidence in their judgments of script-relevant items.

Overconfidence. Drawing upon the model's predictions involving accuracy and confidence, less overconfidence for script-conforming items would be expected under working memory load. This is because accuracy should not be affected by working memory load, whereas confidence should decrease. The model makes a contrary prediction for script-nonconforming items, however. For these items, more overconfidence would be expected because under working memory load, the proportion of familiarity-based responses should increase. For familiarity-based responses, accuracy for script-nonconforming items was expected to be below the level of chance, whereas perceived familiarity and, therefore, confidence should always be above chance. In other words, overconfidence was expected to increase under working memory load in the experimental condition because accuracy and confidence were expected to diverge when working memory load was high. Regarding script-neutral items, the model predicts reductions in both accuracy and confidence under working memory load due to an increase in the number of familiarity-based responses. Therefore, overconfidence was not expected to increase under working memory load.

To test the specific predictions derived from MDM, we conducted a crime simulation experiment in which we manipulated working memory load and script conformity of crime details to examine the impact of these variables on eyewitnesses' accuracy, confidence, and overconfidence. To the extent that the predicted complex pattern of results could indeed be observed, MDM would receive empirical support as a viable model of the cognitive processes involved in eyewitness event recognition memory.

Method

Participants

Eighty-one psychology students from the University of Düsseldorf participated in the study. They were recruited in lectures or on campus and received course credit and a chocolate bar for their participation. Two participants had to be excluded from the study because they failed to comply with the instructions. Of the remaining 79 students, 69 were female (87.3%). Their average age was 22.8 years (SD = 5.7). None of the participants indicated impaired eyesight or hearing.

Design

A 2 x 3 mixed factorial design was employed with working memory load (no distraction task vs. distraction task) as a between-subjects factor and script conformity of the items (script-conforming vs. script-nonconforming vs. script-neutral) as a within-subjects factor. Participants either performed a distraction task during the study phase (working memory load) or did not perform a distraction task (no working memory load). Participants were randomly assigned to the between-subjects conditions. Power analyses using the software G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that the sample size was sufficient to detect medium between-subjects effects and small within-subjects and interaction effects (f = 0.20-0.25, 1-$\beta \approx .80$).

Materials

During the study phase, participants viewed a video of a simulated crime. In the subsequent questioning phase, they were presented with 102 statements about the content of this video.

Crime video. A video showing a robbery was created, staged by amateur actors and actresses. In the video, two members of a gang, a woman and a man, rob a young man. The perpetrators steal the victim's belongings and punch him before leaving the scene. The entire video lasts about eight minutes and contains both script-relevant (script-consistent and script-

inconsistent) and script-neutral details. An example of a script-consistent detail is that the perpetrator had a knife. An example of a script-inconsistent detail is that the lead perpetrator was female. A script-neutral detail is, for example, that a car drove by the crime scene during the robbery.

Items. To generate script-conforming, script-nonconforming, and script-neutral items, three sets of statements had to be created: script-consistent, script-inconsistent, and script-neutral statements, half of which were targets and half of which were distractors, respectively. This resulted in six sets of statements. The sets were matched in terms of item difficulty because overconfidence has been found to be influenced by how easily items can be solved (hard-easy effect; cf. Juslin, 1993).

Pilot study. We conducted a pilot study to generate the required item sets. In a first step, we generated 290 statements referring to the content of the video, of which approximately half were true, whereas the other half were false. The items addressed actions, conversations, the appearance of the persons involved, and the general surroundings. Forty students participated in the pilot study. All participants completed the pretest in approximately 30 min and received course credit for their participation. Half of the participants (n = 20) answered all items after watching the video to determine the item difficulties. The other half (n = 20) read a brief description of the robbery and were then asked to rate on a 7-point Likert scale how typical each detail referred to in the items was for a robbery such as the one depicted in the video. Across items and participants, the mean item difficulty (proportion of correct responses) was .69 (SD = .29), and the mean item typicality was 4.03 (SD = 0.31) on a scale ranging from 1 to 7.

Item subsets. On the basis of the typicality ratings, the items were divided into three preliminary sets: script-consistent statements (mean typicality ratings above 5.0), script-inconsistent statements (mean typicality ratings below 3.0), and neutral statements (mean typicality ratings between 3.5 and 4.5). These preliminary subsets were further divided into

true and false statements, resulting in a total of six subsets. Across all subsets, we selected groups of items that were similar in item difficulty, resulting in six final item subsets of 17 items each. The complete item set thus consisted of a total of 102 items (see Table 1). We then combined all subsets to generate the three factor levels of script conformity, thus making sure that script-consistency and statement correctness were manipulated orthogonally. The *script-conforming* item subset comprised items for which script-based responding resulted in a correct response. Of these items, 17 were script-consistent targets and 17 were script-inconsistent distractors. An example of a script-consistent target was "A member of the robbers' gang smokes a cigarette"; an example of a script-inconsistent distractor was "One of the robbers shakes hands with the victim during the robbery." The average difficulty of script-conforming items was .68 (*SD* = .24).

The *script-nonconforming* item subset contained items for which script-based responding resulted in an incorrect response. It comprised 17 script-consistent distractors and 17 script-inconsistent targets. An example of a script-consistent distractor was "One of the robbers steals the victim's wallet"; an example of a script-inconsistent target was "The victim scratches the paint on a car before the robbery." The average difficulty of the items in this subset was .69 (*SD* = .26).

Finally, the *script-neutral* item subset consisted of items for which script-based responding was expected to result in an accuracy rate at the level of chance. It comprised 17 neutral targets and 17 neutral distractors. An example of a script-neutral target was "At least one car drives by the crime scene during the robbery"; an example of a script-neutral distractor was "At least one person walking past the crime scene carries a backpack." The average item difficulty of the script-neutral items was .66 (*SD* = .24). Confirming the equivalence of all six item sets, a 3 x 2 between-subjects ANOVA on item difficulty with script conformity (script-conforming, script-nonconforming, script-neutral) and statement accuracy (true, false) as factors revealed no main effects and no interaction effect of these

factors on item difficulty (all Fs < 1). For the final item list, all statements were arranged chronologically by the order of events shown in the video.

- Place Table 1 about here -

Procedure

Data were collected in the laboratory of the Department of Psychological Assessment and Differential Psychology at the University of Düsseldorf. Participants were seated individually in an experimental cubicle. The software E-Prime 2 was used to present the stimuli on 19-inch monitors. After providing informed consent, participants answered a set of demographic questions and indicated whether they suffered from any eyesight or hearing impairments. Participants in the working-memory-load condition first received instructions for the distractor task they had to work on during the presentation of the video. A complex mental arithmetic task was employed to distract participants from the video enough to induce script-based responding. The task consisted of two steps that started with a four-digit number. Participants first subtracted the last number from the second-to-last number and then subtracted the resulting number from the original four-digit number. If the first difference was zero, participants were requested to subtract 1 to increase working memory load and to ensure that the task would not terminate prematurely. Participants were instructed to repeat both steps of the task until the video ended.

To ensure that participants had properly understood the distractor task, they were given three practice trials. The experiment continued only after participants had successfully completed the practice trials. To avoid excessive demands on working memory capacity, participants were provided with a piece of paper containing the rules for the secondary task at the top of the page and sufficient blank space to note the required four-digit numbers after each round of the two calculation steps outlined above. Participants were not allowed to make

notes about any intermediate results. To maximize comparability in study duration and cognitive burden, participants in both the working memory load and control conditions were asked to complete the practice trials for the distractor task. However, only participants in the working-memory-load condition were asked to complete the distractor task during the study phase.

Upon completion of the practice trials, all participants were instructed to put on headphones connected to their computer. Then, the study phase began, and the crime simulation was presented as an incidental learning task. Participants in the working-memory-load condition were additionally instructed to carefully complete the distractor task while watching the video but to nevertheless make sure they did not miss too much information from the video.

After the video ended, participants were told that the police had requested their support as witnesses of the crime they had just observed. Following a procedure detailed in Waubert de Puiseau, Assfalg, Erdfelder, and Bernstein (2012), participants were first asked to remember and think of what they had just seen in the video for 2 min. In the subsequent questioning phase, participants judged the correctness of 102 statements about the video by classifying all statements as either “True” or “False.” After classifying each statement, participants rated their confidence in their response on an 11-point confidence scale ranging from 0 (*guess*) to 100 (*absolutely certain*) in steps of 10. Upon completing all items, participants were thanked and debriefed. The study took 25-30 min to complete.

Results

We analyzed the data for accuracy, confidence, and overconfidence separately. We computed 2 x 3 mixed factorial ANOVAs for each of these three dependent variables with working memory load (load vs. no load) as a between-subjects factor and script conformity (script-conforming vs. script-nonconforming vs. script-neutral) as a within-subjects factor.

The specific model predictions were tested with planned contrasts. Unless stated otherwise, an α-level of .05 was employed for all tests of statistical significance.

Accuracy was computed for each participant as the proportion of correct responses across items. Mean accuracy for the total item set ranged from 49.01% to 75.49% with a mean of 64.25% (*SD* = 6.06). A performance at the level of guessing was indicated by a value of 0 on the confidence scale and a value of 50% on the accuracy scale. To use the same scale for both variables, we linearly transformed the confidence ratings from their original range (0-100) to the scale that was used for accuracy (50-100). Confidence was then computed for each participant across all items and ranged from 66.81 to 98.24, with a mean of 86.69 (*SD* = 5.03). Overconfidence was computed by subtracting accuracy from confidence separately for each participant. The resulting overconfidence scores ranged from 9.95 to 41.37 with a mean of 22.44 (*SD* = 6.19). The fact that all values were positive indicates that all participants showed over- rather than underconfidence.

Accuracy

A mixed factorial ANOVA on accuracy with Greenhouse-Geisser corrections for all tests involving the within-subjects factor script conformity showed a significant main effect of the working memory load manipulation, $F(1, 77) = 43.49$, $p < .001$, $\eta_p^2 = .36$. The mean level of accuracy was lower for participants in the working-memory-load condition ($M = 60.58$, $SE = 0.78$) than in the control condition ($M = 67.82$, $SE = 0.77$). This shows that the manipulation of working memory load successfully reduced accuracy.

The main effect of the script conformity factor was also significant, $F(1.85, 142.72) = 9.17$, $p < .001$, $\eta_p^2 = .11$. The mean accuracy was highest for script-conforming items ($M = 67.22$, $SE = 0.83$) and was comparatively low for script-nonconforming ($M = 62.91$, $SE = 0.93$) and script-neutral items ($M = 62.47$, $SE = 0.93$).

The two-way interaction between working memory load and script conformity was also significant, $F(1.85, 142.72) = 9.66$, $p < .001$, $\eta_p^2 = .11$ (see Figure 1). Under working

memory load and in line with the predictions derived from the model, accuracy was significantly reduced for script-nonconforming items, $F(1, 77) = 49.41$, $p < .001$, $\eta_p^2 = .39$, and for script-neutral items, $F(1, 77) = 11.83$, $p = .001$, $\eta_p^2 = .13$. The difference in accuracy between the working memory load and the control conditions was 13.00% (i.e., 4.4 items) for script-nonconforming, and 6.39% (i.e., 2.0 items) for script-neutral items. The reduction in accuracy under working memory load was, thus, three times larger for script-nonconforming than for script-neutral items. This finding is in line with the model-based prediction that relying on familiarity would lead to incorrect responses for script-nonconforming items. Accuracy for script-conforming items was not affected by working memory load, $F(1, 77) = 1.95$, $p = .167$, $\eta_p^2 = .02$. Thus, when a crime detail was similar to the details in most of the comparable crimes stored in memory, familiarity-based recognition judgments were as accurate as direct retrievals. This is exactly the pattern that was predicted by the model.

When the analysis of accuracy was restricted to participants in the no-working-memory-load control condition, there was no effect of the script conformity factor, $F(2, 154) = 2.53$, $p = .083$, $\eta_p^2 = .03$, confirming that any differences in difficulty between the item sets had successfully been removed. For the participants in the working-memory-load condition, script conformity significantly impacted accuracy, $F(2, 154) = 16.13$, $p < .001$, $\eta_p^2 = .17$. As predicted by the model, the participants answered the script-conforming items ($M = 66.06$, $SE = 1.22$) significantly more accurately than the script-neutral items ($M = 59.28$, $SE = 1.20$), $F(1, 38) = 13.84$, $p = .001$, $\eta_p^2 = .27$. Accuracy for the script-nonconforming items ($M = 56.41$, $SE = 1.43$) was significantly lower than for the script-neutral items, $F(1, 38) = 4.77$, $p = .035$, $\eta_p^2 = .11$.

All of these findings are in line with the model's prediction that familiarity-based recognition judgments are usually accurate for script-conforming items and usually inaccurate for script-nonconforming items. The model also predicts that familiarity-based recognition

judgments for script-neutral items should approach the level of chance under working memory load. Given that mean accuracy for script-neutral items was higher than 50% even under working memory load, participants seemed to have had some working memory capacity left to encode a few script-neutral details in a way that later allowed them to make direct retrievals. The mean accuracy for script-nonconforming items also exceeded the level of chance under working memory load, which implies that direct retrievals also occurred for some of the script-nonconforming items. The reduction in accuracy for these two kinds of items is in line with the model's prediction that under working memory load, the proportion of direct retrievals should be reduced. Under load, witnesses increasingly have to base their responses on familiarity, which, however, does not provide the correct answer. No reduction in accuracy occurred for script-conforming items, a finding that complies with the model's prediction that familiarity-based retrievals produce mostly correct recognition judgments.

- Place Figure 1 about here -

Confidence

Second, we examined confidence ratings. A 2 x 3 mixed factorial ANOVA on confidence ratings yielded significant main effects for the between-subjects manipulation of working memory load, $F(1, 77) = 22.74$, $p < .001$, $\eta_p^2 = .23$. In line with the model's prediction that under working memory load, people make fewer direct retrievals and are therefore less confident, participants in the no-working-memory-load control condition were more confident about their responses ($M = 89.04$, $SE = 0.70$) than participants in the working-memory-load condition ($M = 84.27$, $SE = 0.71$).

The within-subjects manipulation of script conformity was significant as well, $F(2, 154) = 24.87$, $p < .001$, $\eta_p^2 = .24$. Again confirming the model's prediction, participants were comparably confident in their responses to the script-conforming ($M = 87.28$, $SE = 0.60$) and

script-nonconforming items (M = 87.26, SE = 0.51), $F(1, 77) < 1$. This complies with the model-based assumption that for both script-conforming and script-nonconforming items, familiarities are based on the same features encoded in the memory traces. Participants were less confident in their responses to the script-neutral items (M = 85.44, SE = 0.48) than in their responses to the script-relevant items, $F(1, 77) = 53.80$, $p < .001$, $\eta_p^2 = .41$, presumably because they paid less attention to and, therefore, made fewer direct retrievals of script-neutral details.

As predicted by the model, the two-way interaction between working memory load and script conformity was not significant, $F(2, 154) = 1.15$, $p = .321$, $\eta_p^2 = .02$ (see Figure 2).

- Place Figure 2 about here -

Overconfidence

A third 2 x 3 mixed factorial ANOVA was computed on overconfidence. Even though there was no significant between-subjects main effect of working memory load, there was a tendency for higher overconfidence to occur under working memory load (M = 23.69, SD = 6.22) compared with the no-working-memory-load control condition (M = 21.22, SD = 6.00), $F(1, 77) = 3.21$, $p = .077$, $\eta_p^2 = .04$.

The within-subjects main effect of script conformity was significant, $F(2, 154) = 6.66$, $p = .002$, $\eta_p^2 = .08$. Overall overconfidence was lowest for script-conforming items (M = 20.05, SE = 0.98) and comparably high for script-neutral (M = 25.93, SE = 1.66) and script-nonconforming items (M = 24.35, SE = 0.97).

The two-way interaction between working memory load and script conformity was significant too, $F(2, 154) = 9.77$, $p < .001$, $\eta_p^2 = .011$ (see Figure 3). The script conformity of the items had no effect on overconfidence in the no-working-memory-load control condition,

$F(2, 154) < 1$. In the control condition, the average level of overconfidence was similar for script-conforming ($M = 21.39$, $SE = 1.36$), script-nonconforming ($M = 20.38$, $SE = 1.22$), and script-neutral items ($M = 21.90$, $SE = 1.42$). As differences between item sets had intentionally been minimized in the control condition, this result was to be expected. Under working memory load, however, script conformity significantly influenced overconfidence, $F(2, 154) = 15.81$, $p < .001$, $\eta_p^2 = .17$. As the model predicted, participants were significantly more overconfident in their responses to script-nonconforming items ($M = 28.31$, $SE = 1.50$) than in their responses to script-neutral items ($M = 24.04$, $SE = 1.37$), $F(1, 38) = 10.22$, $p = .003$, $\eta_p^2 = .21$, and overconfidence in response to script-neutral items significantly exceeded overconfidence in response to script-conforming items ($M = 18.71$, $SE = 1.41$), $F(1, 38) = 8.28$, $p = .007$, $\eta_p^2 = .18$.

As predicted by the model, overconfidence was affected by working memory load for script-nonconforming items. For these items, participants in the working memory load condition showed much more overconfidence than participants in the control condition, $F(1, 77) = 16.91$, $p < .001$, $\eta_p^2 = .18$. The accuracy for script-nonconforming items was expected to be reduced under working memory load due to an increase in the proportion of familiarity-based retrievals that were misleading because they favored an incorrect script-consistent response rather than the correct script-inconsistent response. According to MDM, however, the confidence ratings were predicted to be above the level of chance regardless of the accuracy of the recognition judgments. The accuracy of familiarity-based retrievals for script-nonconforming items was expected to be below the level of chance because perceived familiarity would be misleading for these items. With an increasing proportion of familiarity-based retrievals, accuracy was expected to drop, and the accuracy rate for retrievals that were based solely on familiarity was expected to approach zero. Confidence was expected to decrease, too, with an increasing proportion of familiarity-based retrievals because, different from direct retrievals, familiarity-based retrievals are usually not reported with maximum

confidence. According to the model, however, the confidence ratings were predicted to exceed the level of chance whenever perceived familiarity could be used to differentiate between the answer options; this was predicted to be the case for all script-relevant items. The decrease in confidence associated with the increase in familiarity-based retrievals that had to be expected under working memory load was, therefore, predicted to be smaller than the decrease in accuracy for script-nonconforming items. Thus, overconfidence was expected to increase with an increasing proportion of familiarity-based retrievals. This is exactly what we found.

As predicted by the model, overconfidence for script-conforming items was reduced under working memory load. However, this reduction was small and did not reach statistical significance, $F(1, 77) = 1.86$, $p = .176$, $\eta_p^2 = .02$. A small reduction in overconfidence is in line with the model's prediction that for script-conforming items, familiarity-based recognition judgments produce accuracy levels that are similar to those produced by direct retrievals, whereas confidence ratings should be somewhat lower for familiarity-based retrievals than for direct retrievals. More precisely, direct retrievals should be associated with maximum confidence, whereas for familiarity-based retrievals, confidence should be proportional to familiarity and, therefore, it should be below the maximum confidence level. The predicted and observed decrease in confidence due to an increase in the proportion of familiarity-based retrievals, however, was not large enough to be statistically significant.

In line with the model's prediction, overconfidence for script-neutral items was not affected by the working-memory-load manipulation, $F(1, 77) = 1.17$, $p = .282$, $\eta_p^2 = .02$. For script-neutral items, the model predicted that accuracy would approach the level of chance under working memory load and, thus, both accuracy and confidence were expected to decrease as the proportion of direct retrievals decreased.

- Place Figure 3 about here -

Discussion

We proposed and experimentally tested a theoretical account of the cognitive processes underlying accuracy, confidence, and overconfidence in eyewitness event recognition memory. Our approach drew upon the exemplar-based MINERVA-Decision-Making (MDM; Dougherty, 2001; Dougherty et al., 1999) model, a theoretical model of overconfidence that was first formulated to account for overconfidence in judgments and decision making. This model-based approach to eyewitness event memory has several advantages. Comprehensive theoretical accounts such as MDM can be generalized to multiple situations as they detail not only *what* factors influence overconfidence but also *how* they interact in doing so (Ogloff, 2000). Our model can therefore not only be used to explain the effects of the manipulations applied in the present study, but as we will discuss further below, it can also explain other research findings in the domain of eyewitness event recognition memory, and it can be used to derive and test further predictions. In this respect, our approach complies with the requests of various researchers to conduct more theory-driven forensic psychological research to bridge the gap between basic and applied research (Lane & Meissner, 2008; Ogloff, 2000; Turtle, Read, Lindsay, & Brimacombe, 2008). Choosing an exemplar-based approach seems appropriate as several studies have suggested that for complex tasks (e.g., eyewitness observations) that involve nonlinear and nonadditive combinations of cues or cue-correlations, memory-based decision-making processes can be better described as exemplar-based processing than as a result of cue-abstraction (Bonham & Gonzalez-Vallejo, 2009; Juslin, Karlsson, & Olsson, 2008; Juslin, Olsson, & Olsson, 2003; Karlsson, Juslin, & Olsson, 2008).

To validate the model, we experimentally tested the model's predictions with regard to the addition of working memory load and the effect of the items' script conformity on witnesses' accuracy, confidence, and overconfidence. The directions of all effects were in

accordance with the model's predictions and, with one exception, all effects were statistically significant. As predicted, only the accuracy for script-nonconforming and script-neutral items decreased under working memory load, whereas familiarity-based responses for script-conforming items were usually correct. Also as predicted, the confidence ratings for script-conforming and script-nonconforming items were virtually identical and exceeded those for the script-neutral items. Moreover, and as expected, confidence decreased under working memory load for all item sets. Also in line with the model's predictions, we found a high overall level of overconfidence that the MDM model explains as the result of error variance in memory traces. As predicted, overconfidence further increased under working memory load for script-nonconforming items, for which script-based responding led to incorrect recognition judgments. By contrast, and also in line with our prediction, although failing to reach statistical significance, overconfidence decreased under working memory load for script-conforming items, for which accuracy was expected to remain stable, whereas confidence was expected to decrease. The small size of this effect may have been caused by the fact that accuracy did not remain completely stable but was slightly reduced when working memory load was applied. For script-neutral items, working memory load had no effect on overconfidence; for these items, however, no effect of memory load was expected according to the MDM.

The model can thus account for all of the present findings. However, it can also account for several additional findings that have been observed in other experiments on overconfidence in eyewitness event memory. For example, the presentation of misinformation subsequent to observing a crime has been found to decrease accuracy and increase overconfidence (e.g., Bonham & Gonzalez-Vallejo, 2009; Jack, Zydervelt, & Zajac, 2014). MDM can account for this finding because the model proposes that witnesses generate a memory trace each time they observe information regarding a crime. Witnesses are therefore expected to generate an additional memory trace for each piece of misinformation they

receive. This results in an increase in the number of memory traces containing features that are incorrect with respect to the crime under investigation, and this in turn leads to a higher probability of giving an incorrect response due to a false feeling of familiarity. In line with this reasoning, Jack et al. (2014) found that decreases in accuracy were largest when the same piece of misinformation was received twice (from both a co-witness and the interviewer) instead of only once. Another finding that can be explained by this reasoning is an increase in confidence without an accompanying change in accuracy as has been reported to result from repeated postevent questioning (Shaw & McClure, 1996).

Given that people usually view high confidence ratings as valid predictors of accuracy, overconfidence poses a serious threat to the fairness of legal and judicial decision making (McClure, Myers, & Keefauver, 2013; Potter & Brewer, 1999; Simons & Chabris, 2011). It is therefore desirable to figure out how to decrease overconfidence in eyewitness reports. One way to reduce overconfidence that has been suggested is to provide witnesses with a respective warning. MDM, however, predicts that warnings are ineffective because the model posits that accuracy and confidence result from largely automatic processes that occur during encoding (cf. Dougherty, 2001) and that cannot be consciously controlled. In line with this prediction, García-Bajos and Migueles (2003) found that cautioning witnesses to report only facts and to not make inferences decreased both the number of correctly and incorrectly recalled details but did not decrease overconfidence.

Additional and testable predictions for future research can also be derived from MDM. For example, the model predicts that the accuracy of recognition judgments for script-consistent aspects should remain stable over time, whereas accuracy should decrease for script-inconsistent details. This is because the proportion of direct retrievals is expected to decrease as time goes by, and familiarity-based responding should lead to incorrect responses when familiarity is misleading, which is to be expected for script-inconsistent details. In line with this reasoning, empirical studies have found that script-consistent details are retained

more accurately than script-inconsistent details (Tuckey & Brewer, 2003a, 2003b). However, this latter research did not measure overconfidence, which our model makes a specific prediction about: Overconfidence is predicted to increase over time for script-inconsistent but not for script-consistent details. For the former, accuracy is expected to remain stable, whereas confidence is expected to decrease. Thus, time is expected to impact overconfidence in a manner resembling the impact of working memory load.

MDM further predicts that overconfidence should be positively related to response time because response time is larger for familiarity-based retrievals than for direct retrievals and so is overconfidence. Direct responses are usually provided with high accuracy and high confidence and thus low overconfidence, whereas familiarity-based retrievals are associated with lower accuracy and therefore potentially higher levels of overconfidence. Investigating response time as a potential indicator of accuracy, Robinson, Johnson, and Herndon (1997) indeed found negative relationships between reaction time and both testimony accuracy and confidence. This is consistent with the model's prediction that direct retrievals, which require less extensive cognitive processing than familiarity-based retrievals and are therefore expected to be quicker than familiarity-based retrievals, are associated with maximum accuracy and confidence. However, Robinson and colleagues (1997) did not measure overconfidence, which MDM predicts to be positively related to response time.

MDM models the cognitive processes underlying accuracy and confidence in event memory and addresses both episodic and semantic aspects. This allows researchers to define the conditions under which overconfidence should be small or large. As many empirical studies including the current one have shown, scripts are particularly relevant under working memory load. The model's explanation for this finding is that under working memory load, episodic memory representations are less strong, and the likelihood of direct retrievals is reduced. Consequently, the proportion of familiarity-based retrievals necessarily increases. The model then predicts that crime details that conflict with the script will be recognized less

accurately, and witnesses will be more overconfident because they are misled by familiarity when testifying about script-conflicting crime details. Thus, for a very atypical crime, high overconfidence is expected, especially when witnesses had little working memory capacity available to process information when observing the crime. In the worst case, under extreme working memory load and if all questions are script-nonconforming, eyewitnesses would be expected to produce only incorrect responses with high confidence because they would be expected to report only the content of their scripts rather than the actual crime. For fairly typical crimes, the model still predicts a substantial level of overconfidence due to the error variance in the cognitive processes as outlined above, but this level of overconfidence is predicted to be much lower than for atypical crimes. Generally, the model predicts overconfidence to be an almost ubiquitous phenomenon in eyewitness memory. Anyone participating in the judicial process should therefore always be aware of the potential for overconfident eyewitnesses. According to the model, warning such overconfident witnesses not to overestimate their own performance cannot be expected to solve this problem, though.

In the present study, the overall level of overconfidence was high, which is in line with the model's predictions. However, it seems possible that the overconfidence observed in the present study was higher than the level of overconfidence that can be expected in real-life police interviews because we controlled for item difficulty in the item sets to avoid confounding script conformity and item difficulty. To this end, we increased the proportion of script-nonconforming and script-neutral items that were associated with a higher level of overconfidence. However, such items can be expected to be less frequent than script-conforming items in a natural environment. The items in the present study were, thus, not a representative sample of items, and this condition is known for accentuating overconfidence (Gigerenzer, Hoffrage, & Kleinbolting, 1991; Juslin, 1994). We nevertheless found substantial levels of overconfidence also for script-conforming items in our study, for which familiarity-based responding was predicted to result in correct responses. MDM's error

variance account for overconfidence (Dougherty, 2001; Erev et al., 1994) can explain this finding as the result of a large amount of variability in previously observed crimes or as the result of a high working memory load during previous crime observations. Most items employed in the present investigation were rather difficult, with item difficulties ranging from .64 to .70 in the no-working-memory-load control condition. As overconfidence is generally higher for more difficult items, this high level of item difficulty most likely also contributed to the overall high level of overconfidence.

To sum up, witnesses play an important role in legal fact-finding processes. Their confidence is often used to judge their accuracy. Therefore, eyewitness overconfidence may lead legal fact-finders to overestimate the correctness of witnesses' statements. To minimize the threat posed by overconfidence, it is important to understand the cognitive processes underlying this frequent phenomenon. We argue that model-based studies such as the present one offer a promising approach to a better understanding and an accurate forensic assessment of eyewitness event recognition memory.

References

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist, 36*, 715-729. doi: 10.1037/0003-066x.36.7.715

Allwood, C. M., Granhag, P. A., & Johansson, M. (2003). Increased realism in eyewitness confidence judgements: The effect of dyadic collaboration. *Applied Cognitive Psychology, 17*, 545-561. doi: 10.1002/acp.888

Allwood, C. M., Innes-Ker, A. H., Homgren, J., & Fredin, G. (2008). Children's and adults' realism in their event-recall confidence in responses to free recall and focused questions. *Psychology, Crime & Law, 14*, 529-547. doi: 10.1080/10683160801961231

Allwood, C. M., Knutsson, J., & Granhag, P. A. (2006). Eyewitnesses under influence: How feedback affects the realism in confidence judgements. *Psychology, Crime & Law, 12*, 25-38. doi: 10.1080/10683160512331316316

Bonham, A. J., & Gonzalez-Vallejo, C. (2009). Assessment of calibration for reconstructed eyewitness memories. *Acta Psychologica, 131*, 34-52. doi: 10.1016/j.actpsy.2009.02.008

Bornstein, B. H., & Zickafoose, D. J. (1999). "I know I know it, I know I saw it": The stability of the confidence-accuracy relationship across domains. *Journal of Experimental Psychology: Applied, 5*, 76-88. doi: 10.1037/1076-898x.5.1.76

Bothwell, R. K., Brigham, J. C., & Deffenbacher, K. A. (1987). Correlational of eyewitness accuracy and confidence. Optimality hypothesis revisited. *Journal of Applied Psychology, 72*, 691-695. doi: 10.1037/0021-9010.72.4.691

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11*, 177-220. doi: 10.1016/0010-0285(79)90009-4

Brewer, N. (2006). Uses and abuses of eyewitness identification confidence. *Legal and Criminological Psychology, 11*, 3-23. doi: 10.1348/135532505x79672

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*, 11-30. doi: 10.1037/1076-898X.12.1.11

Buratti, S., & Allwood, C. M. (2012). The accuracy of meta-metacognitive judgments: Regulating the realism of confidence. *Cognitive Processes, 13*, 243-253. doi: 10.1007/s10339-012-0440-5

Buratti, S., Allwood, C. M., & Johansson, M. (2014). Stability in the metamemory realism of eyewitness confidence judgments. *Cognitive Processes, 15*, 39-53. doi: 10.1007/s10339-013-0576-y

Clark, S. E. (1997). A familiarity-based account of confidence-accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 232-238. doi: 10.1037/0278-7393.23.1.232

Cutler, B. L., & Penrod, S. (1989). Forensically relevant moderators of the relation between eyewitness identification accuracy and confidence. *Journal of Applied Psychology, 74*, 650-652. doi: 10.1037/0021-9010.74.4.650

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General, 130*, 579-599. doi: 10.1037//0096-3445.130.4.579

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*, 180-209. doi: 10.1037/0033-295x.106.1.180

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous overconfidence and underconfidence: The role of error in judgment processes. *Psychological Review, 101*, 519-527. doi: 10.1037/0033-295x.101.3.519

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. doi: 10.3758/BF03193146

Fisher, R. P. (1995). Interviewing victims and witnesses of crime. *Psychology, Public Policy, and Law, 1*, 732-764. doi: 10.1037/1076-8971.1.4.732

Fisher, R. P., Geiselman, R. E., & Raymond, D. S. (1987). Critical analysis of police interview techniques. *Journal of Police Science and Administration, 15*, 177-185.

Garcia-Bajos, E., & Migueles, M. (2003). False memories for script actions in a mugging account. *European Journal of Cognitive Psychology, 15*, 195-208. doi: 10.1080/09541440244000102

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models. A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528. doi: 10.1037/0033-295x.98.4.506

Goodman-Delahunty, J., Granhag, P. A., Hartwig, M., & Loftus, E. (2010). Insightful or wishful: Lawyers' ability to predict case outcomes. *Psychology, Public Policy, and Law, 16*, 133-157. doi: 10.1037/a0019060

Granhag, P. A., Jonsson, A. C., & Allwood, C. M. (2004). The Cognitive Interview and its effect on witnesses' confidence. *Psychology Crime & Law, 10*, 37-52. doi: 10.1080/1068316021000030577

Granhag, P. A., Strömwall, L. A., & Allwood, C. M. (2000). Effects of reiteration, hindsight bias, and memory on realism in eyewitness confidence. *Applied Cognitive Psychology, 14*, 397-420. doi: 10.1002/1099-0720(200009)

Greenberg, M. S., Westcott, D. R., & Bailey, S. E. (1998). When believing is seeing: The effect of scripts on eyewitness memory. *Law and Human Behavior, 22*, 685-694. doi: 10.1023/a:1025758807624

Hashtroudi, S., Mutter, S. A., Cole, E. A., & Green, S. K. (1984). Schema-consistent and schema-inconsistent information: Processing demands. *Personality and Social Psychology Bulletin, 10*, 269-278. doi: 10.1177/0146167284102013

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*, 96-101. doi: 10.3758/BF03202365

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411-428. doi: 10.1037/0033-295X.93.4.411

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528-551. doi: 10.1037/0033-295x.95.4.528

Holst, V. F., & Pezdek, K. (1992). Scripts for typical crimes and their effects on memory for eyewitness testimony. *Applied Cognitive Psychology, 6*, 573-587. doi: 10.1002/acp.2350060702

Hudson, J. A., Fivush, R., & Kuebli, J. (1992). Scripts and episodes. The development of event memory. *Applied Cognitive Psychology, 6*, 483-505. doi: 10.1002/acp.2350060604

Jack, F., Zydervelt, S., & Zajac, R. (2014). Are co-witnesses special? Comparing the influence of co-witness and interviewer misinformation on eyewitness reports. *Memory, 22*, 243-255. doi: 10.1080/09658211.2013.778291

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology, 5*, 55-71. doi: 10.1080/09541449308406514

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*, 226-246. doi: 10.1006/obhd.1994.1013

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition, 106*, 259-298. doi: 10.1016/j.cognition.2007.02.003

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132*, 133-156. doi: 10.1037/0096-3445.132.1.133

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

Karlsson, L., Juslin, P., & Olsson, H. (2008). Exemplar-based inference in multi-attribute decision making: Contingent, not automatic, strategy shifts? *Judgment and Decision Making, 3*, 244-260. Retrieved from: http://www.sjdm.org/~baron/journal/bn5.pdf

Kleider, H. M., Pezdek, K., Goldinger, S. D., & Kirk, A. (2008). Schema-driven source misattribution errors: Remembering the expected from a witnessed event. *Applied Cognitive Psychology, 22*, 1-20. doi: 10.1002/acp.1361

Lane, S. M., & Meissner, C. A. (2008). A 'middle road' approach to bridging the basic-applied divide in eyewitness identification research. *Applied Cognitive Psychology, 22*, 779-787. doi: 10.1002/acp.1482

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159-183. doi: 10.1016/0030-5073(77)90001-0

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*, 149-171. doi: 10.1016/0030-5073(80)90052-5

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology, 7*, 560-572. doi: 10.1016/0010-0285(75)90023-7

Luna, K., & Martín-Luengo, B. (2012). Confidence-accuracy calibration with general knowledge and eyewitness memory cued recall questions. *Applied Cognitive Psychology, 26*, 289-295. doi: 10.1002/acp.1822

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23*, 77-87. doi: 10.1002/ejsp.2420230107

McClure, K. A., Myers, J. J., & Keefauver, K. M. (2013). Witness vetting: What determines detectives' perceptions of witness credibility? *Journal of Investigative Psychology and Offender Profiling, 10*, 250-267. doi: 10.1002/jip.1391

Migueles, M., & García-Bajos, E. (2006). Influence of the typicality of the actions in a mugging script on retrieval-induced forgetting. *Psicologica, 27*, 119-135. Retrieved from: http://eric.ed.gov/?id=EJ803974

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*, 502-517. doi: 10.1037/0033-295X.115.2.502

Ogloff, J. R. P. (2000). Two steps forward and one step backward: The law and psychology movement(s) in the 20th century. *Law and Human Behavior, 24*, 457-483. doi: 10.1023/A:1005596414203

Palmer, M. A., 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. doi: 10.1037/a0031602

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology Public Policy and Law, 1*, 817-845. doi: 10.1037/1076-8971.1.4.817

Perfect, T. J. (2002). When does eyewitness confidence predict performance? In T. J. Perfect & B. L. Schwartz (Eds.). *Applied Metacognition* (pp. 95-120). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511489976.006

Perfect, T. J., Watson, E. L., & Wagstaff, G. F. (1993). Accuracy of confidence ratings associated with general knowledge and eyewitness memory. *Journal of Applied Psychology, 78*, 144-147. doi: 10.1037//0021-9010.78.1.144

Peterson, C., & Grant, M. (2001). Forced-choice: Are forensic interviewers asking the right questions? *Canadian Journal of Behavioural Science (Revue Canadienne Des Sciences Du Comportement), 33*, 118-127. doi: 10.1037/h0087134

Potter, R., & Brewer, N. (1999). Perceptions of witness behaviour-accuracy relationships held by police, lawyers and mock-jurors. *Psychiatry, Psychology and Law, 6*, 97-103. doi: 10.1080/13218719909524952

Rattner, A. (1988). Convicted but innocent: Wrongful convictions and the criminal justice system. *Law and Human Behavior, 12*, 283-293. doi: 10.1007/BF01044385

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*, 416-425. doi: 10.1037/0021-9010.82.3.416

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shaw, J. S., & McClure, K. A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior, 20*, 629-653. doi: 10.1007/BF01499235

Sherman, J. W., Groom, C. J., Ehrenberg, K., & Klauer, K. C. (2003). Bearing false witness under pressure: Implicit and explicit components of stereotype-driven memory distortions. *Social Cognition, 21*, 213-246. doi: 10.1521/soco.21.3.213.25340

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the U.S. population. *Plos One, 6*, 1-7. doi: 10.1371/journal.pone.0022757

Smith, V. L., Ellsworth, P. C., & Kassin, S. M. (1989). Eyewitness accuracy and confidence: Within- versus between-subjects correlations. *Journal of Applied Psychology, 74*, 356-359. doi: 10.1037/0021-9010.74.2.356

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. doi: 10.1037//0033-2909.118.3.315

Stangor, C., & Duan, C. (1991). Effects of multiple task demands upon memory for information about social groups. *Journal of Experimental Social Psychology, 27*, 357-378. doi: 10.1016/0022-1031(91)90031-Z

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin, 111*, 42-61. doi: 10.1037/0033-2909.111.1.42

Towfigh, E. V., & Glöckner, A. (2011). GAME OVER: Empirical support for soccer bets regulation. *Psychology, Public Policy, and Law, 17*, 475-506. doi: 10.1037/a0023402

Tuckey, M. R., & Brewer, N. (2003a). How schemas affect eyewitness memory over repeated retrieval attempts. *Applied Cognitive Psychology, 17*, 785-800. doi: 10.1002/acp.906

Tuckey, M. R., & Brewer, N. (2003b). The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied, 9*, 101-118. doi: 10.1037/1076-898x.9.2.101

Turtle, J., Read, J. D., Lindsay, D. S., & Brimacombe, C. A. E. (2008). Toward a more informative psychological science of eyewitness evidence. *Applied Cognitive Psychology, 22*, 769-778. doi: 10.1002/acp.1481

Tyszka, T., & Zielonka, P. (2002). Expert judgments: Financial analysts versus weather forecasters. *The Journal of Psychology and Financial Markets, 3*, 152-160. doi: 10.1207/S15327760JPFM0303_3

Waubert de Puiseau, B., Assfalg, A., Erdfelder, E., & Bernstein, D. M. (2012). Extracting the truth from conflicting eyewitness reports: A formal modeling approach. *Journal of Experimental Psychology: Applied, 18*, 390-403. doi: 10.1037/a0029801

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence. Improving its probative value. *Psychological Science in the Public Interest, 7*, 45-75. doi: 10.1111/j.1529-1006.2006.00027.x

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54*, 277-295. doi: 10.1146/annurev.psych.54.101601.145028

Footnotes

1) Note that “script-conforming” refers to the items, whereas “script-consistent” refers to the details of the observed crime. The same distinction applies to “script-nonconforming” items and “script-inconsistent” details.

Table 1

Typicality Ratings and Item Difficulties for all Six Item Subsets (M, SD)

Statement category	Statement accuracy	Item subset	*m*	Typicality		Item difficulty	
				M	*SD*	*M*	*SD*
Script-consistent	True	Conforming	17	6.06	0.37	.68	.25
	False	Nonconforming	17	5.99	0.47	.67	.26
Neutral	True	Neutral	17	4.08	0.31	.64	.25
	False	Neutral	17	4.11	0.26	.67	.24
Script-inconsistent	True	Nonconforming	17	2.22	0.56	.70	.27
	False	Conforming	17	2.26	0.54	.67	.26
All statements			102	4.12	1.61	.67	.25

Figure 1. Mean accuracy (in percent) and its standard error by working memory load (between-subjects) and script conformity (within-subjects). Every item set (script-conforming, script-neutral, script-nonconforming) comprised 34 questions.

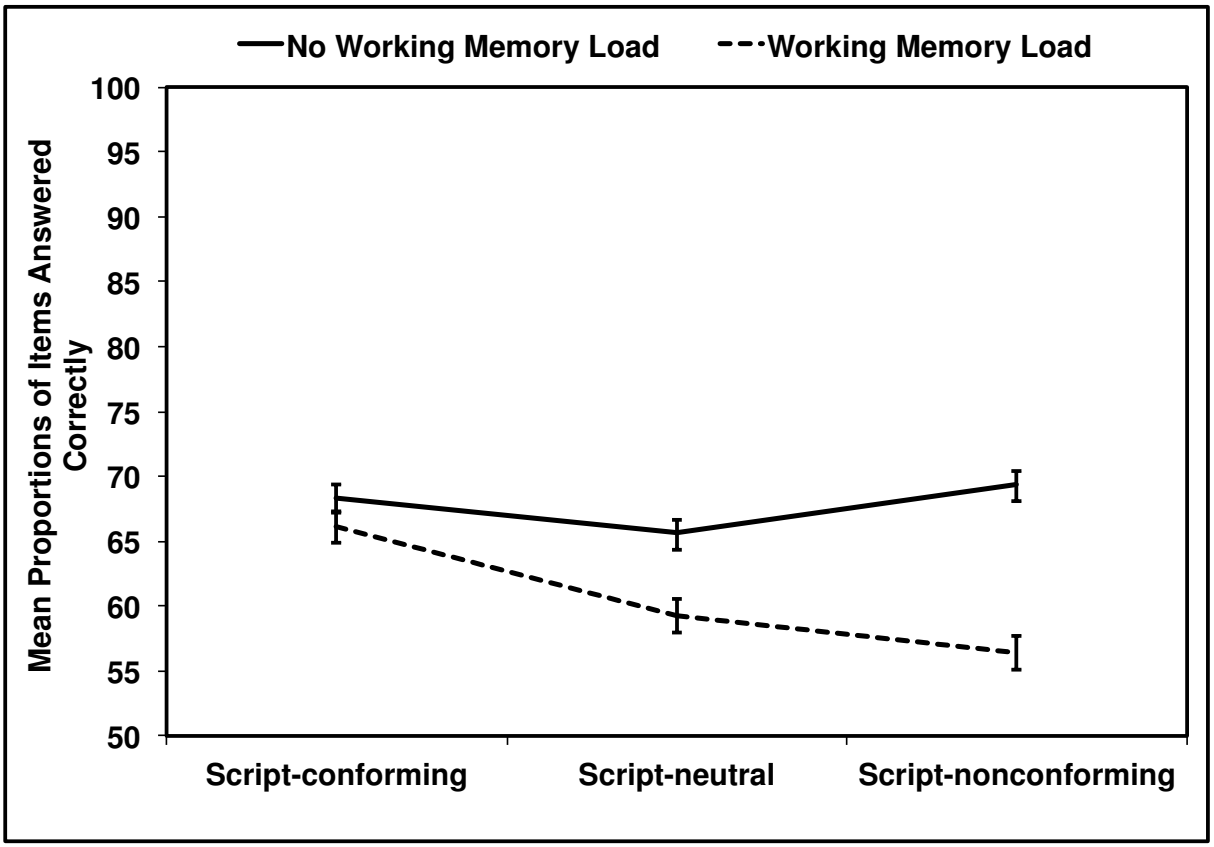


Figure 2. The mean level of confidence and its standard error by working memory load (between-subjects) and script conformity (within-subjects). The confidence scale ranged from 50 to 100. Every question set (script-conforming, script-neutral, script-nonconforming) comprised 34 questions.

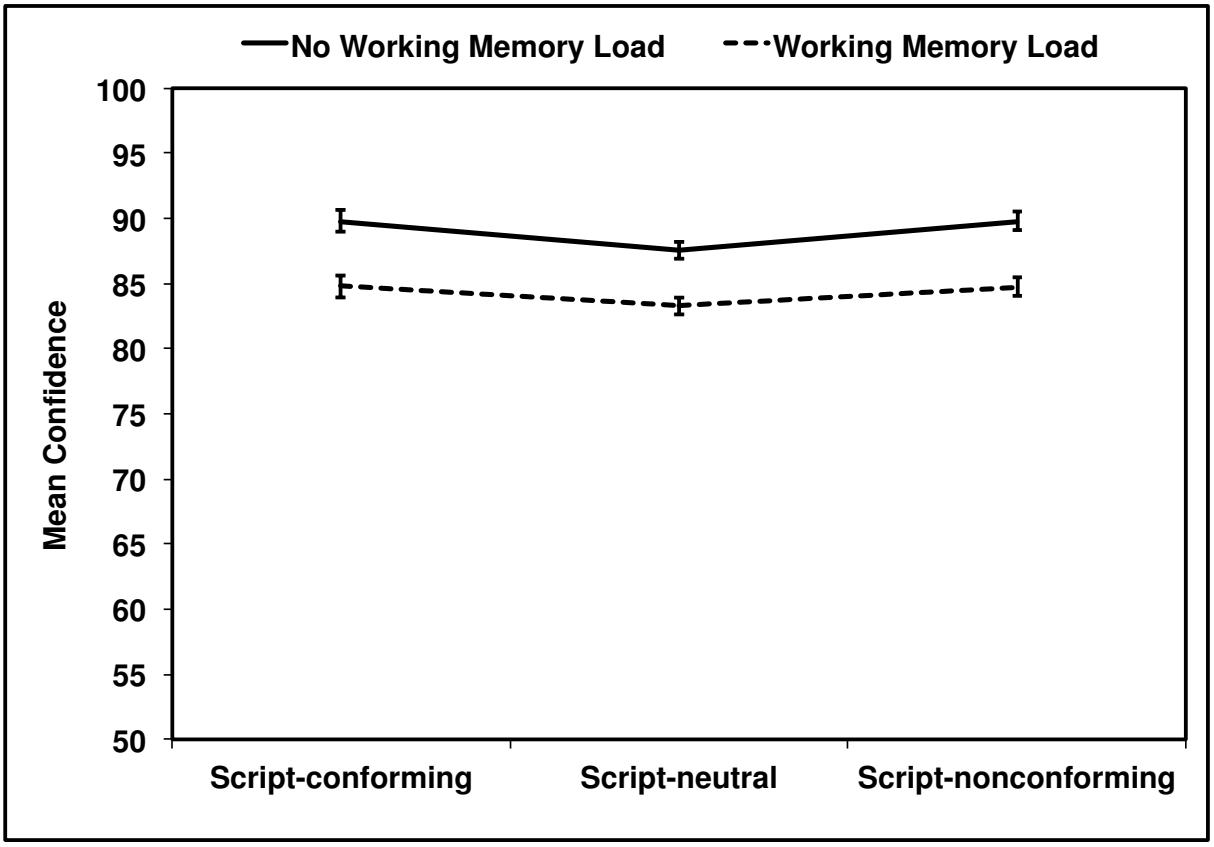
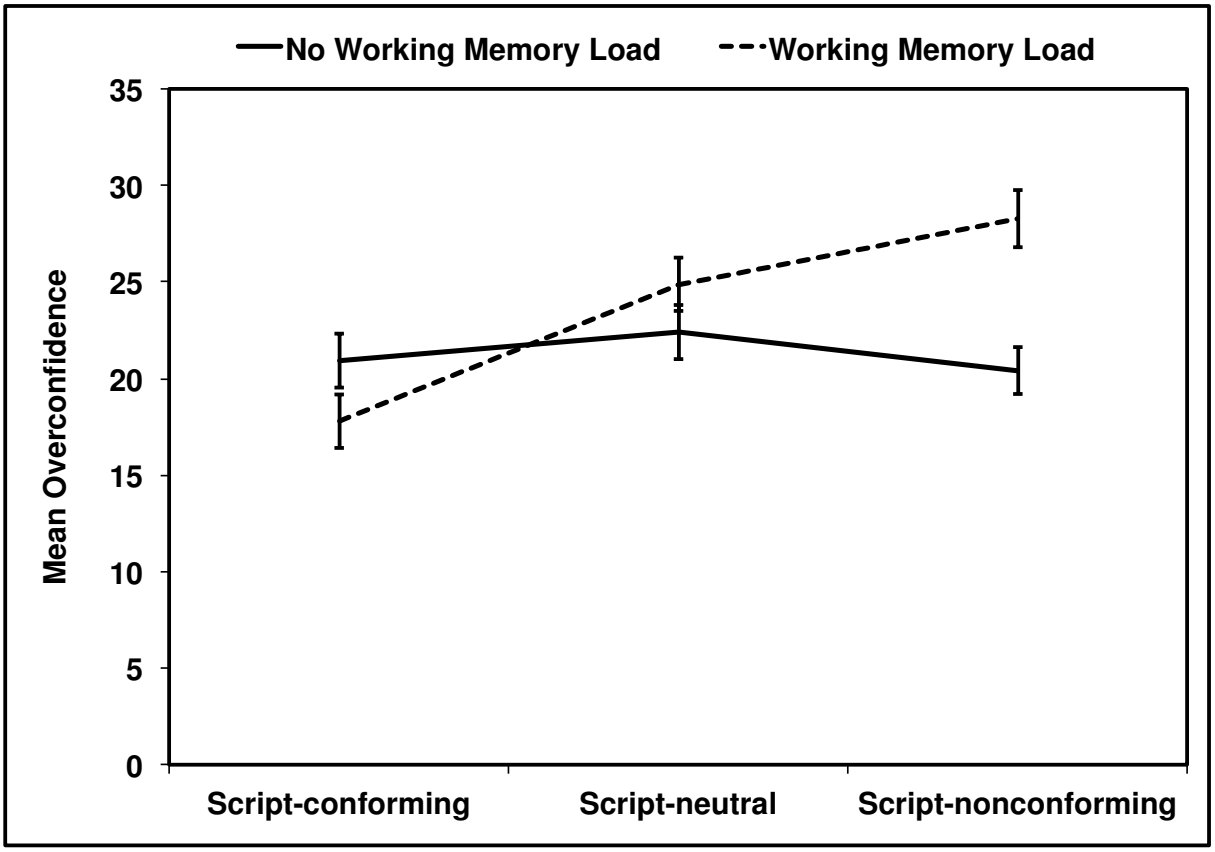


Figure 3. Mean overconfidence and its standard error by working memory load (between-subjects) and script conformity (within-subjects). Overconfidence was computed as the difference between mean confidence and proportion correct with a positive score indicating overconfidence. Every question set (script-conforming, script-neutral, script-nonconforming) comprised 34 questions.



Appendix C

Study 3:

Waubert de Puiseau, B., Platzek, S., Aßfalg, A., & Musch, J. (2016). On the importance of considering heterogeneity in witnesses' competence levels when reconstructing crimes from multiple witness testimonies. *Manuscript submitted for publication.*

On the Importance of Considering Heterogeneity in Witnesses' Competence Levels When Reconstructing Crimes From Multiple Witness Testimonies

Berenike Waubert de Puiseau*[1], Sven Platzek*[1,2], André Aßfalg[3], Jochen Musch[1]

[1]University of Düsseldorf, Germany

[2]University of Kassel, Germany

[3]University of Freiburg, Germany

Word count (excluding Abstract, Tables, Figures, and References): 7,715

«fn»* B. Waubert de Puiseau and S. Platzek contributed equally to this work.

Author Note

Berenike Waubert de Puiseau and Jochen Musch, Department of Experimental Psychology, University of Duesseldorf, Germany; Sven Platzek, Department of Psychology, University of Kassel, Germany; André Aßfalg, Department of Psychology, University of Freiburg, Germany.

Correspondence concerning this article should be addressed to Berenike Waubert de Puiseau, Department of Experimental Psychology, University of Duesseldorf, Universitaetsstrasse 1, Building 23.03, 40225 Duesseldorf, Germany, phone: +49 – (0)211 – 8112063, fax: +49 – (0)211 – 8111753, e-mail: bwdp@hhu.de

Abstract

Aggregating information across multiple testimonies may improve crime reconstructions. However, different aggregation methods are available, and research on which method is best suited for aggregating multiple observations is lacking. Furthermore, little is known about how variance in the accuracy of individual testimonies impacts the performance of competing aggregation procedures. We investigated the superiority of aggregation-based crime reconstructions involving multiple individual testimonies and whether this superiority varied as a function of the number of witnesses and the degree of heterogeneity in witnesses' ability to accurately report their observations. Moreover, we examined whether heterogeneity in competence levels differentially affected the relative accuracy of two aggregation procedures: a simple Majority Rule, which ignores individual differences, and the more complex General Condorcet Model (Romney, Weller, & Batchelder, 1986; Batchelder & Romney, 1988), which takes into account differences in competence between individuals. 121 participants viewed a simulated crime and subsequently answered 128 true/false questions about the crime. We experimentally generated groups of witnesses with homogeneous or hetereogeneous competences. Both the Majority Rule and the General Condorcet Model provided more accurate reconstructions of the observed crime than individual testimonies. The superiority of aggregated crime reconstructions involving multiple individual testimonies increased with an increasing number of witnesses. Crime reconstructions were most accurate when competences were heterogeneous and aggregation was based on the General Condorcet Model. We argue that a formal aggregation should be considered more often when eyewitness testimonies have to be assessed and that the General Condorcet Model provides a good framework for such aggregations.

Word count (Abstract): 249

Keywords: witness testimony; wisdom of the crowd; Cultural Consensus Theory; Majority Rule; heterogeneity

On the Importance of Considering Heterogeneity in Witnesses' Competence Levels When Reconstructing Crimes From Multiple Witness Testimonies

Many crimes are observed by more than just a single witness. Empirical studies have found the number of co-witnesses present at a crime scene to range from only one to more than 100, with a median of around three (Paterson & Kemp, 2006; Skagerberg & Wright, 2008). A well-known example of a large number of co-witnesses is the assassination of US President John F. Kennedy. When Kennedy was shot at a parade in Dallas in 1963, more than 500 attendees of the parade subsequently served as witnesses (President's Commission on the Assassination of President Kennedy, 1964).

Human memory is fallible (Clark & Wells, 2008; Loftus, 1996). When multiple witnesses are interviewed about the same crime, their testimonies may therefore disagree even when all witnesses aim to provide accurate accounts of their observations. Disagreement has even been found to occur with respect to the central aspects of a crime. For example, the witnesses of Kennedy's assassination disagreed on two of the most important details of the crime: the number of shots fired and the assassin's hiding place (President's Commission on the Assassination of President Kennedy, 1964). If at least some witnesses are unable to accurately remember even the core aspects of a crime, efficient means for distinguishing between correct and incorrect recollections are important. However, judging the competence of witnesses (i.e., the probability that their testimony is accurate) is difficult and has been referred to as "one of the biggest challenges in human memory research" (Bernstein & Loftus, 2009, p. 379). Assessment of witness competence may be improved by comparing multiple reports from individual witnesses who testify repeatedly (Fisher, Vrij, & Leins, 2013) or by collecting confidence ratings immediately following each response (e.g., Allwood, Ask, & Granhag, 2005; Roberts & Higham, 2002; Vredeveldt & Sauer, 2015). When witnesses only provide a single report and fail to give confidence ratings, alternative means to assessing their competences are however required.

Research has shown that if a witness is present at all, one or more co-witnesses are usually also available (Paterson & Kemp, 2006; Skagerberg & Wright, 2008). Several studies have investigated how the presence of multiple witnesses improves or impairs the quality of individual testimonies (e.g., Gabbert, Memon, & Wright, 2006; Meade & Roediger, 2002; Vredeveldt, Hildebrandt, & van Koppen, 2015). In contrast, studies aiming to identify predictors of witness competence have focused on individual testimonies, whereas only few studies have taken into account the level of agreement between different witnesses that can inform the reconstruction of crimes and estimates of witness competence. Many empirical studies have shown that aggregate judgments are superior to the judgments of individuals, a phenomenon that in other contexts has been referred to as the *wisdom of the crowd* (Armstrong, 2004; Clemen, 1989; Davis-Stober, Budescu, Dana, & Broomell, 2014; Galton, 1907; Krause, Ruxton, & Krause, 2010; Surowiecki, 2004). Despite its success in other domains of psychological research, the aggregation approach has received little attention in forensic psychology (Clark & Wells, 2008). To our knowledge, only three studies have investigated the usefulness of aggregation when assessing the accuracy of witness testimonies. Two of these studies restricted their analyses to eyewitness identification decisions. One of these studies aggregated identification decisions across groups of three witnesses and found that aggregated identification decisions based on a majority vote among witnesses were more reliable than those of a single witness (Clark & Wells, 2008). A second study found that when groups comprised four or more witnesses, applying a simple rule such as choosing the modal response when determining an aggregate identification decision was sufficient for outperforming individual testimonies (Sanders & Warnick, 1982). A third study investigated eyewitness event memory and revealed that aggregated responses provided more valid crime reconstructions when using an aggregation method that considered estimates of witness competence, compared with an aggregation that was based on simply choosing the majority response (Waubert de Puiseau, Aßfalg, Erdfelder, & Bernstein, 2012).

To summarize, aggregation seems to be a promising approach for improving the reconstruction of a crime and may also be used to assess the competence of witnesses. However, different aggregation methods exist, and it is unknown which method is best suited for reconstructing crimes from individual testimonies. For three reasons, we argue that when aggregating information across multiple witnesses, it is also important to consider the potential level of heterogeneity in the witnesses' competences (cf. Lindsay, Nilsen, & Read, 2000; Loftus, 1996). The first reason is that competence heterogeneity has theoretically and empirically been linked to the superiority of aggregated judgments over individual judgments. Crowd wisdom has been found to increase with diversity of knowledge in the crowd (e.g., Davis-Stober et al., 2014). A second reason is that competence heterogeneity may benefit different aggregation methods to different degrees. In view of Davis-Stober et al.'s (2014) finding that even simple aggregation rules (e.g., choosing the modal response) benefit from competence heterogeneity, it seems plausible that aggregation methods capable of considering differences in competences between witnesses may benefit even more from competence heterogeneity. However, more complex methods of aggregation are based on a number of assumptions. Empirical studies testing the robustness of these methods against violations of the assumptions in the context of eyewitness event memory are lacking.

Third, empirical studies on witness memory usually employ student samples and are mostly conducted in university laboratories under highly standardized conditions. The true degree of heterogeneity in witnesses' competence may be underestimated in such settings (Lindsay et al., 2000). Moreover, studies with highly homogeneous samples likely underestimate the usefulness of aggregating across multiple witness testimonies if aggregation benefits from competence heterogeneity. In forensic psychology, competence homogeneity and environmental invariance in laboratory studies have been suggested to be major reasons for the weak correlation between the confidence and accuracy observed in empirical research on eyewitness identification (cf. Gruneberg & Sykes, 1993; Read, Lindsay, & Nicholls,

1998). Studies using student samples can however provide baseline measures of heterogeneity in witnesses' competence and, thus, can help to investigate the benefit of aggregation. The impact of heterogeneity in competence on measures of witness memory performance can best be examined when a sample that is highly homogeneous is compared with a sample in which large heterogeneity has been successfully induced. It is for this reason that we compared a control group of students that was rather homogeneous in competences with another group of students, for which heterogeneity in competence was experimentally increased.

The goal of the present study was twofold. First, we aimed to add to the currently small body of literature on crowd wisdom in witness testimonies. In particular, we wanted to test the superiority of two aggregation methods over individual testimonies in reconstructing crimes observed by multiple witnesses. To this end, we compared (a) the simple Majority Rule, which provides a *majority reconstruction*, and (b) the Cultural Consensus Theory (Romney, Weller, & Batchelder, 1986), which provides a *consensus reconstruction*. Cultural Consensus Theory can account for individual differences in competence between witnesses when reconstructing a crime and has therefore recently been suggested as a valid aggregation method for eyewitness event memory (Waubert de Puiseau et al., 2012). However, the increased computational effort associated with the use of the Cultural Consensus Theory can be justified only if the consensus reconstruction is more accurate than the majority reconstruction. We therefore aimed to investigate whether a consensus reconstruction improves the superiority of the aggregated response over individual judgments even more than a majority reconstruction. The second goal of the present study was to provide the first empirical investigation of the impact of competence heterogeneity on the relative performances of majority and consensus aggregation in witness testimony.

The paper is structured as follows. We first introduce the two competing aggregation methods used to compute the majority and consensus reconstructions, respectively. We then discuss the potential influence of competence heterogeneity on the performance of these two

aggregation methods. Finally, we present data from a witness experiment in which we manipulated competence heterogeneity. We conducted this experiment to evaluate whether the majority and consensus reconstructions would be superior to individual reports and whether this superiority would be moderated by competence heterogeneity.

Majority Reconstruction

The *majority reconstruction* is usually determined by using the simple Majority Rule and provides an unweighted aggregation of testimonies across witnesses. In the special case of recognition memory for binary events, which is the focus of the present study, the majority reconstruction is defined as the modal response across all witnesses. For example, when the more than 500 witnesses were asked where John F. Kennedy's assassin was hiding, the most frequent response given by about half of the witnesses (46.5%) was that the shots came from the nearby school book depository. About one fifth of the witnesses claimed that the shots had been fired from a grassy knoll (20.2%), whereas the remaining witnesses reported shots from both or other locations (President's Commission on the Assassination of President Kenney, 1964). In this example, the modal response (i.e., the school book depository) would determine the majority reconstruction (cf. Sanders & Warnick, 1982) and was included as the alleged true hiding place in an official report (President's Commission on the Assassination of President Kennedy, 1964).

The Majority Rule is conceptually easy to grasp and is computationally inexpensive. The Majority Rule has been shown to allow a better reconstruction of events than individual statements even if the recollections of a group of witnesses are distorted or not independent of each other (Davis-Stober et al., 2014; Sanders & Warnick, 1982). The accuracy of crime reconstructions based on the Majority Rule increases with the number of witnesses (Grofman, Owen, & Feld, 1983) provided that their average competence (commonly measured as their proportion correct) is above the level of chance (Grofman et al., 1983; Romney & Batchelder, 1999; Sanders & Warnick, 1982). However, the Majority Rule suffers from a number of

shortcomings. For example, the Majority Rule ignores whether a majority is strong (e.g., 99% vs. 1%) or weak (e.g., 51% vs. 49%; Weller, 2007), and in the case of a tie, a majority reconstruction is not defined. Further, the Majority Rule neglects the potentially substantial differences in competence between individuals. As a consequence, responses from a person with high competence are given no more weight than responses from a person with low competence. This can lead to a false majority reconstruction, for example, when the majority is less competent than the minority.

Consensus Reconstruction

If more competent witnesses allow for a more accurate reconstruction of a crime, it is desirable to weigh individual testimonies by the witnesses' competence. However, when the truth is unknown, competence cannot easily be determined, for example, because intuitive judgments of a witness's competence by police officers may suffer from unidentified sources of subjective bias (Lindsay et al., 2000). Because it is difficult to accurately assess individual differences in witness competence, it is desirable to weigh individual testimonies by objectively rather than subjectively determined competences. Cultural Consensus Theory (CCT; Romney et al., 1986) provides an appropriate framework for such an approach because it estimates competence directly from witness testimonies. CCT was originally introduced to investigate the culture of unknown ethnic communities and has since evolved into a standard tool in anthropological research. The theory is based on the assumption that competence produces agreement among individuals, whereas a lack of competence results in stochastically independent (random) responses. Consequently, individual competence can be estimated on the basis of the observed agreement between individual witnesses. Estimates of witnesses' competences inform the *consensus reconstruction*, that is, a type of crime reconstruction based on the General Condorcet Model (Waubert de Puiseau et al., 2012).

For dichotomous-response items, CCT can be formalized as a General Condorcet Model (GCM; Batchelder & Romney, 1986; Karabatsos & Batchelder, 2003; Romney et al.,

1986). The GCM is based on the two-high-threshold model (2-HTM; Bredenkamp & Erdfelder, 1996; Snodgrass & Corwin, 1988). The Appendix includes the mathematical foundations of the 2-HTM and the GCM.

The 2-HTM and the GCM both formally describe the relationship between recognition judgments and witness competence. In the present article, participants were presented with several statements about the crime under investigation. The *answer key* describes which of these statements are true and which are false. A core assumption of the 2-HTM is that when witnesses do not recognize an item, they may guess whether it is true. Importantly, participants may differ in their tendency to guess that an item is true in the absence of recognition. Based on the 2-HTM, the witnesses' competence to correctly judge the statements and their tendency to guess can be estimated. However, whereas the 2-HTM assumes that the answer key to each item is known, the GCM includes the answer key as a latent model parameter that can be estimated along with the witness competence and tendency to guess. Moreover, the GCM accounts for variability in item difficulty.

Estimates for the GCM parameters can be computed given sufficient individual responses from multiple witnesses. The GCM assumes that there is a *common truth*, that is, that all witnesses refer to the same event and make converging descriptions, provided that they correctly remember what they have seen. In more technical terms, the answer key is assumed to be constant across witnesses. This assumption is inherent to any aggregation method. Second, the GCM assumes that the items are *locally independent*. This implies that witnesses' responses are independent of each other once all model parameters have been taken into account (Romney, 1999).

The GCM can be used to reconstruct an event on the basis of eyewitness testimony because there are a number of structural similarities between anthropologists trying to understand a culture and legal experts trying to reconstruct a crime. For example, witnesses are commonly not in perfect agreement with each other, and their competences may differ

widely, if only because some witnesses had a limited view of the crime. More important, the truth is initially unknown both when an anthropologist tries to understand a foreign culture and when a legal expert tries to reconstruct a crime. In one previous study, the GCM was successfully employed to reconstruct a crime from multiple witness testimonies and outperformed a reconstruction that was based on the Majority Rule. However, this study did not compare aggregated with individual responses and did not investigate the impact of the heterogeneity of witness competences on the accuracy of crime reconstructions (Waubert de Puiseau et al., 2012).

Limitations of the Aggregation Approach

Despite their merits, the majority and consensus reconstruction methods have a number of limitations. Both the Majority Rule and the variant of the GCM that we employed in the present study only apply to discrete data. In particular, the variant of the GCM that we used in the present study only applies to dichotomous responses (e.g., true vs. false). Other variants of the GCM however exist that, for example, can also accommodate continuous responses (Anders, Oravecz, & Batchelder, 2014; Batchelder, Kumbasar, & Boyd, 1997; Batchelder & Romney, 1988). Further, conclusions drawn from the witnesses' reports—and by implication from aggregations thereof—are biased if witnesses engage in co-witness talk (Gabbert et al, 2003; Meade & Roediger, 2002; Shaw, Garven, & Wood, 1997) or rely on a script of the crime (Greenberg, Westcott, & Bailey, 1998; Holst & Pezdek, 1992). In other words, if the assumptions of a common truth and local independence are violated, both the Majority Rule and the GCM are expected to produce biased aggregates. Unlike the GCM, the Majority Rule is based on the additional assumptions that all witnesses are equally competent regarding their knowledge of the answer key, that they have equal guessing tendencies and that items are equally difficult. The Majority Rule thus makes stricter assumptions than the GCM. One drawback of the GCM is that it is more difficult to implement and computationally more expensive than the Majority Rule. Various software implementations of

the GCM are however now freely available (Aßfalg & Erdfelder, 2012; Oravecz, Vandekerckhove, & Batchelder, 2014).

The Effect of Competence Heterogeneity on Majority and Consensus Reconstructions

Several studies have investigated the impact of competence heterogeneity on majority reconstructions, but their predictions have differed. In most studies, groups were considered heterogeneous when the competences of the group members were unequal, and a crowd was said to be wise if the majority reconstruction matched an a priori known outcome. With respect to witness testimony, this is equivalent to an accurate crime reconstruction. Grofman et al. (1983) predicted that heterogeneity in competences does not impact the accuracy of majority reconstructions if three conditions are met: (a) mean competence is above the level of chance (i.e., .5 given two answer options), (b) heterogeneity does not affect mean competence, and (c) competences are normally distributed around the mean (cf. also Kazmann, 1973). Kanazawa (1998) formally showed that heterogeneous groups are more likely to choose the correct answer to a binary question if mean individual competences are larger than $(1/2) + (1/2n)$, where n is the number of individuals in the group (cf. also Boland, 1989).

For the GCM, predicting the influence of competence heterogeneity on reconstruction performance is more complex than for the Majority Rule. In principle, choosing the GCM should be advisable when witnesses' competences are heterogeneous and sample size is sufficiently large (Romney et al., 1986). In the only study investigating the impact of competence heterogeneity on consensus reconstruction, Weller (1987) conducted a computer simulation with groups that were homogeneous or heterogeneous in their competences. Consensus reconstructions were unaffected by competence heterogeneity. However, this finding has not been replicated with human participants. Moreover, Weller (1987) employed a restricted version of the GCM, which assumed homogeneous guessing biases and item difficulties. Weighting witness testimonies by estimated competences, guessing biases, and

item difficulties should improve the accuracy of crime reconstructions. When an unrestricted version of the GCM is employed, consensus reconstructions should be more accurate than majority reconstructions if competences are heterogeneous. By contrast, in groups of witnesses with equal competences, weighting should not affect the quality of crime reconstructions.

The Present Study

In the present study, we investigated whether aggregated witness testimonies would outperform individual witness reports in the reconstruction of crimes. Moreover, we compared the performance of the Majority Rule and the GCM in groups of witnesses with homogeneous versus heterogeneous competences. Participants first viewed a video of a simulated crime and subsequently answered true/false questions about the content of the video. In the experimental condition, heterogeneity was induced by selectively impairing performance to various degrees at different stages of the recognition test. We tested two main hypotheses: First, we expected that crowds of witnesses would outperform individual witnesses. Second, we predicted that the advantage of aggregating testimonies would increase with increasing heterogeneity in the witnesses' competences and that this increase would be larger for consensus than for majority reconstructions.

Method

Sample

One hundred twenty-seven psychology undergraduates from the University of Düsseldorf participated in the experiment in exchange for course credit. Five participants were excluded from the analyses because they did not follow the instructions and failed to perform the distractor task that was used to impose a load on working memory. Another participant was excluded because of red/green color blindness, which may have negatively impacted the participant's perception of the stimulus material. Most participants were female (n = 105, 87%), and their ages ranged from 18 to 45 (M = 23.08, SD = 5.21). None of the

subjects were proficient in reading Hebrew (which was necessary because we used Hebrew words to disrupt retrieval as detailed below).

Design and Procedure

Competence heterogeneity (homogeneous vs. heterogeneous) was manipulated between subjects. The goal of the competence heterogeneity manipulation was to generate two samples with similar means but different variances on the measurement of competence. We were thus able to assess whether increasing individual differences between the witnesses' competences influenced the accuracy of the reconstruction of a crime even when witnesses' average competence was held constant (for the relation between competence and aggregation in the Majority Rule and CCT, see Grofman et al., 1983, and Weller, 1987, respectively). We aimed for competences in the heterogeneous condition to be symmetrically distributed around the mean. Therefore, we created 27 cells by manipulating three experimental between-subjects factors to be orthogonal and to have three levels each. To generate similar means in the heterogeneous (n = 67) and homogeneous (n = 54) conditions, all participants in the homogeneous condition participated in the cell that constituted the middle level on all three factors that were manipulated in the heterogeneous condition. Participants were allocated randomly to either the homogeneous or the heterogeneous condition with the constraint that participants in the heterogeneous condition were randomly and evenly distributed across the 27 cells that were manipulated in this condition.

Three manipulations were applied in the heterogeneous condition to increase the variance in the witnesses' competences. The first competence manipulation affected the amount of information a participant could visually extract from the video. This manipulation simulated that witnesses have different viewing angles and distances to the crime scene, and that witnesses likely differ in their cognitive capacities available to perceive and encode their observations, all of which likely introduces variability in the validity of the individual witnesses' testimonies. To decrease the competence of some of the witnesses, parts of the

video were blurred to obscure details. Three factor levels were employed: (a) no blurring at all, (b) blurring of a random 15% of the video, or (c) blurring of a random 30% of the video. For blurring, the video was divided into 100 segments of 3 s each. For all 100 parts, we generated the blurry version by dividing each frame of the video into a 12x16 matrix of equally sized rectangles. The color values of all pixels were then set to the average color value of the rectangle the pixel belonged to. To generate multiple versions of a video showing the same crime (as was necessary to satisfy the assumption of a shared pool of knowledge underlying the Majority Rule and the GCM), we randomly replaced 0, 15, or 30 of the 100 parts of the video with their blurry version.

The second competence manipulation involved the application of a distractor task during the video. With this task, we aimed to impair the encoding of the video content (Troyer & Craik, 2000). For the duration of the video, participants either (a) completed no additional exercise at all, (b) repeatedly wrote down a 4-digit number, or (c) performed a complex calculation exercise in which they had to write down the results of each calculation without being permitted to take notes for intermediate calculation steps. This complex exercise comprised two steps: Participants were presented with a 4-digit number and asked to subtract the last digit from the second-to-last digit. Next, participants were asked to subtract this difference from the original 4-digit number. If the first difference was zero, participants were instructed to subtract 1 from the original 4-digit number to ensure that the task would not terminate prematurely. Participants were then instructed to repeat the previous two steps until the video ended.

The third competence manipulation was aimed at disrupting memory retrieval during the recognition test. Using a technique introduced by Vredeveldt, Hitch, and Baddeley (2011), we presented Hebrew characters for 1 s each in a random location on the screen. We made sure that the visual presentation of the Hebrew characters did not interfere with the visual presentation of the items or the response options. We manipulated the number of distortions

by showing Hebrew characters (a) never, (b) for a random 50% of the items, or (c) for all 128 items.

After giving informed consent and answering several demographic questions, participants put on headphones for the auditory part of the video. All participants first completed trial runs of the distractor task. Upon successfully completing the trial runs, they were presented with the video of the simulated crime. Following a procedure detailed in Waubert de Puiseau et al. (2012), the experiment paused after the video had ended and the participants were given 2 min to mentally recapitulate the crime they had just seen. Participants were then given a surprise recognition test with 128 true/false questions about the video content. Subjects chose *true* or *false* by clicking on the respective button in the lower half of the screen.[1] Upon completing all items, participants were thanked and debriefed.

Material

Video. The video showed a re-enactment of a bank robbery that was taken from a German TV show on unresolved crimes. The video consisted of two parts. In the first part, two people, a woman and a male police officer, are in an office at a police station. The woman reports her car as stolen, and the police officer poses several questions about the theft. The second part depicts a bank robbery. The car that was previously reported stolen stops in front of the building of the branch bank. Two masked men with large guns leave the car and enter the bank that they then rob at gunpoint. Afterwards, the robbers flee the crime scene in the stolen car. The video had a total length of about five minutes.

Items. The item set comprised 128 items. Each item consisted of a statement about the video content and the two answer options *true* or *false*. The items targeted different aspects of the video, including the appearance of the people who were involved (e.g., “The bank tellers wear name tags”), actions (e.g., “One of the robbers turns off the surveillance camera”),

[1] In addition to the true/false responses, participants rated their confidence with respect to each response. This was done for an unrelated study that is not part of the present article.

spoken conversation (e.g., "The robbers threaten the bank tellers by saying: 'Hands up, this is a robbery!'"), surroundings (e.g., "During the robbery, there is a white Audi parked in front of the bank"), and objects (e.g., "There are small bags from the German Federal Bank in the vault"). All items were presented in chronological order, corresponding to the events in the video.

Results

All analyses were implemented with the R statistics software (R Development Core Team, 2014). For the computation of the GCM parameter estimates, we used computer code that was generously made available by George Karabatsos (Karabatsos & Batchelder, 2003).[2]

Descriptive Results and Manipulation Check

The mean proportion of correct responses was slightly lower in the homogeneous condition (61%, SD = 5%) than in the heterogeneous condition (64%, SD = 7%), $t(118.80) = 2.38$, $p = .019$, $d = 0.49$. Witness competences and guessing biases were assessed using the 2-HTM. Mean witness competence differed slightly between the two experimental conditions with participants in the heterogeneous group exhibiting more competence ($D = .27$, $SD = .13$) than participants in the homogeneous group ($D = .23$, $SD = .10$), $t(118.76) = 2.38$, $p = .019$, $d = 0.42$. An F test for equality of variances was computed to test the difference in the variability in competence between the two conditions. As intended, heterogeneity in the competence levels was significantly larger in the heterogeneous condition than in the homogeneous condition, $F(53, 66) = 1.68$, $p = .026$.

As the descriptive results showed, witnesses in the homogeneous condition were on average less competent than witnesses in the heterogeneous condition. Previous studies however found that aggregation outcomes are influenced by differences in competence (e.g., Batchelder & Romney, 1988). To ensure a fair comparison between the homogeneous and the

[2] Parameter estimates were based on 11,000 iterations, of which the first 1,000 iterations were used as burn-ins and therefore discarded.

heterogeneous condition, we simulated groups of witnesses in a series of random sampling analyses.

Random Sampling Analyses

In random sampling analyses, we drew pairs of samples of different sizes (n = 10, 20, and 40) by randomly choosing participants from the homogeneous and heterogeneous conditions, respectively. By comparing the samples thus generated, we tested whether there were differences in the validity of the crime reconstructions between the two aggregation procedures. We were thus able to examine the influence of sample size on aggregation outcomes. The random sampling also ensured equal mean competences in the homogeneous and heterogeneous samples, and thus allowed for a fair comparison. This was important because previous studies have found that aggregation outcomes are influenced by differences in competence (e.g., Batchelder & Romney, 1988). We therefore forced the sampling algorithm to draw sample pairs that were matched in competence (means were allowed to differ only in the third decimal place). Thus, any differences in accuracy between the majority reconstruction and the consensus reconstruction could be attributed to the experimental manipulation of competence heterogeneity. We also made sure that in each sample pair, the variance of competences in the heterogeneous sample significantly exceeded the variance of competences in the corresponding homogeneous sample according to an F test. To this end, we drew 67 sample pairs for each of the three sample sizes (n = 10, 20, and 40). This was the number of sample pairs that was sufficient for performing the required sampling procedures in less than 48 hr per sample size condition while still allowing us to achieve a large power of .89 for a McNemar test that was computed to compare the proportions of correctly reconstructed crime details using the majority and consensus reconstructions, and to compare

the respective proportions in the homogeneous and the heterogeneous samples, respectively.[3] The sampling algorithm ensured that no sample pair was drawn more than once.

To investigate whether our sampling was successful, we checked whether variances in competence estimates that were based on the GCM were larger in the heterogeneous than in the homogeneous samples. *T* tests confirmed that for all three samples sizes (n = 10, 20, and 40), the standard deviations were significantly larger in the heterogeneous than in the homogeneous groups (all ts $\geq$ 10.72, all ps $<$ .001).

On the basis of existing theoretical accounts, we expected the Majority Rule to provide a better reconstruction of the crime with increasing sample size (cf. Grofman et al., 1983; Kazmann, 1973). Previous studies found that the performance of the GCM also improved with increasing sample size (Batchelder & Romney, 1988; Romney et al., 1986; Waubert de Puiseau et al., 2012; Weller, 1987, 2007). We therefore expected more reliable crime reconstructions for larger sample sizes when aggregations were based on the GCM. Because the validity of the individual crime reconstructions was not predicted to be affected by sample size, we did not expect the magnitude of the superiority of the majority and consensus reconstructions over the individual testimonies to vary as a function of sample size.

Aggregated versus Individual Reconstruction. To assess the performance of the Majority Rule, we computed an estimate of the answer key (*majority reconstruction*) and compared it with the actual answer key, that is, the true and known course of events. To evaluate the performance of the GCM[4], we first compared the answer key with an estimate of the answer key computed with the GCM (*consensus reconstruction*). We further compared

[3] In estimating the statistical power, we assumed an odds ratio of 3 and a proportion of discordant pairs of .55. The odds ratio is determined by the ratio of the two cells in the 2x2 table in which the aggregation methods did not perform equally well.

[4] To determine whether all model parameters were needed to explain the observed data, we computed the badness-of-fit Deviance Information Criterion (DIC; cf. Karabatsos & Batchelder, 2003) for the GCM. In both conditions, the most complex variant of the GCM showed the best trade-off between model fit and the number of parameters and was therefore used in all analyses.

the GCM competence estimates[5] with the corresponding 2-HTM competence estimates. Note that the GCM estimates rely on the participants' responses alone, whereas the 2-HTM also requires knowledge of the answer key. Finally, to assess the wisdom of the crowd, we compared the validity of the majority and consensus reconstructions with the average validity of the individual reports (*individual reconstruction*).

The majority and consensus reconstructions were more accurate than the individual reconstructions for all sample sizes (Figures 1a and 1b). In line with our expectations, we found that only the consensus but not the majority reconstructions increased in accuracy when the competences were heterogeneous. The consensus reconstructions were more accurate than the majority reconstructions when the competences were heterogeneous (see Figure 1b) but not when the competences were homogeneous (see Figure 1a), suggesting that the GCM suffered from competence homogeneity. The most accurate crime reconstructions were obtained when competences were heterogeneous and aggregation was based on the GCM. The accuracies of both the majority and consensus reconstructions increased with sample size. The average validity of the individual reports did not increase with the number of witnesses.

- Please insert Figure 1 about here -

In addition, to examine the competence estimates provided by the GCM, we computed correlations between the parameter estimates using the GCM and the 2-HTM (i.e., competences based on the true answer key) for all samples and averaged the resulting coefficients across each combination of sample size and condition (Figure 2). The mean correlation coefficients and proportions of significant correlations increased with increasing sample sizes. However, this pattern occurred only for the heterogeneous samples and not for

[5] The GCM further considers differences in guessing bias and item difficulty. However, because these parameters were not important for present purposes, we do not discuss them any further.

the homogeneous samples. For the homogeneous samples, the competence estimates that were based on the GCM and the true competences were significantly correlated in only a few of the 3*67 samples; more precisely, such a correlation was found to be significant in 11, 18, and 4 of the 67 samples for sample sizes of n = 10, 20, and 40, respectively. The respective numbers were 59, 66, and 67 out of 67 for the heterogeneous samples.

- Please insert Figure 2 about here -

Majority Reconstruction versus Consensus Reconstruction. Separately for each sample size, we assessed whether the consensus reconstruction was more accurate than the majority reconstruction. We assessed for each sample pair whether this was the case only in the homogeneous sample, only in the heterogeneous sample, in both samples, or in neither sample. This resulted in a dependent 2 x 2 matrix containing the validities of the majority and consensus reconstructions for both the homogeneous and heterogeneous conditions (Tables 1a-1c).

To assess whether the advantage of the consensus reconstructions over the majority reconstructions significantly depended on competence heterogeneity, we computed McNemar tests for each of the three sample sizes (n = 10, 20, and 40). To avoid inflating the α error, we used a Bonferroni-corrected significance level of $\alpha_{cor} = \alpha / 3 = .0167$. As the results depicted in Tables 1a-1c show, the McNemar test results were significant for all sample sizes. The consensus reconstructions were more accurate than the majority reconstructions significantly more often in the samples drawn from the heterogeneous condition. The number of sample pairs in which the GCM outperformed the Majority Rule in the heterogeneous but not in the homogeneous samples was significantly larger than the number of sample pairs in which the GCM outperformed the Majority Rule in the homogeneous but not in the heterogeneous

samples. A clear link between competence heterogeneity and the (relative) accuracy of the consensus reconstructions was thus established.

- Please insert Table 1 about here -

Discussion

We investigated whether the wisdom of the crowd could be tapped by interviewing multiple witnesses to improve crime reconstructions. More specifically, we determined how well crimes could be reconstructed from multiple witness statements on the basis of two competing methods of aggregating testimonies, the simple Majority Rule and the Cultural Consensus Theory (CCT) as formalized in the General Condorcet Model (GCM). We manipulated the degree of heterogeneity in witnesses' competence and sample size and determined the extent to which reconstructions that were based on individual responses were outperformed by reconstructions that were based on aggregations, that is, by the majority reconstruction based on the Majority Rule and the consensus reconstruction based on the GCM.

Three conclusions are supported by our results. First, in line with our prediction, groups of witnesses reconstructed crimes more accurately than individual witnesses. This implies that when there are multiple witnesses to a crime, the legal fact-finding process can be improved by considering aggregated reconstructions. Aggregated responses were superior to individual responses regardless of the aggregation method employed.

Our second main conclusion refers to the impact of sample size on the aggregated outcomes. Aggregated responses were superior to individual responses for all samples sizes that were considered, and, in line with our prediction, the wisdom of the crowd increased with increasing sample size. This was true for both the majority reconstruction and the consensus reconstruction.

Third, in line with our expectations, heterogeneity in the competence levels differentially affected the majority and consensus reconstructions. The consensus reconstruction clearly benefited from competence heterogeneity. In the heterogeneous groups, the correlations between the competence estimates that were based on the GCM and the true competences were higher than in the homogeneous groups. Moreover, the consensus reconstruction was more accurate than the majority reconstruction regardless of the sample size when the competences were heterogeneous. This was not true for the homogeneous samples, for which we even observed a tendency for the majority reconstruction to be more accurate than the consensus reconstruction. One potential explanation for this finding is that when competence heterogeneity is low, there is a larger proportion of error variance that is fit by the GCM. The validity of the consensus reconstruction may thus be impaired.

In contrast to the consensus reconstruction, the majority reconstruction was only marginally affected by heterogeneity. Thus, if we look back at Grofman et al.'s (1983) and Kazmann's (1973) prediction that heterogeneity would not affect the accuracy of the majority reconstruction and Kanazawa's (1998) prediction that heterogeneity would affect it, it is interesting to note that neither prediction was clearly supported. More research is needed to determine the impact of competence heterogeneity on the performance of simple aggregation rules. As the present study revealed, the accuracy of aggregated crime reconstructions was maximized when competences were heterogeneous and individual differences were accounted for.

A number of limitations of the aggregation approach have to be mentioned. When individuals have no competence and therefore can provide only random answers, aggregation cannot be beneficial to the fact-finding process (cf. Sanders & Warnick, 1982). As Clark and Wells (2008) put it: "If a response is nondiagnostic, little truth can emerge by simply having more of them" (p. 418). When no competent witnesses are available, however, testimony from a single witness would also be of limited use to the fact-finding process. In the present study,

the average proportion correct was around 61% and therefore not much above the level of chance. Nevertheless, aggregation still largely improved the crime reconstructions, suggesting that even with witnesses with low competence, aggregating individual reports may still be useful. Aggregation models also have to rely on the assumption that all witnesses respond independently of each other. Any relation between witness reports that is not based on knowledge may seriously distort aggregated reconstructions of crimes (cf. Clark & Wells, 2008; Waubert de Puiseau et al., 2012). Examples of such influences include scripts that individuals hold of crimes (Greenberg et al., 1998; Holst & Pezdek, 1992), wrongful information obtained through discussions with fellow witnesses (cf. Gabbert et al., 2003; Meade & Roediger, 2002; Shaw et al., 1997), and leading questions posed by interviewers (cf. Loftus, 1975; Sharman & Powell, 2012). Aggregation methods ignoring witness and item characteristics such as the Majority Rule have been proven to be fairly robust against the interdependence of individual responses (Davis-Stober et al., 2014; Estlund, 1994; Ladha, 1992). Because the GCM is also based on the assumption that responses are locally independent, the consensus reconstruction may be vulnerable to such systematic distortions (Waubert de Puiseau et al., 2012).

Aggregated reconstructions rely only on the witnesses' responses. Aggregation thus neglects other information that may be indicative of the accuracy of witnesses' reports such as confidence ratings assessed immediately after each individual response (Allwood et al., 2005; Roberts & Higham, 2002; Vredeveldt & Sauer, 2015). Interestingly, a recent study has shown that majority opinions can be false (Koriat, 2012). This study however employed a simple aggregation rule. Future studies may want to combine various measures of witnesses' competence, such as confidence ratings and competence estimates based on the GCM, to further improve the reconstruction of crimes.

For samples smaller than 10 witnesses, it was not possible to randomly draw a sufficiently large number of sample pairs that complied with the restriction of equal means

and unequal standard deviations on competence, a requirement that was however necessary to create groups that differed in competence homogeneity. Therefore, we used groups of 10 witnesses each as the minimum sample size. This exceeds the median number of witnesses that have usually been found to be present at a crime scene (Paterson & Kemp, 2006; Skagerberg & Wright, 2008). However, previous research has already established that aggregated reconstructions are more accurate than individual testimonies even in small samples (Clark & Wells, 2008; Sanders & Warnick, 1982; Waubert de Puiseau et al., 2012).

The present study was restricted to true/false questions. This type of question is commonly used in forensic interviewing despite recommendations to use other item types such as cued or free recall (Fisher, Geiselman, & Raymond, 1987; Peterson & Grant, 2001). Different variants of the GCM have been proposed that are capable of aggregating more complex types of data such as continuous responses (see for example Anders, Oravecz, & Batchelder, 2014; Batchelder, Kumbasar, & Boyd, 1997; Batchelder & Romney, 1988). Future research should therefore investigate whether the present findings can be generalized to other question formats that are more favorable in legal practice.

The present study used a student sample. As several studies have pointed out, student samples simulating eyewitnesses tend to be more homogeneous in competences than samples from the general population. Employing a student sample however provided us with the required baseline measure to draw causal inferences regarding the impact of heterogeneity in competences on the advantage of aggregated over individual witness reports. It seems likely that aggregating witness reports may be even more useful in samples drawn from the general population at large that can be expected to be more heterogeneous in competences and more representative of real witnesses.

Practical Implications

Research has demonstrated that lay people, including potential jurors but also legal professionals, commonly hold false beliefs about how witness memory works (Brigham &

Bothwell, 1983; Frenda, Nichols, & Loftus, 2011; Schmechel, O'Toole, Easterly, & Loftus, 2006; Simons & Chabris, 2012; Wells, Memon, & Penrod, 2006). Given that eyewitness memory has generally been found to have a tremendous influence on judicial decision making (cf. Simons & Chabris, 2011), it is not surprising that false witness testimony has been identified as a major contributor to wrongful convictions that were later overturned by the application of DNA evidence (for some exemplary cases, see Scheck, Neufeld, & Dwyer, 2000). Improving the assessment of witness competence and the accuracy of crime reconstructions is, therefore, of central importance to legal justice.

The present study demonstrates the general benefit of aggregation in assessing testimony accuracy. It also provides empirical support for a technique that is important to forensic psychology, because when witnesses are present at all, it is highly likely that co-witnesses are also available. Two features of aggregation are particularly important. First, discrepancies between individual reports and, thus, the unreliability of memory are more visible in aggregations than in individual reports; unanimous reports are much less likely when more than four witnesses are present (Sanders & Warnick, 1982). The introductory example of the assassination of John F. Kennedy can be used to illustrate this implication. When interrogated, witnesses of the assassination reported hearing between two and six shots (President's Commission on the Assassination of President Kennedy, 1964). This large range is surprising given that most of the people attending presidential parade had most likely focused their attention on the victim of the crime, thus creating ideal conditions for an accurate encoding of their observations. The fact that considerable discrepancies nevertheless occurred between testimonies led the fact-finders to carefully scrutinize the statements of all witnesses.

A second important feature of aggregation is that aggregated reconstructions may help reduce the number of judicial errors. It has been demonstrated that there always exists an aggregation rule that makes a crowd wiser than the individuals comprising the group (Davis-

Stober et al., 2014). This superiority of aggregated reconstructions was confirmed in the present study. We therefore recommend that the aggregation approach be used more often to reconstruct crimes and to obtain estimates of witnesses' competences. Due to (a) its superiority over majority reconstructions, (b) the structural similarities between anthropological research and crime observations, and (c) the likely heterogeneity in the competences of actual witnesses, the GCM seems to provide a particularly appropriate aggregation rule for witnesses' event memory.

Compliance with Ethical Standards

All procedures performed in the study reported in this manuscript were in accordance with the 1964 Helsinki declaration and its later amendments. Informed consent was obtained from all individual participants included in the study. The authors declare that they have no conflict of interest.

References

Allwood, C. M., Ask, K., & Granhag, P. A. (2005). The Cognitive Interview: Effects on the realism in witnesses' confidence in their free recall. *Psychology, Crime & Law, 11*(2), 183–198.

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve naming problems. *Journal of Memory and Language, 53*(1), 60-80.

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural conseus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology, 61,* 1-13.

Armstrong, J. S. (2004). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting. A handbook for researchers and practitioners* (pp. 417-439). Boston: Kluwer.

Aßfalg, A., & Erdfelder, E. (2012). CAML – Maximum likelihood consensus analysis. *Behavior Research Methods, 44*(1), 189-201.

Bailenson, J. N., Shum, M. S., Atran, S., Medin, D. L., & Coley, J. D. (2002). A bird's eye view: Biological categorization and reasoning within and across cultures. *Cognition, 84*(1), 1-53.

Barg, F. K., Huss-Ashmore, R., Wittink, M. N., Murray, G. F., Bogner, H. R., & Gallo, J. J. (2006). A mixed-methods approach to understanding loneliness and depression in older adults. *Journal of Gerontology B: Social Sciences, 61*(6), 329-339.

Batchelder, W. H., Kumbasar, E., & Boyd, J. P. (1997). Consensus analysis of three-way social network data. *Journal of Mathematical Sociology, 22*(1), 29-58.

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (pp. 103-112). Greenwich, CT: JAL.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika, 53*(1), 71-92.

Batchelder, W. H., & Romney, A. K. (1989). New results in test theory without an answer key. In E. E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 229-248). Berlin, Heidelberg: Springer.

Bernstein, D. M., & Loftus, E. F. (2009). How to tell if a particular memory is true or false. *Perspectives on Psychological Science, 4*(4), 370-374.

Boland, P. J. (1989). Majority systems and the Condorcet Jury Theorem. *The Statistician, 38*(3), 181-189.

Bornstein, B. H., & Zickafoose, D. J. (1999). "I know I know it, I know I saw it": The stability of the confidence-accuracy relationship across domains. *Journal of Experimental Psychology: Applied, 5*(1), 76-88.

Bredenkamp, J., & Erdfelder, E. (1996). Methoden der Gedächtnispsychologie [Methods of the psychology of memory]. In D. Albert & K.-H. Stapf (Eds.), Gedächtnis (Enzyklopädie der Psychologie, Themenbereich C, Serie II, Band 4, S. 1–94)

[Memory (Encyclopedia of Psychology, Topics C, Series II, Issue 4, pp.1-94)]. Göttingen: Hogrefe.

Brigham, J. C., & Bothwell, R. K. (1983). The ability of prospective jurors to estimate the accuracy of eyewitness identifications. *Law and Human Behavior, 7*(1), 19-30.

Clark, S. E., & Wells, G. L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior, 32*(5), 406-422.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

Crowther, C. S., Batchelder, W. H., & Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logic model of perception. *Psychological Review, 102*(2), 396-408.

Davis-Stober, C., Budescu, D., Dana, J., & Broomell, S. (2014). When is a crowd wise? *Decision, 1*(2), 1-4.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior, 28*(6), 687-706.

Estlund, D. M. (1994). Opinion leaders, independence, and Condorcet's Jury Theorem. *Theory and Decision, 36*(2), 131-162.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Fisher, R. P., Geiselman, R. E., & Raymond, D. S. (1987). Critical analysis of police interview techniques. *Journal of Police Science and Administration, 15*(3), 177-185.

Fisher, R. P., Vrij, A., & Leins, D. A. (2013). Does testimonial inconsistency indicate memory inaccuracy and deception? Beliefs, empirical research, and theory. In B. S.

Cooper, D. Griesel, & M. Ternes (Eds.), *Applied Issues in Investigative Interviewing, Eyewitness Memory, and Credibility Assessment* (pp. 173–189). New York, NY: Springer New York.

Frenda, S. J., Nichols, R. M., & Loftus, E. F. (2011). Current issues and advances in misinformation research. *Current Directions in Psychological Science, 20*(1)*,* 20-23.

Gabbert, F., Memon, A., & Allan, K. (2003). Memory conformity: Can eyewitnesses influence each other's memories for an event? *Applied Cognitive Psychology, 17*(5), 533-543.

Gabbert, F., Memon, A., & Wright, D. B. (2006). Memory conformity: Disentangling the steps toward influence during a discussion. *Psychonomic Bulletin & Review, 13*(3), 480-485.

Galton, F. (1907). Vox populi. *Nature Photonics, 75*(1949), 450-451.

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

Greenberg, M. S., Westcott, D. R., & Bailey, S. E. (1998). When believing is seeing: The effect of scripts on eyewitness memory. *Law and Human Behavior, 22*(6), 685-694.

Grofman, B., Owen, G., & Feld, S. (1983). Thirteen theorems in search of the truth. *Theory and Decision, 15*(3), 261-278.

Gruneberg, M. M., & Sykes, R. B. (1993). The generalisability of confidence-accuracy studies in eyewitnessing. *Memory, 1*(3), 185-189.

Harmon, L., & Julesz, B. (1973). Masking in visual recognition: effects of two-dimensional filtered noise. *Science, 180*(4091), 1194-1197.

Hollins, T. S., & Perfect, T. J. (1997). The confidence-accuracy relation in eyewitness event memory: The mixed question type effect. *Legal and Criminological Psychology, 2*(2), 205-218.

Holst, V. F., & Pezdek, K. (1992). Scripts for typical crimes and their effects on memory for eyewitness testimony. *Applied Cognitive Psychology, 6*(7), 573-587.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America, 101*(46), 16385-16389.

Kanazawa, S. (1998). A brief note on a further refinement of the Condorcet Jury Theorem for heterogeneous groups. *Mathematical Social Sciences, 35*(1), 69-73.

Karabatsos, G., & Batchelder, W. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika, 68*(3), 373-389.

Kazmann, R. G. (1973). Democratic organization: A preliminary mathematical model. *Public Choice, 16*(1), 17-26.

Koriat, A. (2012). When are two heads better than one and why? *Science, 336*(6079), 360-362.

Krause, J., Ruxton, G. D., & Krause, S. (2010). Swarm intelligence in animals and humans. *Trends in Ecology and Evolution, 25*(1), 28-34.

Ladha, K. K. (1992). The Condorcet Jury Theorem, free speech, and correlated votes. *American Journal of Political Science, 36*(3), 617-634.

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior, 24*(6), 685-697.

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science, 9*(3), 215-218.

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

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

Meade, M. L., & Roediger, H. L. (2002). Explorations in the social contagion of memory. *Memory & Cognition, 30*(7), 995-1009.

Oravecz, Z., Vandekerckhove, J., & Batchelder, W. H. (2014). Bayesian Cultural Consensus theory. *Field Methods, 26*(3), 207-222.

Paterson, H. M., & Kemp, R. I. (2006). Co-witness talk: A survey of eyewitness discussion. *Psychology, Crime & Law, 12*(2), 181-191.

Perfect, T. J., & Hollins, T. S. (1996). Predictive feeling of knowing judgements and postdictive confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology, 10*(5), 371-382.

Peterson, C., & Grant, M. (2001). Forced-choice: Are forensic interviewers asking the right questions? *Canadian Journal of Behavioural Science (Revue Canadienne Des Sciences Du Comportement), 33*(2), 118-127.

President's Commission on the Assassination of President Kennedy. (1964). Report of the President's Commission on the Assassination of President Kennedy. Washington, DC: U.S. Government Printing Office. Retrieved from http://www.archives.gov/research/jfk/warrencommission-report/letter.html

R Development Core Team. (2014). The R-project for statistical computing. Retrieved from http://www.r-project.org/

Read, J. D., Lindsay, D. S., & Nicholls, T. (1998). The relation between confidence and accuracy in eyewitness identification studies: Is the conclusion changing? In C. P. Thompson, D. J. Herrmann, J. D. Read, & D. Bruce (Eds.), *Eyewitness memory: Theoretical and applied perspectives* (pp. 107-130). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Roberts, W. T., & Higham, P. A. (2002). Selecting accurate statements from the cognitive interview using confidence ratings. *Journal of Experimental Psychology: Applied, 8*(1), 33-43.

Romney, A. K. (1999). Consensus as a statistical model. *Current Anthropology, 40*(S1), 103-115.

Romney, A. K., & Batchelder, W. H. (1999). Cultural Consensus Theory. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 208-209). Cambridge, Mass: MIT Press.

Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus theory. *American Behavioral Scientist, 31*(2), 163-177.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist, 88*(2), 313-338.

Sanders, G. S., & Warnick, D. H. (1982). Evaluating identification evidence from multiple eyewitnesses. *Journal of Applied Social Psychology, 12*(3), 182-192.

Scheck, B., Neufeld, P., & Dwyer, J. (2000). *Actual innocence: Five days to execution and other dispatches from the wrongly convicted*. New York, NY: Doubleday.

Schmechel, R. S., O'Toole, T. P., Easterly, C., & Loftus, E. F. (2006). Beyond the ken? Testing jurors' understanding of eyewitness reliability evidence. *Jurimetrics, 46*(2), 177-214.

Sharman, S. J., & Powell, M. B. (2012). A comparison of adult witnesses' suggestibility across Various types of leading questions. *Applied Cognitive Psychology, 26*(1), 48-53.

Shaw, J. S., Garven, S., & Wood, J. M. (1997). Co-witness information can have immediate effects on eyewitness memory reports. *Law and Human Behavior, 21*(5), 503-523.

Shaw, J. S., & Zerr, T. K. (2003). Extra effort during memory retrieval may be associated with increases in eyewitness confidence. *Law and Human Behavior, 27*(3), 315-329.

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the U.S. population. *Plos One, 6*, doi: 10.1371/journal.pone.0022757

Simons, D. J., & Chabris, C. F. (2012). Common (mis)beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *Plos One, 7*, doi: 10.1371/journal.pone.0051876

Skagerberg, E. M., & Wright, D. B. (2008). The prevalence of co-witnesses and co-witness discussions in real eyewitnesses. *Psychology Crime & Law, 14*(6), 513-521.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 34-50.

Surowiecki, J. (2004). *The wisdom of crowds*. New York, NY: Doubleday.

Troyer, A. K., & Craik, F. I. (2000). The effect of divided attention on memory for items and their context. *Canadian Journal of Experimental Psychology (Revue Canadienne de Psychologie Expérimentale), 54*(3), 161-171.

Vredeveldt, A., Hildebrandt, A., & van Koppen, P. J. (2015). Acknowledge, repeat, rephrase, elaborate: Witnesses can help each other remember more. *Memory.* doi: 10.1080/09658211.2015.1042884

Vredeveldt, A., Hitch, G. J., & Baddeley, A. D. (2011). Eye closure helps memory by reducing cognitive load and enhancing visualisation. *Memory & Cognition, 39*(7), 1253-1263.

Vredeveldt, A., & Sauer, J. D. (2015). Effects of eye-closure on confidence-accuracy relations in eyewitness testimony. *Journal of Applied Research in Memory and Cognition, 4*(1), 51–58.

Wallbott, H. G. (1992). Effects of distortion of spatial and temporal resolution of video stimuli on emotion attributions. *Journal of Nonverbal Behavior, 16*(1), 5-20.

Waubert de Puiseau, B., Aßfalg, A., Erdfelder, E., & Bernstein, D. M. (2012). Extracting the truth from conflicting eyewitness reports: A formal modeling approach. *Journal of Experimental Psychology: Applied, 18*(4), 390-403.

Weller, S. C. (1987). Shared knowledge, intracultural variation, and knowledge aggregation. *American Behavioral Scientist, 31*(2), 178-193.

Weller, S. C. (2007). Cultural Consensus Theory: Applications and frequently asked questions. *Field Methods, 19*(4), 339-368.

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence. Improving its probative value. *Psychological Science in the Public Interest, 7*(2), 45-75.

Wheatcroft, J. M., Wagstaff, G. F., & Kebbell, M. R. (2004). The influence of courtroom questioning style on actual and perceived eyewitness confidence and accuracy. *Legal and Criminological Psychology, 9*(1), 83-101.

Tables

Table 1

2x2 Contingency Tables and the Results of the McNemar Tests Comparing the Majority with the Consensus Reconstructions for the Homogeneous versus Heterogeneous Samples Separately for Different Sample Sizes (a: 10, b: 20, and c: 40)

a) $n = 10$		Homogeneous samples		Total
		GCM > MR	GCM ≤ MR	
Heterogeneous samples	GCM > MR	17	32	49
	GCM ≤ MR	4	14	28
Total		21	46	67
Odds ratio			8	
Proportion of discordant pairs			54%	
McNemar test			$\chi^2(1, N = 67) = 46.00, p < .001$	

b) $n = 20$		Homogeneous samples		Total
		GCM > MR	GCM ≤ MR	
Heterogeneous samples	GCM > MR	24	39	63
	GCM ≤ MR	2	2	4
Total		26	41	67
Odds ratio			19.5	
Proportion of discordant pairs			61%	
McNemar test			$\chi^2(1, N = 67) = 31.61, p < .001$	

c) $n = 40$		Homogeneous samples		Total
		GCM > MR	GCM ≤ MR	
Heterogeneous samples	GCM > MR	9	48	57
	GCM ≤ MR	0	10	10
Total		9	58	67
Odds ratio			NA[a)]	
Proportion of discordant pairs			72%	
McNemar test			$\chi^2(1, N = 67) = 46.02, p < .001$	

Note. MR = Majority Rule; [a)]no odds ratio could be computed for $n = 40$ because division by 0 is not defined.

Figures

Fig. 1 Mean proportions of agreement (and their standard errors) between the true answer key on the one hand and the answer key estimates that were based on the Majority Rule (majority reconstruction, solid curve), the GCM (consensus reconstruction, dashed curve), and the individual responses (individual reconstruction, dotted curve) on the other hand as a function of competence heterogeneity (a: homogeneous condition; b: heterogeneous condition) and the number of witnesses (n = 10, 20, and 40)

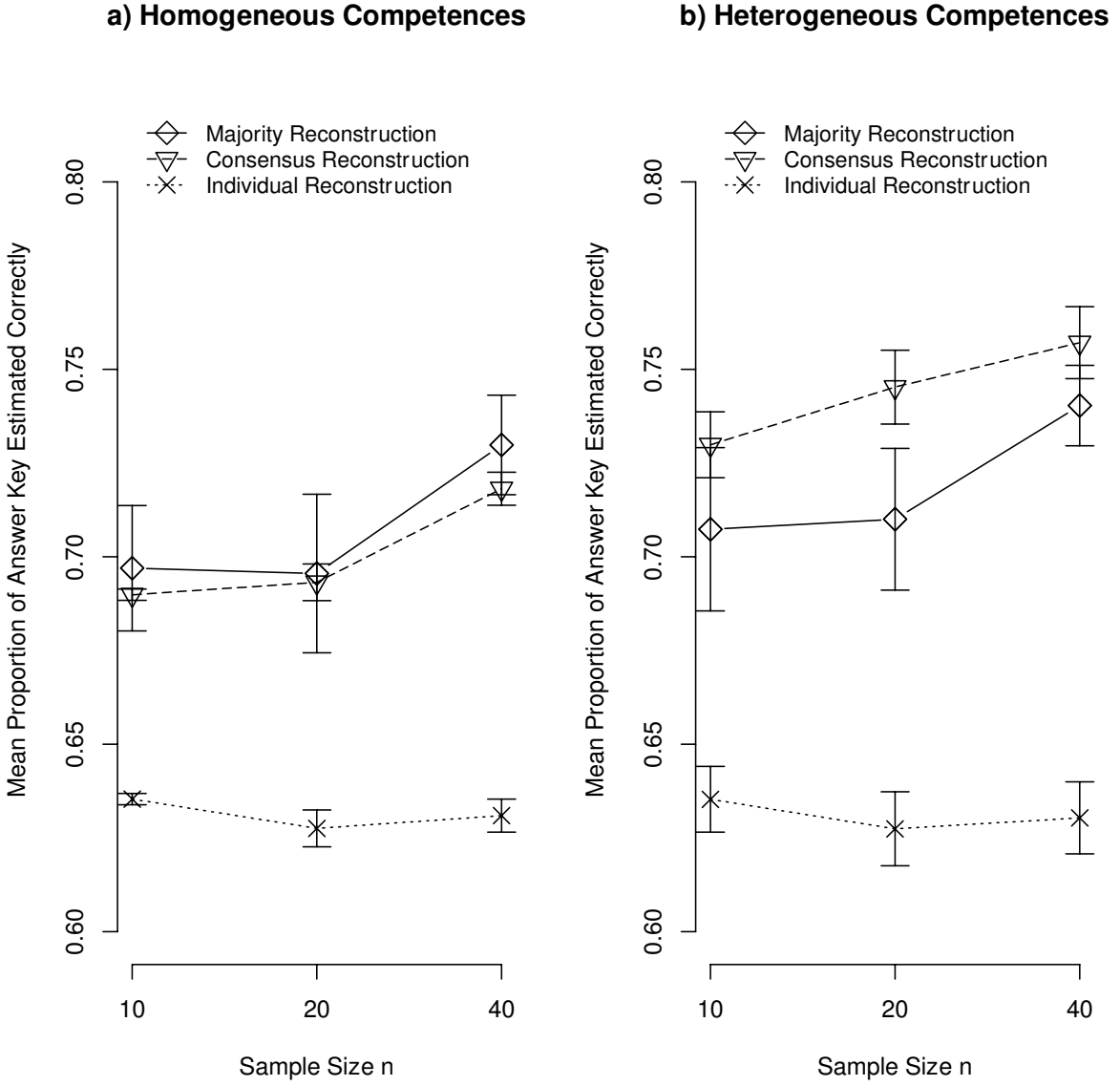
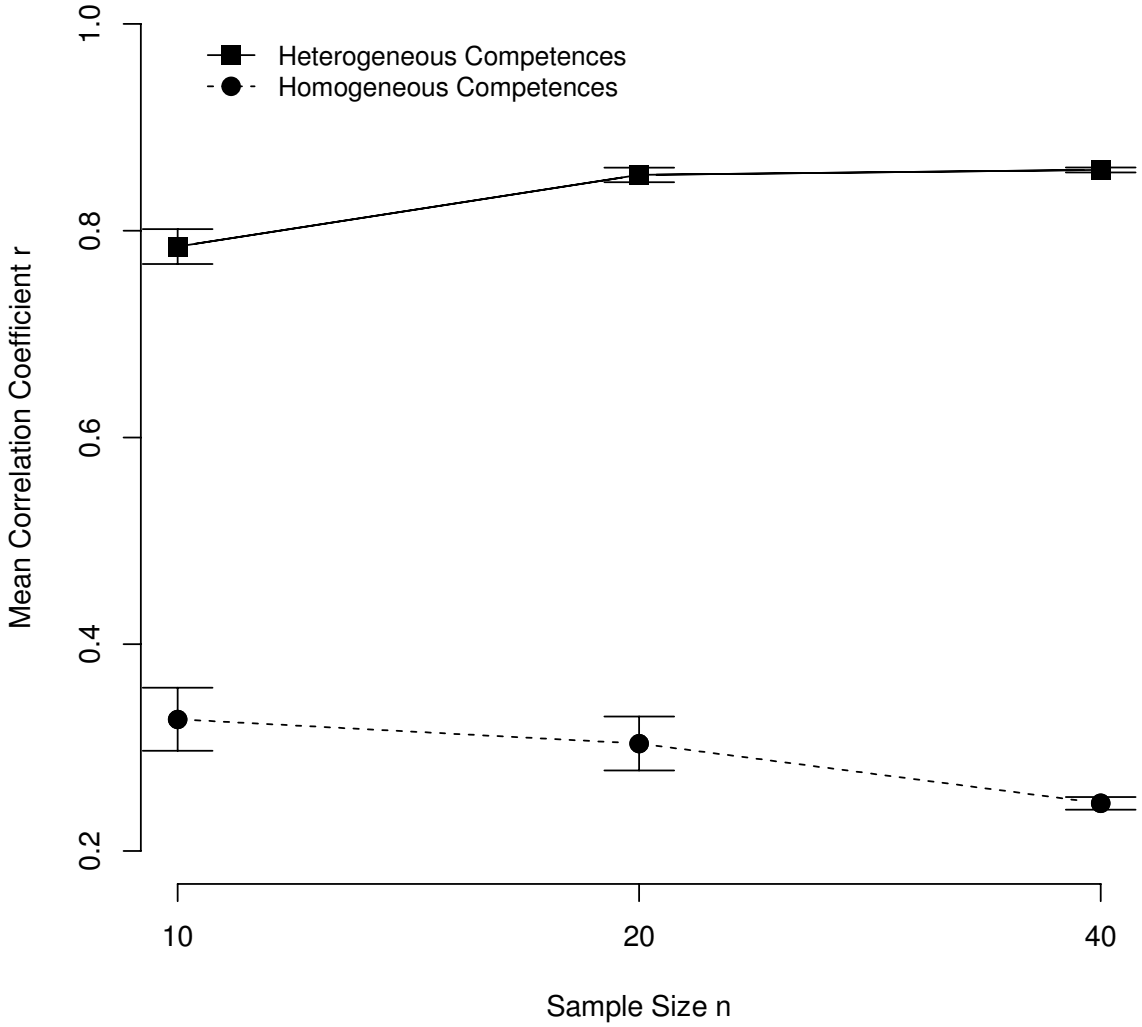


Fig. 2 Mean Pearson product moment correlations (and their standard errors) between the competence parameter estimates that were based on the GCM and the 2-HTM (i.e., the true answer key) in the homogeneous condition (solid line) and the heterogeneous condition (dashed line) as a function of the number of witnesses (n = 10, 20, and 40)



Appendix

Formalization of the General Condorcet Model

In a witness recognition experiment, N witnesses first observe a crime and then make recognition judgments about M statements regarding their observations. In the 2-HTM, responses are modeled as a function of a witness's competence, D_i, $i \in \{1, \ldots, N\}$, and the witness's tendency to guess that a statement is "true," g_i, when the witness does not know the answer. Each witness is assumed to judge a statement as "true" if the witness believes that a detail has occurred, or as "false" if the witness believes that a detail has not occurred. Thus, "true" responses can be classified either as hits if the statement is true or as false alarms if the statement is false. In the 2-HTM, hits occur either because the witness remembers the relevant fact with probability D_i or because the witness does not remember the relevant fact with probability $(1- D_i)$ but guesses correctly with probability g_i. In the 2-HTM, competence and guessing bias are assumed to be constant across questions. Thus, the probability of a hit, H_i, is $H_i = D_i + (1- D_i)\, g_i$. False alarms are assumed to occur when the witness does not remember the relevant fact with probability $(1- D_i)$ and then incorrectly guesses "true" with probability g_i. Thus, the probability of a false alarm, F_i, can be computed as $F_i = (1 - D_i)g_i$. Solving these equations for D_i and g_i yields:

$$D_i = H_i - F_i \tag{1}$$

and

$$g_i = \frac{F_i}{(1 - H_i + F_i)} \tag{2}$$

By using the observed hit and false-alarm rates as estimates of H_i and F_i, respectively, a witness's competence and guessing bias can be estimated with Equations 1 and 2.

Computing competence and guessing bias in the GCM is more complex because the answer key is unknown. The GCM, therefore, extends the 2-HTM by adding a latent variable, the answer key $\mathbf{Z} = (Z_k)_{1xM}$, which is a vector of correct responses for items $k \in \{1, \ldots, M\}$:

$$Z_k = \begin{cases} 1, \text{if the correct judgment of item } k \text{ is "true"} \\ 0, \text{if the correct judgment of item } k \text{ is "false"} \end{cases} \quad (3)$$

Further, the GCM (Karabatsos & Batchelder, 2003; Oravecz et al., 2014) includes another latent variable, the difficulty of item k, δ_k, with $0 < \delta_k < 1$. Taking item difficulty into account, Karabatsos and Batchelder (2003) define the probability of witness i knowing the correct response to item k as

$$D_{ik} = \frac{\theta_i(1-\delta_k)}{\theta_i(1-\delta_k)+(1-\theta_i)\delta_k}, \quad (4)$$

where θ_i denotes the competence of witness i, independent of item difficulty, with $0 < \theta_i < 1$[6].

On the basis of these equations, the GCM defines the probability that witness i correctly recognizes statement k as:

[6] Because different combinations of θ_i and δ_k yield the same D_{ik}, an additional constraint on Equation 4 is necessary (Crowther, Batchelder, & Hu, 1995). Following the procedure employed by Crowther et al. (1995) and Waubert de Puiseau et al. (2012), we therefore set $\delta_1 = .5$ in all analyses.

$$p_{ik} = D_{ik}^{Z_k} + g_i(1 - D_{ik})(2Z_k - 1). \quad (5)$$

The parameter estimates for the latent parameters competence θ_i, guessing bias g_i, item difficulty δ_k, and the answer key Z_k are determined simultaneously from the response matrix $\mathbf{X} = (X_{ik})_{NxM}$,

$$X_{ik} = \begin{cases} 1, \text{if witness } i \text{ answers "true" to item } k \\ 0, \text{if witness } i \text{ answers "false" to item } k \end{cases}. \quad (6)$$

We used the Markov-chain-Monte-Carlo procedure described by Karabatsos and Batchelder (2003) to find parameter estimates that maximize the likelihood function

$$L(\mathbf{X}|\Omega) = \prod_{i=1}^{N} \prod_{k=1}^{M} p_{ik}^{Z_k X_{ik} + (1-Z_k)(1-X_{ik})} \times (1 - p_{ik})^{Z_k(1-X_{ik}) + (1-Z_k)X_{ik}}. \quad (7)$$

where $\Omega = \{\theta_{<i=1,\ldots,N>}, g_{<i=1,\ldots,N>}, \delta_{<k=1>}, Z_{<k=1,\ldots,M>}\}$ are the parameters of the GCM. More detailed descriptions of the 2-HTM and the GCM can be found elsewhere (cf. Aßfalg & Erdfelder, 2012; Batchelder & Romney, 1986, 1988, 1989; Karabatsos & Batchelder, 2003; Oravecz, Vandekerckhove, & Batchelder, 2014; Romney, Batchelder, & Weller, 1987; Romney et al., 1986).

Versicherung an Eides Statt

Hiermit versichere ich an Eides Statt, dass die Dissertation mit dem Titel

On the Assessment of Witnesses' Memory for Events

von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist. Ferner versichere ich, dass die Arbeit in der vorgelegten oder in ähnlicher Form bisher bei keiner anderen Fakultät als Dissertation eingereicht wurde und dass ich bisher keine erfolglosen Promotionsversuche unternommen habe.

Düsseldorf, den

Berenike Waubert de Puiseau