Reliability and Perceived Credibility of Older Eyewitnesses

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Philosophie des Fachbereiches 06 Psychologie

der Justus-Liebig-Universität Gießen

vorgelegt von

Natalie Martschuk

aus Lübeck/ Sydney (Australien)

2019

Dekanin: Prof. Dr. Christiane Hermann

1. Berichterstatter: Prof. Siegfried L. Sporer, Ph.D.

2. Berichterstatter: PD Dr. Kai Hamburger

Tag der Disputation: 7. August 2019

DANKSAGUNG

An dieser Stelle möchte ich mich bei einer ganzen Reihe von Personen bedanken, die mich während der Dissertation unterstützt haben. Mein besonderer Dank gilt Prof. Siegfried L. Sporer, der mir ermöglicht hat, im Bereich der Rechtspsychologie tätig zu werden. Seine fachlich-kompetente Unterstützung und seine hilfreichen und kritischen Rückmeldungen haben deutlich zum Gelingen der vorliegenden Arbeit beigetragen. Bedanken möchte ich mich auch bei PD Dr. Kai Hamburger, der sich bereit erklärt hat, für diese Arbeit als Zweitgutachter zu fungieren. Prof. Jane Goodman-Delahunty danke ich für den fachlichen Austausch, für das konstruktive Feedback an Teilen der Dissertation, und für ihre Motivation, Verständnis und Optimismus. Dr. Kerstin Kocab danke ich für ihre Mitwirkung an den frühen Stadien der Metaanalyse, einschließlich Literaturrecherche, Entwicklung von Kodierungsschema und Kodierung der Primärstudien. Dr. Melanie Sauerland danke ich für das Bereitstellen der Daten ihrer Feldstudie und für die Zusammenarbeit an der Re-analyse dieser Daten. Bedanken möchte ich mich auch bei Dr. Stephen Doherty, Dr. Evianne Grosvenor, Thea Gumbert-Jourjon, Alexandra Lonergan, Prof. Christian Meissner, Stella Palmer und den anonymen Peer-Reviewern für ihre hilfreichen Rückmeldungen und Anregungen zu Teilen dieser Arbeit. Dr. Gang (John) Xie und Lema Seknewna danke ich für die Hilfestellung an der Entwicklung der *R*-Codes, und Dr. Guillaume Jourjon für die Hilfestellung an der Implementierung der Onlinestudie. Meiner langjährigen Freundin Prof. Jacqueline Loos danke ich für ihre emotionale Unterstützung, die vielen Ratschläge und Anmerkungen in Bezug auf diese Arbeit. Für die emotionale Unterstützung danke ich auch meiner Schwester Aljona Martschuk. Ein besonderer Dank gilt meinen Eltern Emma Martschuk und Andre Martschuk, die mich über die Jahre in jeglicher Weise unterstützt und motiviert haben. Ebenso bedanke ich mich herzlich bei meinem Verlobten Andre Hollack für seine Unterstützung in jeglicher Hinsicht, für seinen Optimismus und seine Lebensfreude und dafür, dass er sich jede freie Minute nahm, um unsere Tochter Evolet zu betreuen und mir mehr Zeit zum Schreiben zu verschaffen.

Sydney, im März 2019 *Natalie Martschuk*

ABSTRACT

This dissertation reports three studies that investigated the reliability and perceived credibility of older compared to younger eyewitnesses. The studies were embedded in an integrative model for investigating and evaluating eyewitness testimony which distinguishes an information processing, a metamemorial and a judgmental level (Sporer, 2008). In Study 1 a meta-analysis integrated studies on the influence of age on face recognition at the information-processing level. A total of 19 studies with 79 comparisons of younger and older participants were included. Analyses revealed small to moderate effects for hits, and large effects for false alarms and signal detection theory (SDT) measures. Younger participants outperformed older participants on most face recognition measures. Younger participants made more hits ($g_u = 0.31$) and fewer false alarms ($g_u = 0.95$), and thus had better SDT recognition performance ($g_u = 1.01$) than older participants. These effects were largest for young faces, smaller for mixed-age faces, and smallest for older faces. Furthermore, older participants used a more liberal response criterion, that is, they were more likely to choose a face than younger participants ($g_u = 0.54$). Meta-regression analyses revealed that young faces (vs. mixed-age faces) and longer retention intervals were associated with greater differences between the age groups for hits but not for false alarms. Funnel plot and trim-and-fill analyses indicated the presence of a publication bias.

Study 2 shed light on metamemorial processes across the lifespan in a re-analysis of a field study by Sauerland and Sporer (2009; $N = 720$; $n = 436$ choosers between 15 and 83 years old). Different calibration indices and Bayesian analyses demonstrated a progressive dissociation between identification performance and confidence across age groups. While the confidence expressed following an identification remained unchanged across the lifespan, identification accuracy decreased. Young, highly confident witnesses were much more likely to be accurate than less confident witnesses. With increasing age, witnesses were more likely to be overconfident, particularly at the medium and high levels of confidence, and the postdictive value of confidence and decision times decreased. Overall, the postdictive value of confidence was reduced for participants aged 40 years or older.

At the judgmental level, Study 3 investigated how jurors as fact finders perceive older compared to younger eyewitnesses by manipulating victim age and crime severity in a randomized between-participants design. A total of 204 jury-eligible participants read a summary of a robbery trial and filled out a short questionnaire. Mock jurors believed the young victim more than the older victim when the crime was severe, while no age differences were found for the less severe crime. They evaluated the severe case more in favor of the prosecution for the young victim, while the effect was reversed for the older victim. Whereas previous research demonstrated that juror characteristics were generally associated with culpability we demonstrated that with case-specific information these general views became less important.

In conclusion, the present dissertation demonstrated that memory and metamemory are more fallible in older age than in young age. Therefore, fact finders, that is, police investigators, prosecutors, jurors, and judges need to consider the age of the witness, together with decision time and expressed confidence when they evaluate identification judgments. The present dissertation also showed that mock jurors as fact finders were aware of age-related memory changes only to some extent, and did not consider it sufficiently when evaluating evidence of older eyewitnesses. Following, jurors would benefit from educative information about age-related changes in memory and cognitive capacity in cases involving older eyewitnesses.

TABLE OF CONTENTS

INTRODUCTION

Eyewitnesses provide information about the culprit, the crime, and the surrounding circumstances. Particularly in cases lacking forensic evidence the criminal justice system relies on eyewitness identifications to link a suspect to a crime. Problems with the accuracy of eyewitness identifications have been noted throughout the history of criminal law (Sporer, 1982). Since the beginning of the 1970s, perhaps as a function of some highly publicized legal cases and a shift from “basic” to more “applied” research on memory, several thousand articles on face recognition and person identification as well as several monographs and edited volumes have appeared (e.g., Clifford & Bull, 1978; Loftus, 1979; Shepherd, Ellis, & Davies, 1982; Sporer, 1992; Yarmey, 1979) but have often gone unnoticed by legal decision makers. Yet members of the legal system (i.e., police, prosecutors, and judges) became more sensitive to the problem of mistaken identification when analyses of DNA exoneration cases revealed that a high proportion of these cases involved mistaken identifications, along with other reasons (Garrett, 2011). The problem of wrongful convictions is global—more than 5,700 wrongful convictions have been counted internationally (The Medill Justice Project, 2018), and Innocence Projects have been established in countries such as Australia, France, New Zealand, Norway, and Japan (for a list of projects devoted to overturning wrongful convictions, see http://justicedenied.org/contacts.htm). In most of the cases documented, multiple factors contributed to wrongful convictions, with person identifications often playing a prominent role.

According to the integrative model of eyewitness testimony and its evaluation (Sporer, 2008) problems with person identifications can appear at three levels: (1) the *information-processing level*, (2) the *metamemory level*, and (3) the *judgmental level*—all of which were subject to investigation in the present dissertation. First, at the *information-processing level,* problems can occur during the perception phase, the retention phase, and/or the retrieval phase (Sporer, 2008). Two different research paradigms are distinguished in this context: face recognition and eyewitness identification studies. *Face recognition* studies typically present a large number of faces during the encoding phase. At the recognition phase, these target faces are mixed with an equal number of distractor faces, and the participants indicate whether they saw the face before or not. By contrast, *eyewitness identification* studies examine recognition of faces in the context of a particular event, that is, participants typically watch a staged crime, with one or two targets (Sporer, 1992). In the recognition phase, participants are given the task to identify the target in an array of five to ten people, and are informed (unbiased instruction) or not informed (biased instruction) that the perpetrator may or may not be in the lineup. Participants then indicate whether the perpetrator is in the lineup and express their confidence that their decision was correct.

Witnesses' internal factors, including attention, memory strength and motivation, as well as external factors, namely context and post-event information, can influence the reliability of eyewitness identifications. Furthermore witness factors, such as personality factors and young or old age may affect identification decisions. Adult ageing is associated with a decline in memory, cognitive capacity (Coleman & O'Hanlon, 2008) and systematic reductions in vision and processing speed (Salthouse, 1996; Schieber, 2006). These factors can, in turn, negatively influence perception and recognition of faces (Lott, Haegerstrom-Portnoy, Schneck, & Brabyn, 2005). Further, age-related structural changes of specific brain areas have been associated with a more liberal response bias, that is, the proneness to choose a person regardless of whether the person was the target or not (Huh, Kramer, Gazzaley, & Delis, 2006).

Narrative reviews documented age effects in face recognition and eyewitness identification (Bartlett, 2014; Bartlett & Memon, 2007). Meta-analyses showed a strong age effect in eyewitness identification (Erickson, Lampinen, & Moore, 2015; Kocab, Martschuk, & Sporer, 2019). However, a statistical comparison across studies testing age effect on recognition of faces has been lacking. Study 1 aimed to close this gap by conducting a systematic meta-analysis of 19 studies with 79 comparisons that investigated face recognition by young versus older adults. Effect sizes were calculated for correct recognition (hits), incorrect recognition (false alarms), and the signal detection theory (e.g., Macmillan & Creelman, 2005) measures of recognition performance (i.e., sensitivity) and response bias (i.e., proneness to choose or reject a face). Target age was considered to evaluate a potential own-age bias, that is, better recognition of faces of the in-group (own-age) than the out-group (other-age) (Sporer, 2001). Moreover, the meta-analysis conducted meta-regression analyses to assess the potential effect of predictor variables, while controlling for interdependencies of the predictors. In addition to face recognition, Study 2 investigated performance across the lifespan from adolescence to old age using an identification paradigm in a field study.

Second, at the *metamemory level* witnesses evaluate their own memory process by applying metacognitive strategies and expressing their certainty about their memory (Sporer, 2008). Here, errors can occur when eyewitnesses express high confidence in their judgment, despite making errors (overconfidence), or low confidence when their response is, in fact, correct (underconfidence). An indirect measure of accuracy at the metamemory level is response latency, that is, the time an eyewitness takes to reach an identification decision.

While early reviews and meta-analyses found that confidence is only modestly associated with identification accuracy (Bothwell, Deffenbacher, & Brigham, 1987; Wells & Murray, 1983), analyses which consider only choosers—those who made a positive identification which are forensically more relevant—revealed that confidence was strongly associated identification accuracy (Sporer, Penrod, Read, & Cutler, 1995). Recently, based on re-analyses of previous studies, researchers made even stronger arguments that confidence expressed at the time of an initial identification are reliable indicators of identification accuracy (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted, Read, & Lindsay, 2016; Wixted & Wells, 2017). These analyses were, however, conducted across large datasets without consideration of potential moderators, such as eyewitness age.

Metacognitive models indicate that metacognitive judgments and accuracy may dissociate in older age (Palmer, David, & Fleming, 2014), because older adults may encounter problems in effectively using and retrieving information when making decisions (Hertzog & Hultsch, 2000; Shing, Werkle-Bergner, Li, & Lindenberger, 2009). However, the majority of research was conducted with verbal materials, not with visual materials, such as faces. Study 2 examined the confidence–accuracy relationship across the lifespan in a re-analysis of a large-scale field study by Sauerland and Sporer (2009). Analyses were conducted with choosers only (436 out of 720 participants, aged between 15 and 83 years)—

those who would end up testifying in court. In addition to different calibration indices, Bayesian analyses were performed to assess base rates of the probability that the suspect is the perpetrator for each age group separately.

Further, in line with the *dual-process model* that accurate decisions are fast and automatic, while inaccurate decisions are time-consuming and conscious (Dunning & Perretta, 2002; Dunning & Stern, 1994), correct identifications are supposedly made faster than false identification (Sporer, 1992, 1993, 1994). Study 2 extended this theoretical notion by testing the relationship between decision time and identification accuracy as a function of age. These analyses aimed to provide further evidence to the question whether metamemory declines in older age.

Finally, at the *judgmental level* fact finders or "memory judges" (Sporer, 2008, p. 750) evaluate eyewitness testimony observed at the information-processing and metamemory levels, by using testimony/identification statements, behaviors, and statements of metamemorial processes by an eyewitness. Fact finders are usually police investigators, prosecutors, jurors or judges, and are typically not memory experts. In their decision, fact finders evaluate witness evidence in light of other available evidence. Errors may occur at the judgmental level when eyewitness testimony is interpreted incorrectly or when an alternative hypothesis is not considered (Sporer, 2008). Study 3 investigated the decision processes of jurors—those who decide on a defendant's guilt in a criminal trial in an adversarial system. Of particular interest was the mock jurors' awareness of age-related memory changes, that is, whether they knew that cognitive capacity declines in older age and used this knowledge to render a verdict in the case of an older versus a young eyewitness. A total of 205 jury-eligible citizens from countries holding jury trials were provided a summary of a robbery trial, in which eyewitness age and crime severity were manipulated. In addition to rendering a verdict of the defendant's guilt, mock jurors responded to a series of questions, which aimed at assessing their decision processes as well as beliefs and attitudes about memory in old age and the criminal justice system.

References

Bartlett, J. C. (2014). The older eyewitness. In T. Perfect & D. Lindsay (Eds.), *The SAGE handbook of applied memory* (pp. 654–674). London, UK: SAGE Publications Ltd.

Bartlett, J. C., & Memon, A. (2007). Eyewitness memory in young and older adults. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Vol. 2: Memory for people* (pp. 309–338). Mahwah, NJ: Lawrence Erlbaum.

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). An exploratory study of personality differences in eyewitness memory. *Journal of Applied Psychology*, *72*, 691–695. doi:0021-9010/87

Clifford, B. R., & Bull, R. (1978). *The psychology of person identification.* London: Routledge & Kegan Paul.

Coleman, P. G., & O'Hanlon, A. (2008). Ageing and adaptation. In R. Woods, & L. Clare (Eds.), *Clinical psychology of ageing* (pp. 17-32). West Sussex, England: John Wiley & Sons.

Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10- to 12-second rule for distinguishing accurate from inaccurate positive identification. *Journal of Applied Psychology, 87*, 951–962. doi:10.1037/0021-9010.87.5.951

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, *67*, 818–835. doi:10.1037/0022-3514.67.5.818

Erickson, W. B., Lampinen, J. M., & Moore, K. N. (2016). Eyewitness identifications by older and younger adults: A meta-analysis and discussion. *Journal of Police and Criminal Psychology, 31,* 108–121. doi:10.1007/s11896-015-9176-3

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge: Harvard University Press.

Hertzog, C., & Hultsch, D. F. (2000). Metacognition in adulthood and old age. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp.

417–466). Mahwah, NJ: Lawrence Erlbaum.

Huh, T. J., Kramer, J. H., Gazzaley, A., & Delis, D. C. (2006). Response bias and aging on a recognition memory task. *Journal of the International Neuropsychological Society, 12,* 1–7. doi:10.1017/S1355617706060024

Kocab, K., Martschuk, N., & Sporer, S. L. (2019). *Identification evidence by elderly witnesses: A meta-analysis*. Unpublished manuscript.

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

Lott, L. A., Haegerstrom-Portnoy, G., Schneck, M. E., & Brabyn, J. A. (2005). Face recognition in the elderly. *Optometry and Vision Science, 82,* 874–881. doi:10.1097/01.opx.0000180764.68737.91

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

The Medill Justice Project. (2018). *Wrongful convictions.* Medill Northwestern University. Retrieved from http://www.medilljusticeproject.org/wrongful-convictions/

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, *28*, 151–160. doi:10.1016/j.concog.2014.06.007

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103,* 403–428. doi:10.1037/0033-295X.103.3.403

Sauerland, M., & Sporer, S. L. (2007). Post-decision confidence, decision time, and self-reported decision processes as postdictors of identification accuracy. *Psychology, Crime, & Law*, *13*, 611–625. doi:10.1080/10683160701264561

Schieber, F. (2006). Vision and aging. In J. E. Birren, & K. W. Schaie (Eds.), *The handbook of the psychology of aging* (pp. 129–161). Burlington, USA: Elvesier Academic Press.

Shepherd, J. W., Ellis, H. D., & Davies, G. M. (1982). *Identification evidence: A psychological examination.* Aberdeen: Aberdeen University Press.

Shing, Y. L., Werkle-Bergner, M., Li, S.-C., & Lindenberger, U. (2009). Committing memory errors with high confidence: Older adults do but children don't. *Memory*, *17*, 168–179. doi:10.1080/09658210802190596

Sporer, S. L. (1982). A brief history of the psychology of testimony. *Current Psychological Reviews, 2,* 323–339. doi:10.1007/BF02684465

Sporer, S. L. (1992). *Das Wiedererkennen von Gesichtern [Recognizing faces].* Weinheim: Beltz/Psychologie Verlags Union.

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology*, *78*, 22–33. doi:10.1037/0021-9010.78.1.22

Sporer, S. L. (1994). Decision–times and eyewitness identification accuracy in simultaneous and sequential lineups. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 300–327). New York: Cambridge University Press.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7,* 36–97. doi:10.1037//1076-8971.7.1.36

Sporer, S. L. (2008). Lessons from the origins of eyewitness testimony research in Europe. *Applied Cognitive Psychology, 22,* 737–757. doi:10.1002/acp.1479

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

Wells, G. L., & 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

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*, 515–526. doi:10.1037/a0039510

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence–accuracy relationship. *Journal of Applied Research in Memory and Cognition*, *4*, 329–334. doi:10.1016/j.jarmac.2016.04.006

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

Yarmey, A. D. (1979). *The psychology of eyewitness testimony.* London: MacMillan.

MEMORY FOR FACES IN OLD AGE: A META-ANALYSIS

Older people often complain about memory problems in everyday life. Forgetting other people's names or faces can be particularly embarrassing. Most research on memory of older people has been conducted on verbal memory. Here, we focus on visual memory, in particular, recognition memory of unfamiliar faces. In criminal cases, misidentifications have been one of the main contributing factors in wrongful convictions (*www.innocenceproject.org*).

Current State of Knowledge

To date, narrative literature reviews found evidence for age effects in face recognition and eyewitness identification studies (see Bartlett, 2014a; Bartlett & Memon, 2007). However, no formal meta-analysis has been conducted to statistically test the size of these effects on recognition of faces by young vs. older adults.

Face Recognition Studies

Typically, facial recognition studies have two phases, a presentation phase and a recognition phase. During the presentation phase, a large number of faces is presented, participants are told to remember them (intentional learning) or to judge them with respect to some characteristics (e.g., attractiveness, perceived age; i.e., an incidental learning paradigm). At recognition, previously seen stimuli are mixed with an (equal) number of distractors and participants judge them as "old" = previously seen or "new" = distractors.

Face researchers have repeatedly emphasized the use of real faces varying in pose and expression (e.g., Bruce, 1982; Burton, 2012; Sporer, 1992) that are not identical at study and test, and perhaps not even taken with the same camera (Burton, 2012). In our meta-analysis, we have coded studies to the extent they meet these desirable features.

Here, the focus is on *unfamiliar faces*, that is, faces of people not known to participants prior to the experiment. Using large numbers of faces allows researchers to test general principles of this type of stimuli as well as variations between stimuli based on perceived face qualities like attractiveness, distinctiveness, and memorability (e.g., Bainbridge, Isola, & Oliva, 2013) but also gender, race, and most pertinent for our review, age.

The goal of the present meta-analysis was to test age effects in memory for faces in face recognition studies, without the context of a complex event, by comparing recognition of younger (< 30 years) versus older participants (> 60 years) with faces of different age groups. Although we acknowledge from the outset that age-related declines may be continuous across the lifespan, we chose these age cut-off points in line with the great majority of studies. Whenever possible, however, we consider also more fine-grained distinctions between different age groups among the older adult group (cf. Bartlett, 2014b). Further, although other meta-analyses compared the performance of children and adults, the present meta-analysis focused on young and older adults only, mainly because differences in cognitive and memory processes between children and adults on the one hand and adults and older people on the other hand are not necessarily comparable.

Age-related Changes of Memory: Neurological and Cognitive Aspects

Ageing is often discussed as a deteriorative process. Advanced adult ageing is associated with systematic reductions in the efficiency of the sensory system (Schieber, 2006). Myriad age-related changes occur in the eye, the retina, and the ascending visual pathways in the nervous system. Further, reductions in the efficiency of the sensory system negatively affect visual acuity, contrast sensitivity, spatial vision, and speed perception, resulting in reduced encoding of information. These deficiencies, in turn, are thought to affect face recognition (Lott, Haegerstrom-Portnoy, Schneck, & Brabyn, 2005). Age-related changes of memory have also been attributed to age-related changes of brain structure (Coleman & O'Hanlon, 2008).

The main theoretical frameworks of memory in old age describe the memory processes and age-related changes from different perspectives. According to the *limited processing resources approach* (e.g., Craik, 1986), the available resources for encoding and retrieval of information are limited in older people and used in a different way. Particularly, older people's processing of resource-demanding information seems to be limited, especially when self-initiated processing is required (Maylor, 2005). An implication of this theoretical approach applied to face recognition would be that older people are less able to engage in "deeper" or more "elaborate" processing of faces which has been demonstrated to aid younger people in the recognition of unfamiliar faces (Bower & Karlin, 1974; Sporer, 1991). Relevant data were provided by Smith and Winograd (1978) who found that both young and older participants showed better recognition performance when judging the friendliness of a face than when judging the size of the nose or after a standard intentional learning instruction. Although the college student group had a higher *d'* score (2.46) than the two groups of older participants (high school educated: 1.58; college educated: 1.51) the interaction between encoding instructions and age was not significant. Thus, we need to be careful to apply theories developed with verbal material and tested with recall paradigms to face recognition. Interestingly, in Smith and Winograd's study older people showed a more lenient response criterion *beta* (resulting in more false alarms) than younger people. We will return to this point below.

Another widely cited explanation, the *reduced processing speed hypothesis* (Salthouse, 1996), proposed that cognitive performance deteriorates with the decrease of mental speed in old age because (a) relevant tasks cannot be accomplished in limited time, and (b) relevant information presented simultaneously may no longer be available when processing is completed. Consequently, older people show poor performance in memory even when given unlimited time (Maylor, 2005). Accordingly, when retrieval of information fails because of reduced processing speed, people would be more inclined to produce inaccurate memories. Support for this hypothesis was found in a cross-sectional study that examined event-related correlates of face recognition memory of 19 to 80 year old participants (Wolff, Wiese, & Schweinberger, 2012). This study demonstrated age-related slowing of face-processing speed beginning in early adulthood (aged 30 to 44) compared to a younger group (aged 19 to 29), with a steeper decline after the age of 60. Consequently, we would expect older adults to show weaker face recognition performance, regardless of the length of time allowed for encoding.

In more applied settings, the associated deficit hypothesis (Naveh-Benjamin, 2000) has been invoked to explain deficits of older compared to young people. According to this hypothesis, older people have difficulties linking a seen face with a name, an action or a specific context (e.g., Naveh-Benjamin, Guez, Kilb, & Reedy, 2004; Old & Naveh-Benjamin, 2008). Here we review only studies using recognition of faces irrespective of context cues, so this hypothesis appears less relevant.

Despite differences in emphasis, these theoretical frameworks converge in the sense that they both assume a reduction in information processing with increasing age which is associated with weaker cognitive performance in older age. Moreover, Craik and Rose (2012) argued that the reduced processing speed and efficiency are the consequences of limited attentional resources. A meta-analysis, which in principle is correlational, cannot test which of these theories better fits the available data.

Social-Cognitive Aspects and Own-age Bias

When examining recognition memory for faces, it is important to consider both cognitive (above) and social-psychological approaches. As proposed in Sporer's (2001) in-group/out-group model, people are better in recognizing faces belonging to their in-group than to an out-group. Group membership may be determined by one's ethnicity, gender, or age, and respective in-group performance advantages have been demonstrated (own-race, own-sex, own-age bias: Herlitz & Lovén, 2013; Meissner & Brigham, 2001; Rhodes & Anastasi, 2012; Shapiro & Penrod, 1986; Sporer, 2001; Wright & Sladden, 2003; Wright & Stroud, 2002). According to Sporer's model, when observers encounter an in-group person, they process his or her face in a rather automatic fashion. This entails configural, holistic processing which has been demonstrated to lead to better recognition performance. In contrast, with out-group faces, observers quickly label the face as an out-group member, which is considered of little personal interest for future interactions (cf. Malpass, 1990, social utility explanation). Thus, no further processing is engaged in, which lowers later recognition performance.

To our knowledge, this model is the only theoretical approach that predicts not only a deterioration in performance but also *a more lenient response criterion* for out-group faces. Accordingly, out-group faces are perceived as more similar to each other, that is, more "homogeneous" (out-group homogeneity effect: Quattrone & Jones, 1980). Encoding does not entail a search for individuating, distinctive features but rather relies on stereotypical features (e.g., "a skinhead; long hair; wrinkled face"). As these features are rather unspecific, they are not very helpful at the recognition stage. At recognition, the labeling process enlarges the region of acceptance used to designate a face as previously seen. Thus, more faces will be considered as previously seen (“Yes” responses), irrespective of whether they appeared in the study list, that is, an increase in response bias (for a detailed description of the model and supporting evidence, see Sporer, 2001).

From a related theoretical perspective, older people may rely upon a generalized sense of familiarity (Hancock, Bruce, & Burton, 2000; Vokey & Read, 1992), which is considered sufficient to respond with “Yes, previously seen”. This is tantamount to relying on a larger radius of acceptance in multiple dimension face space models (e.g., Valentine, 1991; see Sporer, 2001, for review). Consequently, a more lenient response bias would be expected with older people (main effect of participant age = age effect). To the extent that out-group faces are perceived as more homogeneous, this response bias of older people may be even more pronounced when recognizing young faces (Participant Age x Face Age interaction).

Empirical evidence for these models.

According to these categorization-individuation models, people tend to categorize and process faces of in- versus out-groups differently (Hugenberg, Miller, & Claypool, 2007; Sporer, 2001; Wiese, Komes, & Schweinberger, 2012). Accordingly, people may be less motivated to process out-group faces individually, perhaps even disregard them cognitively (Rodin, 1987; Sporer, 2001) and process them at a categorical level (Wiese et al., 2012).

In a series of five studies Rodin (1987) gained support for her *cognitive-disregard theory*, using both recall and recognition measures. Although youthfulness per se enhanced memorability, age-discrepancy acted as a disregard cue, which led to inferior face recognition of targets discrepant in age both in a laboratory experiment (Experiment 2)[1] and in a field study after a brief encounter (Experiment 3). Attractiveness as well as gender of targets moderated the effect (Experiment 4). Experiment 5 indicated that the effect was not restricted to recognition but affected also recall of personal attributes.

---

[1] Unfortunately, not enough statistical data were presented to include these studies in our meta-analysis.

More direct evidence for visual attentional processes at encoding comes from a study that examined recognition performance and visual scan patterns of younger and older people while they were observing faces of different age groups (He, Ebner, & Johnson, 2011). Independent of their age, people tended to look longer at faces of their own age group than at other-age faces. Participants also reported to have more contact with people of their own-age than the other-age group. Consequently, people were better in recognizing own-age than other-age faces. Interestingly, this study demonstrated that self-reported amount of contact with young and older people predicted face recognition (He et al., 2011).

The latter findings are compatible with the *contact hypothesis*, namely that recognition performance is related to the amount and quality of social contacts with the particular group (Meissner & Brigham, 2001; Sporer, 2001). Consequently, younger and older people would be expected to recognize faces of their own group better as a function of contact and experience with their own-age than with other-age groups.

Additional support for this assumption comes from a quasi-experimental study with geriatric nurses who had frequent contact with older people (Wiese, Wolff, Steffens, & Schweinberger, 2013). In contrast to a control group, these nurses were equally likely to recognize old-age and own-age faces. A comparable pattern was found with trainee teachers for child faces (Harrison & Hole, 2009; cf. also the field study on the own-race bias of White teachers tested before and after a teaching visit to a Black African community, cited in Valentine, Chiroro, & Dixon, 1995).

On the other hand, one could argue that older people have had a life-long experience with younger people *and* in later years with older people while younger people may predominantly have interacted with younger people. Consequently, younger people may show deficits with older relative to younger faces while older people may not show such differences as a function of face age (e.g., Short, Semplonius, Proietti, & Mondloch, 2014). In other words, there should be a disordinal, not a cross-over interaction (see Sporer, 2001). Contrary to this assumption, in Rhodes and Anastasi’s (2012) meta-analysis the own-age bias was present for hits and discriminability independently of participant age across the lifespan, beginning in early childhood until old age. An own-age bias in false alarms was observed with young adults only. This suggests that not the life-long, but the more recent experience may be crucial. Another possibility is that factors other than the own-age bias might account for people's proneness to a particular response bias, for instance stable "cognitive traits" (Kantner & Lindsay, 2012).

Face Recognition of Unfamiliar vs. Familiar Faces

Neuropsychological studies have demonstrated that certain factors may affect processing and recognition of familiar and unfamiliar faces differently (see Johnston & Edwards, 2009, for a review). For instance, viewpoint, facial expression and context influence recognition of unfamiliar but not of familiar faces. Unfamiliar faces, unlike familiar faces, seem to be encoded in a viewpoint dependent manner, and thus their recognition is sensitive to image changes. To study face recognition, not just picture recognition, stimuli should not be identical at study and test (Bartlett, 2014b; Bruce, 1982; Sporer, 1992). Studies which exclusively relied on picture recognition or used unnaturally cropped or computer-generated faces were excluded (e.g., Germine, Duchaine, & Nakayama, 2011; Hildebrandt, Wilhelm, Schmiedek, Herzmann, & Sommer, 2011).

To test the importance of face changes, we coded studies with respect to changes (e.g., expression and/or view/pose) between encoding and test and conducted moderator analyses with face change as a predictor (see Table 1 for an overview of study characteristics coded).

Table 1

Main Characteristics of Primary Studies in the Meta-analysis

	Study	Mean age of older age group	Face Age	Encoding task	Encoding time (s)	Retention interval (min)	No. of faces at encoding	No. of faces (distractors) at recognition	Gender of faces	Change of faces at test	Recognition task	Visual acuity	Health screening
1.1	Anastasi & Rhodes (2006, Exp. 1)	65.7	Young vs. Middle-aged vs. Old	Rating of attractiveness	7	18	24	48 (24)	Mixed	Identical	Y/N	N/A	SR
1.2	Anastasi & Rhodes (2006, Exp. 2)	69.5	Young vs. Middle-aged vs. Old	Categorizing age groups	7	2 days	24	48 (24)	Mixed	Expression	Y/N	N/A	SR
2.2	Bartlett & Fulton (1991, Exp. 2)	69.0	Mixed-age	Rating of pleasantness	10	0	48	72 (24)	Mixed	Expression	Y/N; Familiarity	Snellen chart 20/40	WAIS
3.1	Bartlett et al. (1989, Exp 1)	66.7	Mixed-age	Rating of pleasantness	10	0	48	72 (24)	Female	Expression, pose	IRJ	Snellen chart 20/40	WAIS
3.2	Bartlett et al. (1989, Exp 2)	69.6	Mixed-age	Rating of pleasantness	10	0	48	72 (24)	Female	Expression, pose	IRJ	Snellen chart 20/40	WAIS
4.1	Bartlett et al. (1991, Exp. 1)	72.0	Mixed-age	Rating of familiarity	12	25	24	48 (24)	Mixed	Expression, pose	IRJ	Snellen chart 20/50	GMT, WAIS
5.1	Bastin & van der Linden (2003)	64.4	Mixed-age	Intentional	Unlimited	30	18	36 (18)	Mixed	Identical	Y/N (or 2-AFC)	N/A	HT, SR, ST
6.1	D’Argembeau & van der Linden (2004)	67.0	Mixed-age	Intentional	5	5	12	24 (12)	Mixed	Expression	Y/N, RKG	N/A	BDI, SR, VT
7.1	Firestone et al. (2007)	70.5	Young vs. Old	Rating of photo quality, age of face	7	5	24	48 (24)	Mixed	Identical	Y/N	SR	PS

Table 1 continued

	Study	Mean age of older age group	Face Age	Encoding task	Encoding time (s)	Retention interval (min)	No. of faces at encoding	No. of faces (distractors) at recognition	Gender of faces	Change of faces at test	Recognition task	Visual acuity	Health screening
9.1	Fulton & Bartlett (1991)	71.4	Young vs. Middle-aged vs. Old	Rating of pleasantness	10	0	24	72 (48)	Mixed	Expression, pose	IRJ	Snellen chart 20/40	SR, WAIS
10.1	Lamont et al. (2005)	74.0	Young vs. Old	Intentional	5	0	30	60 (30)	Male	Identical	Y/N	RTT	N/A
11.1	Mason (1986)	75.0	Young vs. Old	Face-name pairs	10	0	40	80 (40)	Mixed	Identical	Y/N, name recognition	N/A	N/A
12.1	Naveh-Benjamin et al. (2009, Exp. 1)	72.7	Mixed-age	Face-name pairs; intentional	3	0	24	16 (8)	Mixed	Identical	Y/N, name recognition	N/A	SR
12.2	Naveh-Benjamin et al. (2009, Exp. 2)	73.2	Mixed-age	Fit of face-name pairs; incidental	3	0	48	32 (16)	Mixed	Identical	Y/N, name recognition	N/A	SR
13.1	Olofsson & Bäckman (1996)	73.8	Mixed-age	Face-name pairs	10	15	16	32 (16)	Mixed	Identical	Y/N, name recognition	N/A	SR, VT
14.1	Savaskan et al. (2007)	67.3	Mixed-age	Intentional	10	10	30	60 (30)	Mixed	Expression	Y/N, expression recognition	N/A	GDS, MMSE, SR
15.1	Smith & Winograd (1978)	65.0	Mixed-age	Rating of nose/ friendliness/ Intentional	8	0	50	60 (30)	Mixed	Identical	Y/N	N/A	N/A
16.1	Bryce & Dodson (2013)	72.0	Young vs. Old	Intentional	5	5	12	24 (12)	Mixed	Expression	Y/N, RKG	N/A	BDI, SR, VT
17.1	He et al. (2011)	73.9	Young vs. Old	Incidental (inspect pictures)	4	10	48	96 (48)	Mixed	Identical	Y/N	MARS; RPVS	SR, DSST

Note. Face Age: Young; Mixed = Young, (Middle-aged) and Old; Young vs. Middle-aged vs. Old = age of faces as a repeated measurement factor; Young vs. Old = both as a repeated measurement factor; Lamont et al. (2006) varied the number of faces as a between-participants factor (encoding: 20 vs. 40; test: 40 vs. 80); 2-AFC: two-alternative forced choice test; BDI: Beck Depression Inventory (Beck, Rush, Shaw, & Emery, 1979); DSST: Digit-Symbol-Substitution Test: measures visual-motor processing speed (Wechsler, 1981); GDS: Geriatric Depression Scale; GMT: Guild Memory Test (Gilbert, Levee, & Catalano, 1974); HT: Hayling Test (Burgess & Shallice, 1996); IRJ: Identity/recognition judgment with identical/old, old-but changed [expression, pose], new; MARS: MARS Letter Contrast Sensitivity Test (Arditi, 2005); MMSE: Mini-Mental State Examination; N/A: not available; PS: Prescreening through the Rothman Research Institute volunteer pool; RKG: Recognition with remember/know/guess judgment; RPVS: Rosenbaum Pocket Vision Screener (Rosenbaum, Granham-Field Surgical Co Inc, New York, NY); RTT: Reading Test types (Faculty of Ophthalmologists, London, 1987); SR: Self-report; ST: Verbal fluency: modification of the Stroop test (Stroop, 1935); VT: Vocabulary test; WAIS: Vocabulary test of the second half of the Wechsler Adult Intelligence Scale (Wechsler, 1981); Y/N: Yes-No recognition test.

Signal Detection Theory

To measure memory performance and response bias, many studies rely on signal detection theory (SDT; e.g., Macmillan & Creelman, 2005). SDT is based on the assumption that the degree of perceived familiarity (i.e., degree of memory strength) of a new face and a previously seen face are discriminable from each other (O'Toole, Bartlett, & Abdi, 2000). Therefore, the experienced level of familiarity should be higher for a previously seen face compared to a new face. Observers can either correctly recognize a previously seen face as "Old" (*hit*) or incorrectly reject it as "New" (*false rejection*). Similarly, they can either correctly reject (*correct rejection*) or incorrectly recognize a "New" face as previously seen (*false alarm*). Hits and false alarms suffice to describe recognition performance, given that false and correct rejections are merely complementary measures to hits and false alarms (Macmillan & Creelman, 2005).

Hits and false alarms can be integrated into combined measures of recognition performance or "sensitivity", usually *d'*. A *d'* of zero denotes no discrimination, and continuously increasing values of, for example, 1.0, 2.0 or 3.0 (to infinity) indicate stronger recognition performance. Besides sensitivity, people also differ in response tendencies. Most authors who did analyze response biases reported the criterion measure *c*, which is positive when hits exceed false alarms, negative when false alarms exceed hits, and zero if there is no difference (see Macmillan & Creelman, 2005). From an applied perspective, both a person's recognition performance as well as his or her response bias are of importance. For instance, a study that examined recognition memory of middle-aged adults (35–49 years) and older adults (75–89 years) using verbal learning test materials demonstrated that older adults showed a more liberal response criterion than their younger counterparts (Huh, Kramer, Gazzaley, & Delis, 2006). Further, inhibition—an executive function associated with the frontal brain lobe—was a significant predictor of response bias, suggesting a more liberal response criterion with reduced control in inhibition among older adults.

Mirror effect.

One way to test whether age affects only performance or also response bias is to test for a "mirror effect" (Glanzer & Adams, 1985) of hits and false alarms. A mirror effect describes the pattern of results in a Yes-No recognition experiment where a stimulus (here: a face) is accurately recognized as previously seen when it had been presented before (a hit), *and* accurately rejected as new when it had not been seen at presentation. Although some studies did not find such mirror effects in face recognition (e.g., Hancock et al., 2000; Vokey & Read, 1992) or face matching studies (Megreya & Burton, 2007), in a large scale meta-analysis of the own-race bias Meissner and Brigham (2001) found such a symmetrical pattern—although the effect sizes for false alarms were almost twice as large as those for hits.

Some authors have also postulated (and found) changes in response criterion along with a mirror effect (Hirshman, 1995; McClelland & Chappell, 1998). In line with Meissner and Brigham's assumption of a mirror effect and a concomitant change in response criterion, we also expected to find both effects in our meta-analysis.

One way to test for mirror effects is to examine the patterns of hits and false alarms separately for young and older participants. Unfortunately, researchers did not report the respective intercorrelations between dependent variables, so no meta-analytic synthesis of correlations was possible. Hence, we had to rely on visual inspection of unweighted means for hits and false alarms.

Summary of Hypotheses of the Present Meta-analysis

The present meta-analysis tested an overall age effect in face recognition across studies that compared young adults with an older age group, focusing on natural, unfamiliar faces. In line with a general age-related decline in neurological and memory processes considered to be involved in face recognition (Bartlett, 2014a; Coleman & O'Hanlon, 2008; Maylor, 2005) we predicted that across all studies, older people show decreased face recognition performance compared to younger people (i.e., an age effect). This should be reflected in both decreased hits and increased false alarms as well as in smaller *d'* values in older age. This age effect may be smaller in the recognition of older compared to young faces (own-age bias). As predicted from Sporer's (2001) in-group/out-group model, we expected a more liberal response criterion of older compared to younger people. Finally, we hypothesized that variables that make the task more difficult for participants (i.e., number of faces to be recognized, short encoding times, face changes between study and test, longer retention interval, additional tasks at test) would be particularly detrimental for old age participants and should hence be positively associated with effect sizes of the old-age effect.

Method

Search Strategies

Different search strategies were used to obtain studies for the current meta-analysis. Computerized search was conducted of *PsycINFO*, *EDOC*, *MEDLINE*, and *Web of Science* (*Science Citation Index Expanded*, *Social Sciences Citation Index* and *Arts and Humanities Citation Index)*, using free text search and search engines, that is, *ZPID*, *DGPs*, *Google Scholar*, and *Dissertation Abstracts*. To identify all possible studies, different combinations of the keywords *face recognition*, *identification*, *accuracy*, *age*, *elderly*, *old* using Boolean operators "AND" or "OR", and the wildcard character "*" were used.

Further, manual searches of relevant journals and reference sections of published articles were conducted. Finally, after completing the literature research, authors of the existing studies were contacted to obtain unpublished studies. The literature search revealed a large number of studies, including both lineup and face recognition studies. The current meta-analysis is restricted to studies using a recognition paradigm showing a large number of target faces at encoding that were randomly mixed with distractors at testing, displaying one face at a time. Studies reporting only neuropsychological measures were excluded. Person identification studies using single targets that were exposed either live or in a film later to be recognized in a lineup were recently reviewed in a meta-analysis by Kocab, Martschuk and Sporer (2017).

To be included, studies had to compare at least two groups of participants, viz. young people above 16 years and older persons above 60 years. In the primary studies participants with any intellectual or psychiatric impairments had been excluded. Stimuli must have been photographs of strangers of young or older target faces, not of familiar people like celebrities. Experiments with transformed, inverted or computer-generated faces were excluded. Dependent variables were hits, false alarms, and/or signal detection theory measures of performance (*d'*, *A'*) or bias (*c*, *beta*). The reported statistics had to provide data to calculate effect sizes (e.g., proportions, frequencies, means and standard deviations, or other statistical information such as *F*, *t*, or *p* values). If sufficient data were not reported, authors were contacted for providing missing information. Although we searched for published and unpublished studies in English, German, French and Russian, only English studies were found. A total of 19 studies met the inclusion criteria, with $n_Y = 656$ young ($M_Y = 24.1$ years) and $n_O = 755$ older people ($M_O = 70.1$ years).

The following studies were excluded from the analyses: Bartlett and Fulton (1991, Exp. 1) and Flicker, Ferris, Crook, and Bartus (1990) used a continuous recognition paradigm: Participants were instructed to identify repeated faces in an ongoing series of faces without a clear distinction between the encoding and recognition phase; Bastin (2008), Bastin and van der Linden (2006), and Old and Naveh-Benjamin (2008) used a two-alternative-forced-choice recognition paradigm, that is, at test participants were presented with pairs of faces consisting of a target and a distractor, and were instructed to choose the previously studied face. In Bartlett, Strater and Fulton (1991, Exp. 2), the sample was the same as in their first experiment. In order not to violate the assumption of independence of effect sizes, Exp. 2 was excluded. Rodin (1987, Exp. 2) did not report enough information to calculate an effect size. Ebner and Johnson (2009) reported only results for corrected recognition but not for the dependent measures of interest for this meta-analysis.

Study Characteristics and Coded Variables

Table 1 shows the main characteristics of the included studies. To ensure reliability of coding, each study was coded independently by two coders. Inter-coder reliabilities for categorical variables were between *kappa* = .66 and *kappa* = 1.00. Intra-class correlation coefficients (*ICCs*) for continuous variables were between *ICC* = .79 and *ICC* = 1.00. Discrepancies were discussed until full agreement was reached. The most important study characteristics coded are discussed below.

Encoding task.

The encoding task was coded continuously (e.g., number of faces, exposure time in seconds) and categorically (e.g., instruction: intentional vs. incidental). In intentional encoding participants were informed of the impending recognition test beforehand. Incidental encoding tasks involved categorizing faces by age, rating their pleasantness, etc. Encoding times varied between 1.5 (Bastin & van der Linden, 2003) and 12 seconds (Bartlett et al., 1991), with *Mdn* = 7 seconds. The number of faces at encoding was between 12 (D'Argembeau & van der Linden, 2004) and 50 (Smith & Winograd, 1978), *Mdn* = 24.

Retention interval.

The retention interval, that is, the time between encoding and recognition of the faces, varied between none (i.e., immediate; e.g., Bartlett & Fulton, 1991) and 48 hours (Anastasi & Rhodes, 2006, Exp. 2), *Mdn* = 0.

Recognition task.

In addition to the Yes-No recognition paradigm that requires participants to indicate whether each face had been seen before ("Yes") or not ("No") (e.g., Lamont, Steward-Williams, & Podd, 2005) some studies also tested the associations between names and faces (Naveh-Benjamin et al., 2009). Other studies tested the influence of change of expression or change of pose on memory for faces, instructing participants to classify each face as "identical", "changed", or "new".

Change of faces at test was coded with levels *no change*, *change of appearance*, *pose*, *expression*, or both *expression and pose*. Continuous variables were number of targets and number of distractors. The number of faces (targets and distractors) varied between 16 (Naveh-Benjamin et al., 2009) and 96 (He et al., 2011), *Mdn* = 48. Recognition of names, ratings of familiarity or of friendliness were not subject of the present meta-analysis and were therefore not analyzed.

Participant screening.

In the majority of studies (89.5%) participants were screened for visual acuity and/or with performance tests, dementia tests or tests for psychopathological problems (see Table 1).

Data Analyses

Separate effect sizes (Hedges g_u adjusted for sample bias; Borenstein, 2009) were calculated for hits, false alarms, and signal detection measures of performance (*d'*) and response bias (*c*, *beta*). For each dependent variable, one single effect size per study was calculated, if necessary averaging across experimental conditions. Some studies reported several dependent variables for the same participants, but not the correlations between them. To avoid dependencies among effect sizes, only one of the measures was used (e.g., the effect size for changed faces but not the one for identical faces).

Homogeneity of effect sizes was tested using the *Q* statistic (Hedges & Olkin, 1995), supplemented by the descriptive index of heterogeneity I (Shadish & Haddock, 2009). This value indicates the proportion of effect size variation due to heterogeneity between studies rather than to sampling error. Recommendations to interpret values of I are as follows: $I = .25$ is considered as small, $I = .50$ as medium, and $I = .75$ as large heterogeneity (Higgins & Thompson, 2002). However, I should not be interpreted in isolation but only in combination with the amount of true heterogeneity (see Borenstein, Higgins, Hedges, & Rothstein, 2017).

For the global estimate of effect sizes, both fixed effects models (FEM) and random effects models (REM) were calculated. The FEM assumes that a single effect size estimates the same fixed population parameter, with heterogeneity due to within-study sampling error. The REM assumes a random distribution of individual effect sizes that differ from the population mean by subject-level sampling error plus a random variance parameter (Lipsey & Wilson, 2001). Although we report both models, interpretation will focus on the REM results only.

Separate analyses were conducted for young faces, mixed-age faces (results for young and older faces not reported separately), and older faces. Some studies varied face age as a within-participants variable. These were, however, not combined in a single meta-analysis in order not to violate the independence assumption of effect size estimates.

To test for moderator effects, studies with young and mixed-age faces were analyzed together. Hierarchical weighted mixed-model (methods of moments) meta-regression analyses using macros developed by Lipsey and Wilson (2001) were conducted to test the effects of the mean age of the older age group and of face age (young vs. mixed-age) at step 1. At step 2, the number of faces at encoding, retention interval, change of pose and/or expression at test, and the presence of additional tasks beside face recognition were added to the model.

The nonparametric trim-and-fill method was applied to assess possible problems of publication bias (Duval & Tweedie, 2000a, 2000b; Sutton, 2009). Trim-and-fill analyses were carried out for young and mixed-age faces combined to increase the number of studies, and after removing outliers to avoid biased results. Effect sizes were plotted against standard errors to assess the distribution of the funnel plot (Sterne & Egger, 2001; Sutton, 2009). Based on the assumption that an asymmetric funnel plot arises from missing unfavorable and nonsignificant studies, the trim-and-fill method estimates the number of missing studies, imputes them in the funnel plot and estimates an "adjusted" effect size (Duval & Tweedie, 2000a, 2000b). Although trim-and-fill analyses could be carried out with an arbitrary amount of studies (Duval & Tweedie, 2000b), tests for funnel plot asymmetry would be unlikely to be useful in meta-analyses with small number of studies (Sterne, Becker, & Egger, 2005).

In addition to the graphical method, the statistical linear regression test was applied (Egger, Smith, Schneider, & Minder, 1997; Sutton, 2009). Thereby, the standard normal deviate is regressed against the inverse variance, which is similar to adjusting a line to Galbraith's radial plot. We also calculated the commonly used failsafe *N* (Orwin, 1983; Rosenthal, 1979) due to its pioneering role. However, we advise of recent criticisms of failsafe *N* in the meta-analytic literature (see Becker, 2005).

Meta-analyses and meta-regression analyses were conducted using SPSS Macros provided by Lipsey and Wilson (2001) and cross-validated with self-programmed Excel macros. For trim-and-fill analyses and forest plots, the *R* package 'metafor' was used (Viechtbauer, 2010).

Results

Weighted Mean Effect Size Analyses

Figures 1 to 3 display the forest plots of individual effect sizes and their 95% confidence intervals (CIs), as well as the estimated mean effects of fixed and random effects models with and without outliers. Effect sizes for young, middle-aged and older faces are separated to avoid dependencies among effect sizes and to appreciate possible own-age effects.

Author(s), Year [Exp.]	N	M_Y	M_O	Z	VC	Q	I^2	N_{fs}	Hedges g_u [95% CI]
Hits									
Young Faces									
Bryce & Dodson, 2013 [1]	48	0.83	0.81						0.16 [−0.41, 0.72]
Fulton & Bartlett, 1991	69	0.80	0.76						0.26 [−0.21, 0.74]
Mason, 1986	86	0.92	0.88						0.29 [−0.14, 0.71]
Lamont et al., 2005	96	0.79	0.73						0.41 [−0.02, 0.84]
He et al., 2011	49	0.72	0.64						0.46 [−0.10, 1.03]
Firestone et al., 2007	48	0.86	0.78						0.63 [0.05, 1.21]
Anastasi & Rhodes, 2006 [1]	74	0.93	0.71						1.26 [0.76, 1.77]
Anastasi & Rhodes, 2006 [2] #	88	0.88	0.63						1.42 [0.95, 1.89]
FE Model (k = 8)				6.90***		25.84**	.729	90	**0.61 [0.44, 0.79]**
RE Model (k = 8)				3.57***	.170	25.84**	.729	18	**0.61 [0.28, 0.95]**
FE Model (k = 7)				5.05***		12.54	.522	40	**0.48 [0.30, 0.67]**
RE Model (k = 7)				3.52***	.070	12.54	.522	16	**0.49 [0.22, 0.77]**
Mixed−age Faces									
Naveh−Benjamin et al., 2009 [1] #	47	0.72	0.78						−0.44 [−1.02, 0.14]
Bartlett et al., 1989 [2]	32	0.85	0.86						−0.04 [−0.73, 0.66]
Naveh−Benjamin et al., 2009 [2]	84	0.76	0.76						0.00 [−0.43, 0.43]
Bartlett et al., 1989 [1]	30	0.84	0.83						0.01 [−0.71, 0.72]
Bartlett & Fulton, 1991 [2]	70	0.78	0.77						0.13 [−0.34, 0.60]
Savaskan et al., 2007	45	0.35	0.32						0.41 [−0.21, 1.04]
D'Argembeau & van der Linden, 2004	64	0.73	0.65						0.47 [−0.02, 0.97]
Olofsson & Bäckman, 1996	96	0.91	0.84						0.55 [0.08, 1.01]
Bastin & van der Linden, 2003	128	0.72	0.64						0.55 [0.20, 0.90]
Bartlett et al., 1991 [1]	39	0.99	0.89						0.84 [0.19, 1.50]
FE Model (k = 10)				3.49***		16.28	.447	16	**0.29 [0.13, 0.45]**
RE Model (k = 10)				2.34*	.056	16.28	.447	3	**0.27 [0.04, 0.49]**
FE Model (k = 9)				4.06**		9.78	.182	25	**0.35 [0.18, 0.52]**
RE Model (k = 9)				3.55***	.015	9.78	.182	18	**0.34 [0.15, 0.53]**
Old Faces									
He et al., 2011	49	0.68	0.80						−0.79 [−1.37, −0.21]
Bryce & Dodson, 2013 [1]	48	0.70	0.76						−0.40 [−0.97, 0.17]
Fulton & Bartlett, 1991	69	0.78	0.82						−0.26 [−0.74, 0.21]
Lamont et al., 2005	96	0.79	0.82						−0.21 [−0.64, 0.21]
Mason, 1986	86	0.86	0.88						−0.14 [−0.56, 0.28]
Firestone et al., 2007	48	0.89	0.86						0.28 [−0.28, 0.85]
Anastasi & Rhodes, 2006 [1] #	74	0.91	0.79						0.85 [0.37, 1.33]
Anastasi & Rhodes, 2006 [2] #	88	0.90	0.70						1.25 [0.80, 1.71]
FE Model (k = 8)				1.30		51.72***	.865	10	**0.11 [−0.06, 0.29]**
RE Model (k = 8)				0.33	.397	51.72***	.865	0	**0.08 [−0.39, 0.55]**
FE Model (k = 6)				2.30*		7.25	.310	3	**−0.24 [−0.44, −0.03]**
RE Model (k = 6)				1.93	.029	7.25	.310	0	**−0.24 [−0.49, 0.00]**

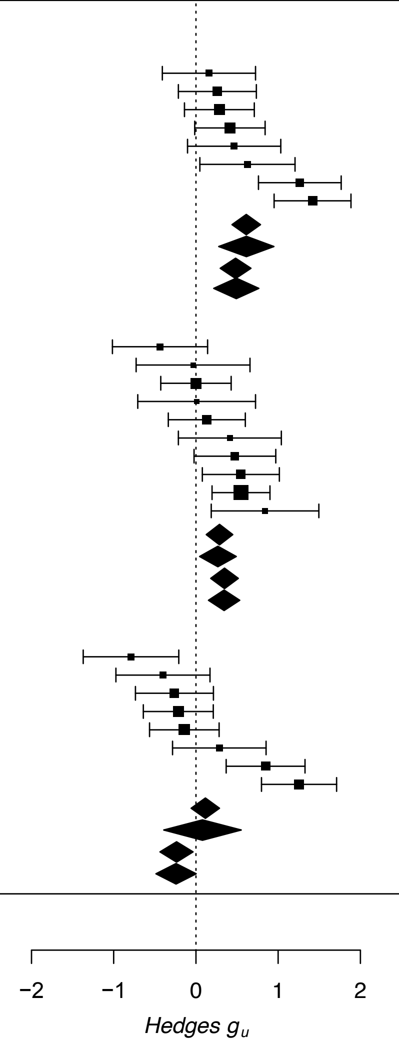


Figure 1. Forest plot for hits and the mean weighted effect sizes with and without outlier.

Author(s), Year [Exp.]	N	M_Y	M_O	Z	VC	Q	I^2	N_{fs}	Hedges g_u [95% CI]
False Alarms									
Young Faces									
He et al., 2011	49	0.28	0.34						0.24 [−0.33, 0.80]
Bryce & Dodson, 2013 [1]	48	0.05	0.15						0.79 [0.20, 1.37]
Lamont et al., 2005	96	0.17	0.30						0.86 [0.41, 1.30]
Fulton & Bartlett, 1991	69	0.25	0.41						0.96 [0.46, 1.46]
Firestone et al., 2007	48	0.08	0.22						0.99 [0.39, 1.59]
Anastasi & Rhodes, 2006 [2]	88	0.11	0.32						1.07 [0.63, 1.52]
Anastasi & Rhodes, 2006 [1]	74	0.03	0.15						1.23 [0.73, 1.73]
Mason, 1986 #	86	0.14	0.70						3.00 [2.38, 3.62]
FE Model (k = 8)				11.59***		48.57***	.856	281	**1.09 [0.90, 1.27]**
RE Model (k = 8)				4.55***	.421	48.57***	.856	35	**1.13 [0.63, 1.63]**
FE Model (k = 7)				9.17***		7.93	.243	140	**0.90 [0.71, 1.09]**
RE Model (k = 7)				7.88***	.022	7.93	.243	103	**0.89 [0.67, 1.12]**
Mixed−age Faces									
Naveh−Benjamin et al., 2009 [1] #	47	0.13	0.15						0.18 [−0.39, 0.75]
Olofsson & Bäckman, 1996	96	0.09	0.20						0.74 [0.27, 1.22]
Bastin & van der Linden, 2003	128	0.17	0.28						0.84 [0.48, 1.20]
D'Argembeau & van der Linden, 2004	64	0.21	0.34						0.90 [0.39, 1.42]
Bartlett & Fulton, 1991 [2]	70	0.28	0.40						0.94 [0.45, 1.43]
Savaskan et al., 2007	45	0.16	0.34						1.12 [0.45, 1.78]
Bartlett et al., 1991 [1]	39	0.04	0.15						1.25 [0.56, 1.94]
Naveh−Benjamin et al., 2009 [2]	84	0.22	0.42						1.42 [0.94, 1.90]
Bartlett et al., 1989 [1]	30	0.21	0.43						1.45 [0.65, 2.26]
Bartlett et al., 1989 [2]	32	0.26	0.58						1.78 [0.96, 2.59]
FE Model (k = 10)				11.24***		18.15*	.504	325	**0.97 [0.80, 1.14]**
RE Model (k = 10)				7.91***	.078	18.15*	.504	158	**1.01 [0.76, 1.26]**
FE Model (k = 9)				11.57***		10.09	.207	315	**1.05 [0.87, 1.22]**
RE Model (k = 9)				10.25***	.020	10.09	.207	247	**1.07 [0.86, 1.27]**
Old Faces									
Lamont et al., 2005	96	0.12	0.17						0.47 [0.04, 0.90]
He et al., 2011	49	0.34	0.44						0.49 [−0.07, 1.06]
Firestone et al., 2007	48	0.07	0.13						0.56 [−0.01, 1.14]
Anastasi & Rhodes, 2006 [1]	74	0.06	0.13						0.64 [0.16, 1.11]
Bryce & Dodson, 2013 [1]	48	0.06	0.15						0.81 [0.22, 1.40]
Fulton & Bartlett, 1991	69	0.35	0.50						0.89 [0.39, 1.38]
Anastasi & Rhodes, 2006 [2]	88	0.08	0.28						1.14 [0.68, 1.59]
Mason, 1986 #	86	0.14	0.63						2.39 [1.84, 2.94]
FE Model (k = 8)				9.83***		37.27***	.812	197	**0.90 [0.72, 1.08]**
RE Model (k = 8)				4.33***	.291	37.27***	.812	31	**0.92 [0.50, 1.34]**
FE Model (k = 7)				7.49***		6.07	.012	93	**0.72 [0.53, 0.91]**
RE Model (k = 7)				7.45***	.001	6.07	.012	91	**0.72 [0.53, 0.91]**

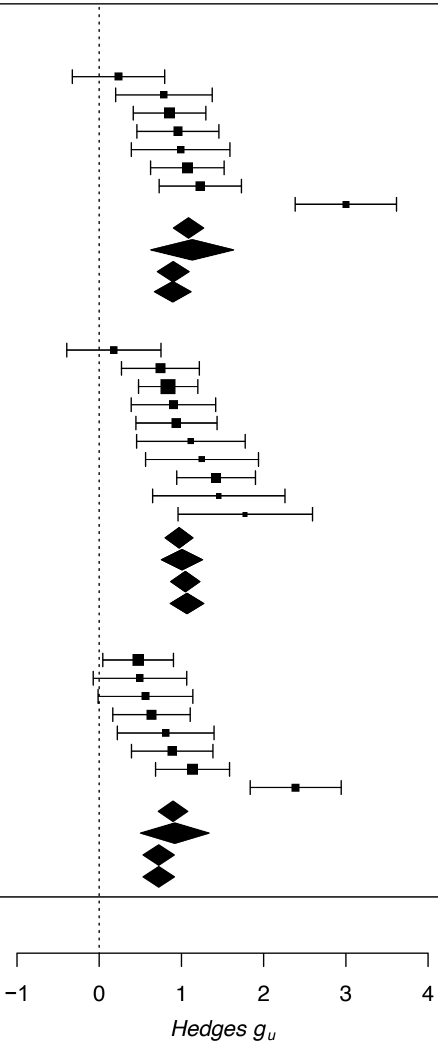


Figure 2. Forest plot for false alarms and the mean weighted effect sizes with and without outlier.

Author(s), Year [Exp.]	N	Type	M_Y	M_O	Z	VC	Q	I^2	N_{fs}	Hedges g_u [95% CI]
Recognition Performance										
Young Faces										
Lamont et al., 2005	96	*d'*	2.03	1.35						0.90 [0.45, 1.34]
Fulton & Bartlett, 1991	69	*d'*	1.76	1.08						0.93 [0.43, 1.43]
He et al., 2011	49	*d'*	1.99	1.07						1.22 [0.61, 1.83]
Firestone et al., 2007	48	*d'*	2.36	1.59						1.44 [0.80, 2.07]
Mason, 1986 #	86	*d'*	3.47	1.06						1.99 [1.47, 2.51]
FE Model (k = 5)					10.43***		12.24*	.673	139	**1.25 [1.02, 1.49]**
RE Model (k = 5)					6.02***	.151	12.24*	.673	43	**1.28 [0.87, 1.70]**
FE Model (k = 4)					7.86***		2.40	.000	63	**1.06 [0.80, 1.32]**
RE Model (k = 4)					7.86***	.000	2.40	.000	63	**1.06 [0.80, 1.32]**
Mixed–age Faces										
D'Argembeau & van der Linden, 2004	64	*d'*	1.99	0.96						0.71 [0.21, 1.22]
Bastin & van der Linden, 2003	128	*d'*	1.70	1.10						0.86 [0.50, 1.23]
Bartlett et al., 1989 [1]	30	*d'*	2.18	1.67						0.93 [0.17, 1.68]
Smith & Winograd, 1978	160	*d'*	2.05	1.30						1.01 [0.67, 1.35]
Savaskan et al., 2007	45	*d'*	1.60	0.97						1.11 [0.45, 1.77]
Bartlett et al., 1989 [2]	32	*d'*	1.05	0.50						1.17 [0.42, 1.91]
Olofsson & Bäckman, 1996	96	*d'*	3.26	2.18						1.28 [0.78, 1.78]
FE Model (k = 7)					10.57***		3.30	.000	184	**0.98 [0.80, 1.16]**
RE Model (k = 7)					10.57***	.000	3.30	.000	184	**0.98 [0.80, 1.16]**
Old Faces										
Lamont et al., 2005	96	*d'*	2.24	2.25						−0.01 [−0.44, 0.41]
He et al., 2011	49	*d'*	1.42	1.24						0.29 [−0.27, 0.85]
Fulton & Bartlett, 1991	69	*d'*	1.30	1.05						0.36 [−0.11, 0.84]
Firestone et al., 2007	48	*d'*	2.59	2.16						0.65 [0.07, 1.23]
Mason, 1986	86	*d'*	2.28	1.20						0.78 [0.34, 1.21]
FE Model (k = 5)					3.57***		7.33	.455	12	**0.40 [0.18, 0.61]**
RE Model (k = 5)					2.67**	.052	7.33	.455	4	**0.40 [0.11, 0.70]**
Response Criterion										
Young and Mixed–age Faces										
D'Argembeau & van der Linden, 2004	64	*c*	0.03	0.04						0.00 [−0.49, 0.49]
Olofsson & Bäckman, 1996	96	*c*	−0.01	−0.11						0.25 [−0.21, 0.71]
Bastin & van der Linden, 2003	128	*c*	0.20	0.11						0.27 [−0.08, 0.62]
Firestone et al., 2007	48	*c*	0.23	0.01						0.27 [−0.29, 0.84]
Smith & Winograd, 1978	218	..	1.35	0.85						0.39 [0.12, 0.65]
Fulton & Bartlett, 1991	69	*c*	−0.12	−0.34						0.63 [0.15, 1.11]
Bartlett et al., 1991	39	*c*	N/A	N/A						0.72 [0.07, 1.37]
Mason, 1986	86	..	2.61	0.72						1.02 [0.57, 1.47]
Bartlett et al., 1989 [1]	30	*c*	−0.24	−0.65						1.17 [0.40, 1.95]
Bartlett et al., 1989 [2]	32	*c*	−0.44	−0.81						1.19 [0.43, 1.94]
FE Model (k = 10)					6.46***		19.97*	.549	111	**0.47 [0.33, 0.61]**
RE Model (k = 10)					4.52***	.069	19.97*	.549	48	**0.52 [0.29, 0.76]**

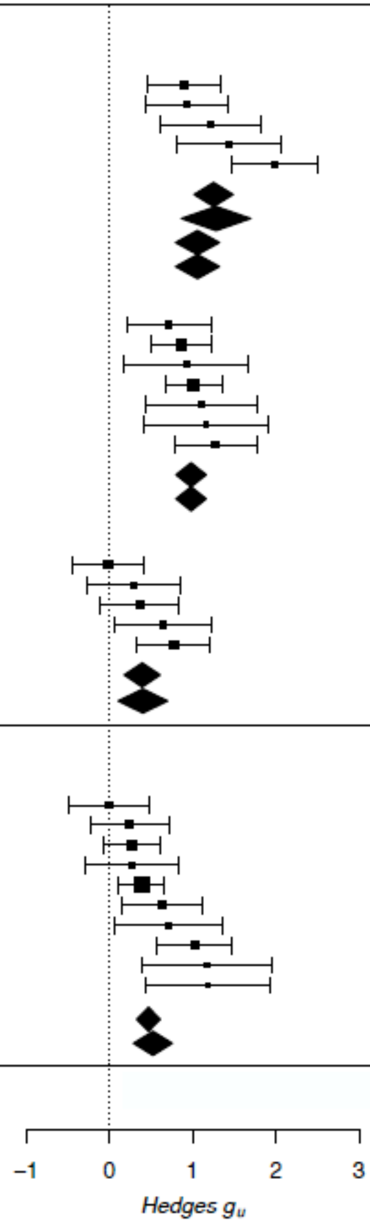


Figure 3. Forest plot for recognition performance d' and response criterion c (or beta), and the mean weighted effect sizes with and without outliers.

Before synthesizing effect sizes, two different methods were applied to check for outliers for each dependent variable: (a) graphic representations of individual effect sizes (and 95% CIs), along with the weighted mean effect sizes (see the forest plots in Figures 1-3); (b) Hedges and Olkin's (1985, Chapter 12) method of calculating standardized residuals and homogeneity tests, excluding potential outliers consecutively using a self-programmed Excel spreadsheet.

Next, FEMs and REMs were calculated for young, mixed-age and older faces. Below we focus on the REMs, which were calculated both for all reported studies and after removal of outliers if necessary. Results without outliers provide a somewhat more conservative estimate. Figure 4 provides an overview of all mean effect sizes (with 95% CIs) using the REM without outliers.

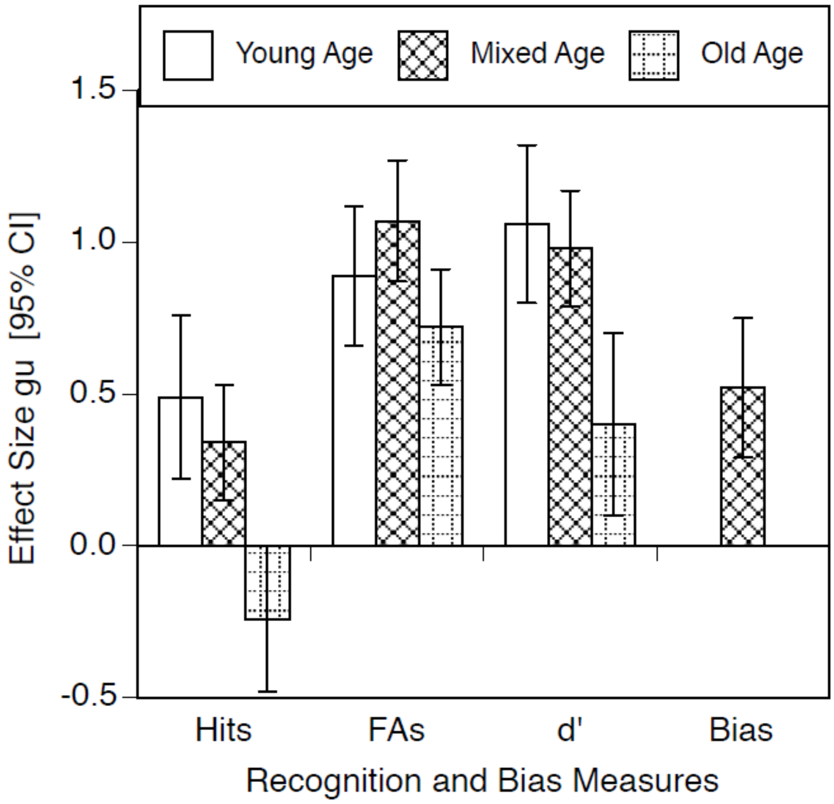


Figure 4. Summary of mean weighted effect sizes (without outliers) using the random effects model for hits, false alarms, recognition performance and response bias for young, mixed-age and old-age faces.

Hits.

Overall, 26 hypothesis tests were identified reporting age effects on hits (8 for young, 10 for mixed-age, and 8 for older faces; Figure 1). Outlier analyses indicated one extreme case for young faces (Anastasi & Rhodes, 2006, Exp. 2, one for mixed-age faces (Naveh-Benjamin et al., 2009, Exp. 1), and two for older faces (Anastasi & Rhodes, 2006, Exp. 1 and 2; see Figure 1). The experiments by Anastasi and Rhodes (2006) differed from the other studies in several ways: (1) the encoding task was incidental, asking participants to rate the face age as young, middle-aged or old (Exp. 1); (2) an age rating was used as the basis for categorizing faces into same-age/other-age group at the retrieval phase, instead of using objective age ranges; (3) pose of target faces was changed between encoding and recognition phase; (4) retention interval was 48 hours in Experiment 2, and therefore much longer than in any other study included. The difference in encoding task and the much longer retention interval may have contributed to the extreme outcomes. Naveh-Benjamin et al. (2009, Exp. 1) instructed their participants to study 24 face-name pairs and tested them (1) with a Yes-No face recognition task, (2) a name recognition task, and (3) recognition of face-name pairs. It is not clear to us why this study produced a negative age effect despite the complexity of these multiple tasks.

After outliers were removed, homogeneity tests were no longer significant (*p*s > .05). The mean weighted effect size indicated higher hits for young participants than their older counterparts when viewing young faces, $g_u = 0.49$ [0.22, 0.76], $p < .001$, $VC = .07$, and to a lesser extent when viewing mixed-age faces, $g_u = 0.34$ [0.15, 0.53], $p < .001$, $VC = .02$. A marginally significant negative effect was found between the age groups for older faces, $g_u = -0.24$ [-0.49, 0.00], $p = .053$, $VC = .03$. While younger participants had higher hits than their older counterparts when viewing young and mixed-age faces, the effect reversed such that older participants tended to have higher hits when viewing older faces.

Three studies reported hits for middle-aged faces. Therefore, no separate meta-analysis was conducted. Anastasi and Rhodes (2006) observed higher hits for young compared to older adults (Exp. 1: $g_u = 1.05$ [0.56, 1.55]; Exp. 2: $g_u = 1.26$ [0.81, 1.72]), whereas Fulton and Bartlett (1991) found no differences between age groups (g_u = 0.10 [-0.37, 0.58] for identical, and g_u = 0.06 [-0.42, 0.53] for changed faces).

False alarms.

Twenty-six hypothesis tests were identified to test the effects on false alarms (8 for young, 10 studies for mixed-age, and 8 for older faces; see Figure 2). Outlier analyses revealed one extreme case for young and for older faces, respectively (Mason, 1986; see Figure 2), and one extreme case for mixed-age faces (Naveh-Benjamin et al., 2009, Exp. 1; see Figure 2). In Mason's study (1986), the task to study not only faces but also their associated names within a very limited time frame, and the large number of faces used, may explain why older adults had many more false alarms than young adults. The study by Naveh-Benjamin et al. (2009, Exp. 1) was described above.

After excluding outliers homogeneity tests were no longer significant, *p*s > .05. Young participants made fewer false alarms than older participants when recognizing young, g_u = 0.89 [0.67, 1.12], $p < .001$, *VC* = .02, mixed-age, g_u = 1.07 [0.86, 1.27], $p < .001$, and older faces, g_u = 0.72 [0.53, 0.91], $p < .001$, *VC* = .00.

Finally, for the three studies using middle-aged faces older participants had significantly higher false alarms than their younger counterparts (Anastasi & Rhodes, 2006, Exp. 1: g_u = 0.90 [0.41, 1.38]; Anastasi & Rhodes, 2006, Exp. 2: g_u = 1.12 [0.67, 1.57]; Fulton & Bartlett, 1991: g_u = 0.96 [0.46, 1.46].

In summary, regardless of target age, older adults were more likely to incorrectly choose a new face as previously seen than young adults (Figures 2 and 4).

Recognition performance.

For signal detection measures of recognition performance (*d'*) five studies were included for young, seven for mixed-age, and five for older faces (Figure 3). For young faces, Mason (1986) was identified as an outlier (see above) and removed from the analyses (see Figure 3). The analyses showed large effects in favor of young participants for young faces: g_u = 1.06 [0.80, 1.32], $p < .001$, *VC* = .00, and mixed-age faces: g_u = 0.98 [0.80, 1.17], $p < .001$, *VC* = .05, and a medium effect for old faces: g_u = 0.40 [0.11, 0.70], $p < .001$, *VC* = .05. Parallel to the results for hits and false alarms, young adults outperformed their older counterparts in face recognition performance, regardless of target age. However, the differences between the age groups were smaller for older faces than for young and mixed-age faces.

Response bias.

Ten studies reported enough information to calculate the mean effect size for response bias for combined young and mixed-age faces (see Figure 3). Outlier analyses revealed no outliers. The mean weighted effect size was positive, $g_u = 0.52$ [0.30, 0.75], $p < .001$, $VC = .07$, indicating that older participants used a more lax decision criterion than their younger counterparts. In other words, compared to young adults, older adults were more likely to choose a face as previously seen, regardless of whether the face was a target or a distractor.

Mirror Effect in Face Recognition

To further grasp the sensitivity and specificity of the age groups, hits (sensitivity) and false alarms (1 – specificity) were plotted against each other (see Figures 5a and 5b). Effects for all face age groups were included (young, middle-aged, older and mixed-age faces). Savaskan et al. (2007) was excluded as an outlier as the hits were unusually low for both age groups ($M_y = .35$; $M_o = .32$; see Figure 1). Young adults showed a negative relationship between hits and false alarms, indicating that with an increase in hits false alarms decreased (Figure 5a), thus demonstrating the expected mirror-effect pattern. In contrast, the relationship was reversed for older adults. Higher hits were not necessarily associated with lower false alarms (Figure 5b). Instead there was either no or a positive relationship between hits and false alarms for older adults, indicating that they did not discriminate well between targets and distractors. Only for young-aged faces Figure 5b implies a mirror effect, whereas for middle-aged, mixed or older faces there appears to be a positive association which would indicate a lenient response bias.

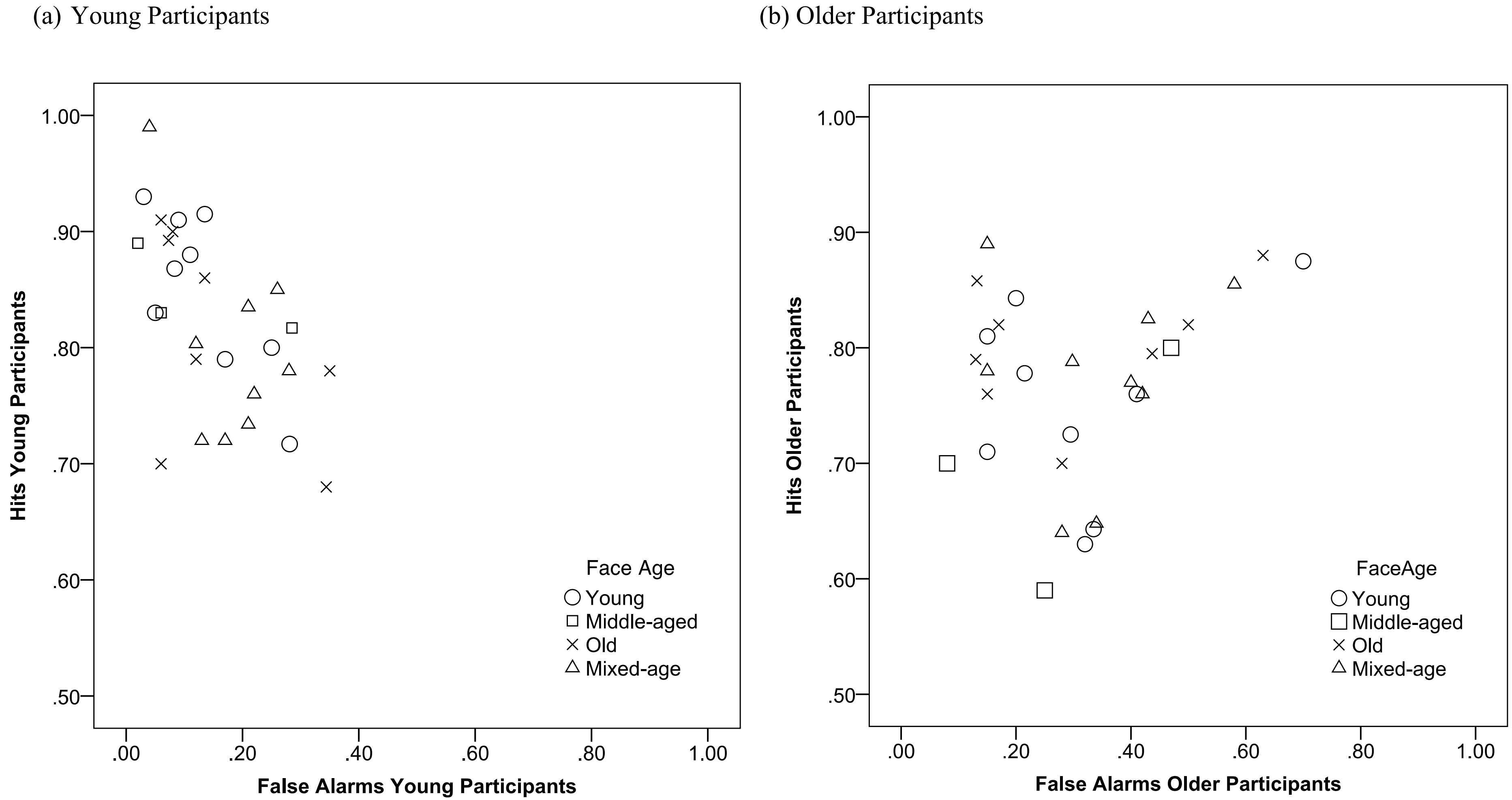


Figure 5a and 5b. Scatterplots of the correlations between hits and false alarms of young (Panel a) and older adults (Panel b).

Moderator Analyses with Meta-regression

The meta-analyses reported above were carried out separately for hits, false alarms and signal detection measures of performance and bias. We also separated analyses for young, mixed-age, and old faces. The latter separation was necessary because all studies that used old faces also used young faces (to test for an own-age bias). Because studies with both younger and older faces used identical methodologies their effect sizes are stochastically dependent (see Gleser & Olkin, 2009), and hence should not be pooled with studies using different methodologies that yield independent effect sizes.[2]

To avoid statistical dependencies, we pooled studies with independent data for young faces and studies with mixed-age faces—none of the studies used both—to conduct moderator analyses.[3]

Most meta-analyses in the (eyewitness and face recognition) literature have used separate blocking analyses (as an analogue to ANOVAs; see Lipsey & Wilson, 2001) with different moderator variables as predictors. However, this approach has been criticized because it does not adequately take the *intercorrelations of predictor variables* into account. To demonstrate, Table 2 displays the intercorrelations of potential moderators (encoding time (in seconds), retention interval (in minutes), number of faces at encoding, number of faces at

---

[2] Because the correlations between the dependent variables were not reported we could not estimate effect sizes taking these dependencies into account. We also decided against robust variance estimation methods (Hedges, Tipton, & Johnson, 2010) as an alternative due to the relatively low number of studies and the ensuing lack of power (a minimum of 40 studies are needed for the robust variance estimation when dependent and nondependent effects are included in the model; Hedges et al., 2010).

[3] For signal detection measures, the number of studies (k = 12 for d' and k = 10 for c or *beta*) was too small (and heterogeneity too small after removal of outliers; see Figure 3) to warrant meta-regression analyses. For SDT performance, the mean effect size was g_u = 1.01 [0.86, 1.16], Z = 13.16, $p < .001$, VC = .000, $Q(10)$ = 5.92, p = .822 (after removing Mason, 1986, as an outlier). For response bias, the mean effect size is reported in Figure 3.

test, change of view (no change, 1 change, 2 changes), as well as the weighted semi-partial and zero-order correlations between the effect sizes (hits, false alarms) and the predictor variables used in the meta-regressions. As is apparent from Table 1, many studies shared some features, for example, using a longer encoding time when a large number of faces was used. Thus, such similarities lead to confounded predictors. While some of the correlations were small, they nonetheless indicated that ignoring them would introduce confounds in blocking analyses (Pigott, 2012). However, the intercorrelations were not too high to create problems of multicollinearity.

As Table 2 shows, the number of faces at encoding were highly correlated with the number of faces at test, thus making the latter redundant as a predictor. Less obvious, and hence a potential confound in moderator analyses, are the medium to high correlations of encoding time with number of faces at test and with the number of changes at test. Presumably, researchers used longer encoding times to compensate when the task became more difficult.

Further support for the interdependencies of the predictor variables became apparent from the semi-partial correlations, which were calculated using ordinary least squares weighted regression analyses to partial out shared variance with other predictor variables. Compared to zero-order correlations, some of the semi-partial correlations differed considerably in strength and direction. For instance, although the correlation between hits and number of faces at encoding was large (r = -.49), the semi-partial correlation that controlled for other predictors was zero (sr_k = -.02). The difference was reversed for change of faces: While the small correlation (r = -.08) increased considerably for hits (sr_k = -.24), the medium to large correlation (r = .40) decreased to a small effect for false alarms after controlling for other variables (sr_k = .14).

Table 2

Intercorrelations of Predictor Variables for all Studies (excluding Smith & Winograd, 1978) and Weighted Semi-partial (and Zero-order) Correlations for Hits and False Alarms

Variable	Mean Age of Older Age Group	Face Age	Encoding Time	Retention Interval	# of Faces at Encoding	# of Faces at Test	# of Foils at Test	Change of Faces	Additional Tasks
Mean Age of Old Age Group	1.00								
Face Age (Young vs. Mixed)	**-.43**	1.00							
Encoding Time (s)	.04	-.01	1.00						
Retention Interval (min)	-.08	-.26	.16	1.00					
# of Faces at Encoding	.20	.21	.17	-.28	1.00				
# of Faces at Test	.13	-.22	**.45**	-.08	**.67**	1.00			
# of Foils at Test	.24	-.40	.34	.01	.36	**.87**	1.00		
Changes of Pose and/or Expression at Test	-.25	.35	**.62**	.09	.15	.26	.12	1.00	
Additional Task(s) besides Face Recognition	**.43**	.00	-.08	-.20	-.03	-.36	-.31	-.34	1.00
Effect Size Hits	-.29 (-.22)	-.30 (-.24)	.07 (.09)	**.49 (.58)**	-.02 (-.49)	__[a]	__[a]	-.24 (-.08)	-.24 (-.39)
Effect Size False Alarms	-.30 (-.25)	-.01 (-.19)	.15 (.34)	.11 (.13)	.30 (.34)	__[a]	__[a]	.14 (.40)	.11 (-.12)

Note. Significant (with *df* = 16) correlations > .40 are highlighted in bold face. Semi-partial correlations are reported only for predictor variables included in meta-regressions. [a] These two variables were removed from meta-regression to avoid collinearity and to reduce the number of predictors in the model.

To avoid these confounds, we only used weighted meta-regression analyses which statistically control for mutual pairwise as well as multivariate dependencies in predictors. This was only possible for studies reporting recognition of young and mixed-age faces which we pooled to obtain a sufficient number of studies to conduct meta-regression analyses.[4] Predictor variables were entered as blocks in hierarchical meta-regressions, after the outliers noted above were removed.[5]

Because we pooled studies using both young faces and mixed-age faces, in step one we included both the mean age of participants in the old age group and age of faces (young vs. mixed-age) as predictors (see Tables 3 and 4). If there is an own-age bias, effect sizes should be larger for recognition of young faces compared to those using mixed-age faces. In step 2 we entered methodological characteristics described in Table 1 into Tables 3 and 4 as predictors.

Hits.

A total of 16 effect sizes were available for the meta-regression analyses for hits (without the two outliers by Anastasi & Rhodes, 2006, Exp. 1 and 2). The mean weighted effect size was $g_u = 0.31$ [0.17, 0.45], $Z = 4.31$, $p < .001$, $VC = .015$, $Q(15) = 18.34$, $p = .245$.

Results of the mixed-model meta-regression for hits are presented in Table 3. At step 1, the effect sizes for hits were surprisingly larger when the mean age of the older participants was younger and for studies using young (vs. mixed-age) faces. That is, the higher the mean age of the older participants the smaller was the difference between them and younger adults. At the same time, young adults did not perform as well when they viewed mixed-age faces compared to young faces only.

---

[4] It is not quite clear what the minimum number of studies should be to conduct meta-regression analyses but there is serious concern about the lack of power of meta-regression analyses when *k* is small (Hedges & Pigott, 2004).

[5] Although including the outliers would result in significant results for some of the predictors we decided to exclude them so results would not be distorted by one or two outliers.

Table 3

Hierarchical Weighted Mixed-Effect Meta-regression for Hits for Young and Mixed-Age Faces (k = 16)

Study characteristics			*B*	*beta*	*p*	R^2
Step 1:						
Model	$Q(2) = 5.62$,	$p = .060$				.306
Residual	$Q(13) = 12.72$,	$p = .470$				
Mean Age of Old Age Participant Group			**-.05**	**-.66**	**.033**	
Face Age (Young vs. Mixed)			**-.37**	**-.68**	**.029**	
Step 2:						
Model	$Q(7) = 14.90$,	$p = .037$				.812
Residual	$Q(8) = 3.44$,	$p = .904$				
Mean Age of Old Age Participant Group			-.05	-.60	.208	
Face Age (Young vs. Mixed)			-.30	-.54	.203	
Encoding Time (s)			.01	.13	.756	
# of Faces at Encoding			-.00	-.04	.917	
Retention Interval			**.03**	**.63**	**.038**	
Changes of Pose and/or Expression at Test			-.13	-.37	.314	
Additional Task(s) besides Face Recognition			-.19	-.33	.301	
Regression constant			4.16	.00	.137	

Note. *Q* statistics shown are from mixed-effects meta-regression models and test whether the independent variables (1) mean age of the old-age participant group and face age, and (2) study characteristics account for an additional portion of the variance in effect sizes. Two outliers (Anastasi & Rhodes, 2006, Exp. 1 and 2) were removed from the model. Before removing the outliers: Step 2 Model $Q(7) = 44.79$, $p < .001$; Residual $Q(10) = 4.55$, $p = .919$; $R^2 = .908$. Significant results are highlighted in boldface.

These effects were no longer significant at step 2 when other variables were entered as predictors and thus controlled for, indicating that other factors mitigated the effect of mean age of the older group and of face age. The only significant predictor was retention interval: With a slightly longer retention interval, effect sizes for hits were higher than when tests followed more or less immediately the encoding phase. Although R^2 was very high (R^2 = 81.2% explained variance, as opposed to R^2 = 30.6% at step 1), none of the other predictors was significant. This may be due to mutually controlling these factors (or to low power). Particularly in the case of low statistical power (many predictors with few studies) moderator effects are difficult to detect (Hedges & Pigott, 2004).

False Alarms.

After removing one outlier (Mason, 1986), the mean weighted effect size was $g_u = 0.95$ [0.78, 1.12], $Z = 11.17$, $p < .001$, $VC = .047$, $Q(16) = 26.37$, $p = .049$. None of the predictors (the same as used for hits above) at step 1 and at step 2 yielded significant contributions (see Table 4).

Table 4

Hierarchical Weighted Mixed-Effect Meta-regression for False Alarms for Young and Mixed-Age Faces (k = 17)

Study characteristics			*B*	*beta*	*p*	R^2
Step 1:						
Model	$Q(2) = 1.80$,	$p = .406$				.112
Residual	$Q(14) = 14.36$,	$p = .423$				
Mean Age of Old Age Participant Group			-.03	-.26	.336	
Face Age (Young vs. Mixed)			.09	.13	.624	
Step 2:						
Model	$Q(7) = 6.02$,	$p = .537$				.408
Residual	$Q(9) = 8.74$,	$p = .462$				
Mean Age of Old Age Participant Group			-.05	-.42	.247	
Face Age (Young vs. Mixed)			.00	.01	.987	
Encoding Time (s)			.03	.24	.517	
# of Faces at Encoding			.01	.33	.274	
Retention Interval			.00	.11	.713	
Changes of Pose and/or Expression at Test			.10	.21	.588	
Additional Task(s) besides Face Recognition			.08	.11	.747	
Regression constant			3.70	.00	.215	

Note. Q statistics shown are from mixed-effects meta-regression models and test whether the independent variables (1) mean age of the old-age participant group and face age, and (2) study characteristics account for an additional portion of the variance in effect sizes. One outlier (Mason, 1986) was removed from the model. Before removing the outlier: Step 2 Model $Q(7) = 5.45$, $p = .606$; Residual $Q(10) = 10.38$, $p = .407$; $R^2 = .344$.

Publication Bias

Sensitivity analyses were conducted to test and adjust for publication bias with the nonparametric trim-and-fill method by plotting effect sizes against standard errors (Duval & Tweedie, 2000a, 2000b). Figures 6a to 6c present the funnel plot for (a) hits, (b) false alarms, and (c) recognition performance for young and mixed-age faces with the imputed studies (circled dots), if applicable. The dashed vertical line indicates the mean weighted effect without the imputed studies.

The funnel plot in Figure 6a showed a symmetric distribution of the effect sizes for hits. The trim-and-fill analyses confirmed the funnel plot analyses, namely absence of a publication bias for hits. The fail-safe N revealed that $k = 74$ studies using the fixed-effects model or $k = 55$ studies using the random-effects model would be necessary to render the effects nonsignificant for hits.

By contrast, the funnel plot in Figure 6b showed an asymmetric tendency for false alarms and in Figure 6c for recognition performance, indicating smaller effect sizes missing at the left. The trim-and-fill method suggested the possibility of four missing studies on the left side of the funnel for false alarms and of three missing studies for recognition performance. The re-estimated weighted mean effect size for false alarms was $g_u = 0.85$ [0.73, 0.97], $Z = 14.06$, $p < .001$, $k = 21$, which only little deviates from the original value of $g_u = 0.94$ [0.81, 1.07]. The fail-safe N indicated that $k = 909$ using the fixed-effects model or $k = 544$ using the random-effects model would be required to make the effect for false alarms nonsignificant. For recognition performance, the re-estimated weighted mean effect was $g_u = 0.93$ [0.79, 1.07], $Z = 13.35$, $p < .001$, $k = 14$, again only slightly smaller than the original value of $g_u = 1.01$ [0.86, 1.16]. The fail-safe N for SDT performance was $k = 472$ using both fixed- and random-effects models.

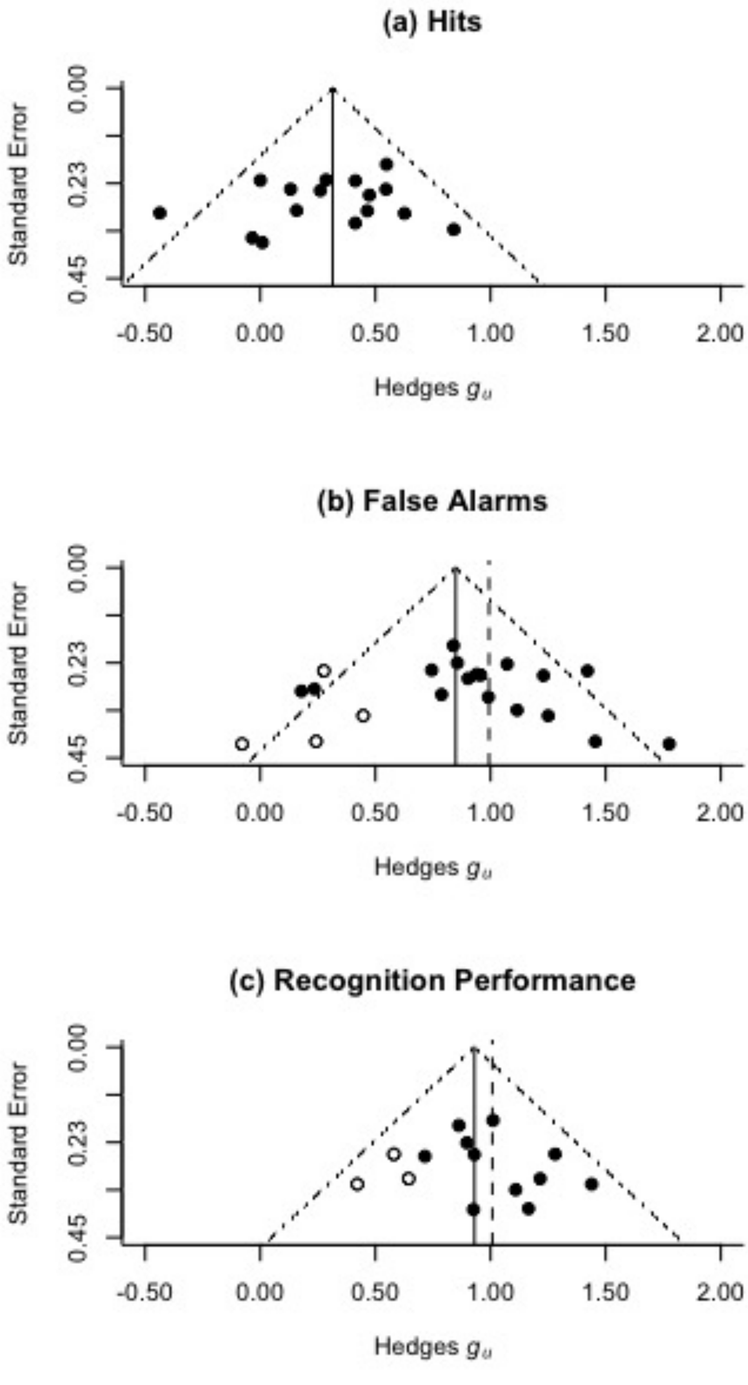


Figure 6a, 6b, and 6c. Funnel plots and trim-and-fill analyses for (a) hits, (b) false alarms, and (c) recognition performance without outliers. Solid circles are included studies, open circles are imputed studies. The solid vertical line denotes the re-estimated mean effect size, while the dashed vertical line indicates the original mean weighted effect size.

Discussion

Different theoretical approaches discussing age-related changes in the sensory system (Lott et al., 2005; Schieber, 2006), in brain structure (Coleman & O'Hanlon, 2008), and in cognitive processes and processing speed (Craik, 1986; Salthouse, 1996) have commonly been invoked to account for deteriorating memory performance at old age. These factors can also be assumed to affect face recognition performance with increasing age. Although comprehensive narrative reviews of face recognition studies of young versus older people exist (Bartlett, 2014a; Bartlett & Memon, 2007), little was known about the size of these effects.

The present meta-analysis examined age effects on recognition memory for faces containing 79 comparisons between young and older adults on different dependent measures (hits, false alarms and signal detection measures *d'* and *c*). Included studies focused solely on the recognition of unfamiliar faces without a complex event. As hypothesized, young people outperformed their older counterparts, showing small to medium effects for hits and large effects for false alarms and SDT performance measures (small: $g_u = 0.2$; medium: $g_u = 0.5$; large: $g_u \geq 0.8$; according to Cohen's, 1988, recommendations). The effects observed here are even more noteworthy in light of the large-scaled meta-analysis of 128 studies by Shapiro and Penrod (1986) who found much smaller effects for most of the variables investigated in the literature. Our effects were even larger than those for the own-race bias in Meissner and Brigham's (2001) meta-analysis who found a mean weighted $g_u = 0.24$ for hits and $g_u = 0.39$ for false alarms. Thus, the effects observed in our meta-analysis can be considered quite robust, in particular because the effects were largely homogeneous once outliers had been removed. Notably, in all studies reviewed older participants had been prescreened to exclude any persons which showed more general cognitive or memory deficits. Thus, our estimates are likely *underestimates* compared to more representative populations. Our results also show that age effects are not only present in recall and cued recall but also in recognition memory paradigms.

Although the effects on hits were not quite as high as those on false alarms, the general pattern followed a mirror effect as postulated by signal detection theory (Glanzer & Adams,

1985) that had also been observed in the literature on the cross-race effect. However, the mirror effect found here was not quite symmetric, and, interestingly, only present with young participants who showed the expected negative association between hits and false alarms across studies (Figure 5a). With older participants, the pattern was non-existent or tended to be even positive, depending on the age of the to-be-recognized faces (Figure 5b). However, these conclusions are based on visual inspection of the graphs only and could not be tested with inferential statistics because the original studies did not report the necessary intercorrelations.

Age effects differed as a function of face age: Differences between young and older adults were larger for young and mixed-age faces than for older faces, and reversed for hits for older faces, indicating the presence of an own-age bias. While younger adults outperformed older adults when recognizing young faces, older adults were better at correctly recognizing older faces. Analogous own-group advantages have been well documented regarding the own-race bias (Meissner & Brigham, 2001; Sporer, 2001). Also analogous to the own-race bias, older people displayed a more liberal response criterion than younger people, with a medium effect size. In summary, compared to older people young people were better in discriminating new faces from previously seen faces and less likely to accept new faces as previously seen.

We further explored potential effects of moderator variables. Due to small to medium size intercorrelations between predictor variables, we employed meta-regression analyses that control for confounding effects (rather than using a series of blocking analyses as is frequently found in the meta-analytic literature; see Pigott, 2012).

Surprisingly, the above-mentioned age effect became weaker when mean age of the older participants was taken into consideration: The higher the age of older people the less was the difference in hits between young and older people. This result contradicts previous research on biological markers of cognitive functioning, which found a decline in memory that is more rapid in older age (Anstey, 2011). One possible explanation are possible confounds inherent in the experimental designs across studies: Mean age of older people was negatively correlated with face age (young vs. mixed-age) (r = -.43). Thus, it seemed that the higher the age of older people the more likely were the primary studies to report results for young and older faces separately. Yet when mean age of older people was controlled for, face age remained a significant predictor, indicating the presence of an own-age bias (Rhodes & Anastasi, 2012).

Perhaps surprisingly, retention interval was the only moderator to reach statistical significance in the meta-regression with hits when controlling for the remaining variables (face age, mean age of older participants, encoding time, number of faces at encoding, changing between encoding and test, and additional tasks at test). Longer retention intervals were associated with larger differences between the age groups for hits but not for false alarms. Considering the different mirror effect patterns of young and older people shown in Figures 5a and 5b, these differences between hits and false alarms indicate that the recognition process and forgetting over time may operate differently in young versus older age. However, practically all retention intervals were minimal in length (less than one hour), clearly indicating the need to study longer retention intervals of days, weeks, or months in future studies. To test whether memory decline over time is linear or curvilinear, future studies should measure recognition at least at three points in time (Deffenbacher, Bornstein, McGorty, & Penrod, 2008; Rubin & Wenzel, 1996), perhaps also with within-participants designs.

None of the other moderators were associated with hits, and none of the moderators were associated with false alarms. These findings, however, do not necessarily rule out moderator effects, considering the relatively low number of studies available and the ensuing low power of our moderator analyses (Hedges & Pigott, 2004).

Recognition Performance

In addition to separate face recognition measures, the current meta-analysis considered sensitivity of recognition performance, namely the SDT discrimination index *d'*, which simultaneously takes both hits and false alarms into account (Macmillan & Creelman, 2005). From an applied perspective, the SDT performance measure is of greater importance than the individual recognition measures hits and false alarms. In other words, making conclusions about someone's face recognition performance based solely on hits would be misleading. If false alarms increase along with hits, this would only indicate a change in response bias but not in recognition performance.

The age groups differed by one standard deviation in *d'* demonstrating a clear advantage of young age. Discrimination between old and new faces, therefore, seemed to be significantly impaired with age. The reduced processing speed hypothesis (Salthouse, 1996) proposes that mental speed declines with age. Cognitive processes that require working memory may therefore decline with age, which results in poorer performance and inaccurate memory of older people (Maylor, 2005). Salthouse (1996) showed that reduced processing speed in age led to decreased ability to accomplish tasks or retrieve relevant information. Similarly, Wolff et al. (2012) demonstrated that face-processing speed declined with age, with the strongest decline after 60. As already indicated in the introduction, meta-analyses cannot provide answers regarding the types of face information extracted at encoding and used in decision processes and the role working memory plays. We do encourage future researchers to test participants regarding their (visual) working memory capacity and to assess response latencies to gain further insight into these processes.

Response Bias

The response criterion measures a participant's tendency to denote a face as "seen" or "not seen" before. As a consequence of setting a more liberal criterion, an individual might correctly recognize most of the faces presented because of his or her proneness to choose. In contrast, another individual might operate more cautiously or conservatively at the face recognition task, leading not only to fewer false alarms, but also fewer hits. In fact, the tendency towards a conservative or liberal response criterion and a tendency to false memory has been linked to personality traits (Kantner & Lindsay, 2013), and to age-related structural changes in the frontal brain lobe (Huh et al., 2006).

Overall, the tendency for a liberal response seemed to be a robust effect in older age, which we could only test for young and middle-aged faces. Older adults may rely more on a generalized sense of familiarity (Hancock et al., 2000; Vokey & Read, 1992) when deciding whether or not they have seen a face before. Future studies should routinely investigate response bias with faces of different ages. More frequent choosing rates were also observed with older compared to young people in studies using a lineup paradigm in the meta-analysis by Kocab et al. (2017), and in a large scale field study by Martschuk et al. (2018).[6]

Own-age Bias

Social-psychological approaches to face recognition stress the importance of in-group and out-group categorization of the target face for later recognition (Hugenberg et al., 2007; Sporer, 2001). Possible reasons for the own-age bias are cognitive processes (He et al., 2011), personal motivation (Wiese et al., 2012) and (recent) contact frequency with out-group members (Rhodes & Anastasi, 2012; Wiese et al., 2013). Moreover, the own-age bias has been shown to be a robust effect across age groups from children to old age (Rhodes & Anastasi, 2012).

The present meta-analysis supported these assumptions, showing decreased differences between age groups with increasing face age, that is, a Participant Age by Face Age interaction (see Figure 4). Participant age differences decreased with increasing target age and became negative for older age faces for hits. Specifically, older participants showed an increase, whereas their younger counterparts showed a slight decrease in hits with older targets. Similarly, differences in false alarms and in recognition performance *d'* between the age groups decreased when participants were shown older faces compared to young faces. With mixed-age faces, young participants chose significantly fewer unknown faces as previously seen than older participants.

Practical Implications

The findings demonstrated that older individuals are not only less likely to correctly recognize a person they previously saw but also have a greater probability of incorrectly

---

[6] But see also Colloff, Wade, Wixted, and Maylor (2017), who used an online study in which the target had a highly distinctive mark [scar or tattoo on his cheek] and in order to create fair lineups in three out of four conditionns the lineups were digitally manipulated to prevent the distinctive target from standing out, and found older adults did not just set an overall more lenient response criterion than young adults. With unfair lineups, older adults were highly overconfident and confidence was not a good predictor of accuracy.

indicating a face as previously seen compared to younger people. This effect was robust across the different measures and was stronger for false alarms than for hits, with little heterogeneity. Although the studies included in this meta-analysis all used old-age groups with people older than 60, declines in recognition performance may occur even earlier (e.g., Germine et al., 2011).

These findings have general and forensic implications. In daily life, social interactions with new acquaintances (as opposed to people we know) might be complicated with increased age by the fact that memory for faces is impaired, and even more when the interaction is with younger people. Of course, this meta-analysis did not test whether these effects hold in live interactions (including conversations) but the present findings indicate that in older age people tend to "recognize" faces of strangers whom they think they have encountered before. As some of the individual studies reviewed show, these effects may be even stronger when the persons are encountered in a different context (or have changed in appearance—although our moderator analyses did not confirm this). Training programs for older adults "to improve your memory" could be designed to counteract age-related deficits, emphasizing the need for "individuation" of faces.

In criminal investigations, the face recognition paradigm reviewed here resembles more a "showup" or "street identification" than a traditional lineup with a suspect and several distractors. Considering the importance of person identifications in criminal investigations, our results should be considered by triers of fact (including police, attorneys, jurors and judges) who evaluate the accuracy of witness identification decisions. When evaluating identification decisions, these evaluators have to consider all possible factors that may affect the accuracy of an identification, especially if witnesses are older. The sensory system, eyesight and cognitive functioning are important, and all these factors should be evaluated individually. For instance, not every young person would be a more reliable observer than an older person when his or her visual ability is limited. Likewise, not every older person would be an unreliable witness just because they have reached a particular age. If possible, medical and psychiatric history should be taken into consideration, as these factors impair memory and recognition.

An important factor is a participant's age in relation to the age of the person-to-be-recognized. The own-age bias has been repeatedly demonstrated, both in the present meta-analysis on studies with the face recognition paradigm and in those using eyewitness identification paradigms (although evidence in the latter studies is not so clear; see Kocab et al., 2017; Rhodes & Anastasi, 2012). Accordingly, older individuals should be more reliable in recognizing older faces than younger faces, and young individuals should be less reliable in recognizing faces as the target face increases in age. As most crimes are committed by younger people, and the proportion of older people in the population, as well as their likelihood of victimization, increases, a possible own-age bias may be an important factor to consider.

Further, older individuals displayed a more liberal response criterion than younger people. Although attitudinal variables may not be related to identification decisions (at least not in the context of own-race bias research: Meissner & Brigham, 2001) older participants have been shown to have more positive attitudes towards the police (Gallagher, Maguire, Mastrofski, & Reisig, 2001) and to be more prosecution prone in a large scale representative study using the *Juror Bias Scale* (Goodman-Delahunty, 2018). Hence we hypothesize that older people may also be more likely to choose someone in an identification/face recognition task even if their memory trace is weak. There is at least some prior research that indicated that personality and age were directly related to setting of response criteria (Huh et al., 2006; Kantner & Lindsay, 2013); for a general review of individual differences and facial recognition, see Hosch, 1994).

However, we have been unable to locate research that directly compared young vs. older people on attitudinal variables like *Belief in a Just World* (Jones & Brimbal, 2017; Lipkus, 1991), *Right-wing Authoritarianism* (Altemeyer, 1998; Zakrisson, 2005), or *Legal Authoritarianism* (Narby, Cutler, & Moran, 1993) in relation to the setting of a more liberal response criterion. While most researchers in the past have favored cognitive explanations for age effects in recognition performance we encourage future researchers to also investigate social and personality variables to test whether they may explain the setting of response criteria in face recognition and lineup studies.

In identification or showup procedures, possible countermeasures against a response bias (e.g., fair lineup instructions: Memon, Hope, Bartlett, & Bull, 2002) should be further empirically tested and optimized for different age groups (children, younger and older adults). Also, explicitly allowing a “don’t know” option could be explored.

Limitations of the Present Meta-analysis

One of the most frequently voiced objections to meta-analysis is the “file-drawer problem” (Rosenthal, 1979), that is, the absence of unpublished data. If unpublished studies are missing, the overall effect size is likely to be overestimated, based on the assumption that unpublished studies are likely to have been negative or not significant.

To counteract the overestimation of effect sizes different methods were applied in the current meta-analysis. First, known researchers in the field and authors of identified primary studies were contacted to obtain unpublished data. Second, the literature search included not only English but also German, French and Russian studies to overcome a language bias. Unfortunately, the search for unpublished and non-English studies did not add any additional studies to the current meta-analysis. Even though some authors responded, these studies did not meet the inclusion criteria. In addition, missing data in some of the published studies led to exclusion of these data even though we had contacted the researchers of the primary studies for clarification. Third, the nonparametric trim-and-fill analysis (Duval & Tweedie, 2000a, 2000b) was conducted to assess the possible impact of publication bias and to calculate an “adjusted” effect size. Due to the small number of studies that reported the response criterion the impact of publication bias using the trim-and-fill analysis (Duval & Tweedie, 2000a, 2000b) could not be assessed for this dependent measure.

The trim-and-fill analysis (Duval & Tweedie, 2000a, 2000b) was conducted when combining young and mixed-age faces for hits, false alarms, and recognition performance, respectively. Results indicated the potential absence of four smaller or negative effect sizes for false alarms, and the potential absence of three smaller effect sizes for recognition performance; thus a possible overestimate of the weighted mean effect size may have been present for these measures. The difference between the results obtained in the meta-analysis and the "adjusted" effect sizes was negligible. There was no indication of a publication bias for hits using this method.

A further limitation of the meta-analysis associated with missing data was nonreporting of inter-effect-size covariances between dependent measures in a study. This made it impossible to include all possible effects in the meta-analysis without violating the assumption of independence of effect sizes (Cheung & Chan, 2014; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). The random variance estimation method was not conducted, because (1) the method assumes equal correlations between measures—but our data indicated that this was not the case; and (2) at least 40 studies are required to yield stable results (Hedges et al., 2010). Multi-level meta-analyses that take within-cluster dependencies into account could not be applied, because some of the studies reported only data for one effect size, and therefore did not meet the criteria for Level 2 clustering (for a critical review, see Cheung & Chan, 2014). The sample-wise approach would lead to biased results due to the assumed heterogeneity of the effect sizes when all measures would have been entered into the same model (Cheung & Chan, 2004; Hunter & Schmidt, 2004).

Finally, our meta-regression findings need to be interpreted with caution for different reasons. While the meta-regression results at step 1 seem to be robust, the absence of significant effects in the full model are likely a function of low power (Hedges & Pigott, 2004) and the fact that some of the intercorrelated predictor variables share mutual variance. To gain further insight into these interdependencies we calculated the semi-partial correlations between predictor variables and hits and false alarms.[7] To increase the validity of our meta-regression analyses (a) we excluded outliers from analyses to prevent distortions of effects; (b) we specified predictors based on theory *a priori* and not based on "*ex post facto* observed correlations" (Schmidt, 2017, p. 470); (c) we reported unstandardized and standardized regression weights, and (d) intercorrelations and semi-partial correlations of predictor variables.

---

[7] These analyses were recommended by an anonymous reviewer.

Methodological Implications and Future Directions

As discussed above, one of the limitations of the meta-analysis was the presence of missing values. This was particularly present for the signal detection measures for performance (*d'*) and response bias (*c*). Less than two-thirds of the studies reported a discrimination index and only a handful of studies reported data on response criterion. Hence, we encourage future researchers not only to report the individual face recognition measures but also the combined statistics showing participants' SDT performance and response criterion measures, as these measures can only be calculated with original data.

Although many studies tested face recognition for faces changed between study and test (e.g., pose, expression), such natural variations should be routinely employed (even in neurological studies). To control for possible confounds of distinctiveness, attractiveness, familiarity, perceived age and other face attributes that covary with face age, large pools of faces should be used. Faces could be pre-rated on these attributes and assigned either randomly or counter-balanced in complex experimental designs. One of the advantages of the face recognition paradigm over eyewitness identification paradigms is the maximization of stimulus sampling (Wells & Windschitl, 1999).

Furthermore, future studies should manipulate and extend the delay between encoding and retrieval; more than half of the studies tested recognition performance immediately, most studies within 25 minutes, and only one study after 48 hours (Anastasi & Rhodes, 2006, Exp. 1). Hence, the implications of the present findings are limited to an immediate or short delay and memory and retrieval processes after longer delays remain unknown. Over extended delays, the role of verbal and visual (rehearsal) processes and of context reinstatement could also be tested with young and older adults.

Finally, future research should include confidence as an additional measure to the "Old"/"New" recognition paradigm or as response alternatives in rating experiments (i.e., high to low certainty that the stimulus was "old"/"new"; Macmillan & Creelman, 2005). Such designs will enable to plot empirical Receiver Operating Curves, which are more fine-grained measures of sensitivity. Confidence has been routinely measured in eyewitness research, and findings revealed that high confidence of choosers is a reliable indicator of identification accuracy (Sporer, Penrod, Read, & Cutler, 1995; Wixted & Wells, 2017). Confidence was also more strongly related to accuracy with younger than with older adults in an identification field study (Martschuk, Sporer, & Sauerland, 2018; but see Colloff et al., 2017, who found older adults could adjust their confidence ratings appropriately to achieve a similar likelihood of being correct at each level of confidence as younger adults in fair but not in unfair lineups). Future studies need to address the confidence-accuracy issue with different targets (e.g., race), procedures, retention intervals and other possible moderators related to old age.

Conclusions

The present meta-analysis demonstrated that face recognition performance was weaker for older than for young adults. Further, older adults tended to choose more often, demonstrating a more liberal response bias which led to more false alarms, whereas young adults tended to use a more conservative response criterion. This is in line with the meta-analytical findings for eyewitness identification studies (Kocab et al., 2017). This conclusion was also supported when hits and false alarms were plotted against each other for each age group separately: While the mirror effect was present for young participants, showing a negative association between hits and false alarms, there was no visible association between hits and false alarms for older people, perhaps even a positive relationship.

Further, our findings showed the importance of face age. In line with previous research (e.g., Meissner & Brigham, 2001; Rhodes & Anastasi, 2012) the present study demonstrated an own-group advantage which was strongest for hits (i.e., older adults made more hits than their younger counterparts when recognizing old faces), and to a lesser extent for recognition performance *d'* (i.e., the age differences were smaller for mixed-age and older faces compared to young faces). Finally, longer retention intervals (as opposed to immediate retrieval) tended to be associated with greater differences between the age groups but future research should test this with longer delay intervals.

References

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

Altemeyer, B. (1998). The other "authoritarian personality". *Advances in Experimental Social Psychology, 30,* 47–92. doi:10.1016/S0065-2601(08)60382-2

*Anastasi, J. S., & Rhodes, M. G. (2006). Evidence for an own-age bias in face recognition. *North American Journal of Psychology, 8,* 237–252. doi:10.1037/2Fe527342012-645

Anstey, K. J. (2011). Biomarkers and memory aging: A life-course perspective. In M. Naveh-Benjamin & N. Ohta (eds.), *Memory and aging: Current issues and future directions* (pp. 349–372). New York: Taylor and Francis Group.

Arditi, A. (2005). Improving the design of the letter contrast sensitivity test. *Investigative Ophthalmology and Visual Science, 46,* 2225–2229. doi:10.1167/iovs.04-1198

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face images. *Journal of Experimental Psychology: General, 142,* 1323–1334. doi:10.1037/a0033872

Bartlett, J. C. (2014a). The older eyewitness. In T. Perfect & D. Lindsay (Eds.), *The SAGE handbook of applied memory* (pp. 654–674). London, UK: SAGE Publications Ltd.

Bartlett, J. C. (2014b). True and false recognition of faces by young and old adults. In M. Toglia, D. Ross, J. Pozzulo, & E. Pica (Eds.), *The elderly witness in court* (pp. 67–92). New York: Psychology Press.

*Bartlett, J. C., & Fulton, A. (1991). Familiarity and recognition of faces in old age. *Memory & Cognition, 19,* 229–238. doi:10.3758/BF03211147

*Bartlett, J. C., Leslie, J. E., Tubbs, A., & Fulton, A. (1989). Aging and memory for pictures of faces. *Psychology and Aging, 4*, 276-283. doi:10.1037/0882-7974.4.3.276

Bartlett, J. C., & Memon, A. (2007). Eyewitness memory in young and older adults. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Vol. 2: Memory for people* (pp. 309–338). Mahwah, NJ: Lawrence Erlbaum.

*Bartlett, J. C., Strater, L., & Fulton, A. (1991). False recency and false fame of faces in young adulthood and old age. *Memory & Cognition, 19,* 177–188. doi:10.3758/BF03197115

Bastin, C. (2008). [Face recognition in young and older adults]. Unpublished raw data.

*Bastin, C., & Van der Linden, M. (2003). The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology, 17,* 14–24. doi:10.1037/0894-4105.17.1.14

Bastin, C., & Van der Linden, M. (2006). The effects of aging on the recognition of different types of associations. *Experimental Aging Research, 32*, 61–77. doi:10.1080/03610730500326291

Beck, A. T., Rush, A. J., Shaw, B., & Emery, G. (1979). *Cognitive therapy of depression.* New York: Guilford.

Becker, B. J. (2005). The failsafe *N* or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 111–126). West Sussex, UK: Wiley.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (Vol. 2, pp. 221–236). New York: Russell Sage Foundation.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8,* 5–18. doi:10.1002/jrsm.1230

Bower, G. H. & Karlin, M. B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology, 103,* 751–757. doi:10.1037/h0037190

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73,* 105–116. doi:10.1111/j.2044-8295.1982

*Bryce, M. S., & Dodson, C. S. (2013). Cross-age effect in recognition performance and memory monitoring for faces. *Psychology & Aging, 28,* 87–98. doi: 10.1037/a0030213

Burgess, P. W., & Shallice, T. (1996). Response suppression, initiation and strategy use following frontal lobe lesions. *Neuropsychologia, 34,* 263–273. doi:10.1016/0028-3932(95)00104-2

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology, 8,* 1467–1485. doi:10.1080/17470218.2013.800125

Cheung, S. F., & Chan, D. K.-S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology, 89,* 780–791. doi:10.1037/0021-9010.89.5.780

Cheung, S. F., & Chan, D. K.-S. (2014). Meta-analyzing dependent correlations: An SPSS macro and an R script. *Behavioral Research, 46,* 331–345. doi:10.3758/s13428-013-0386

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Coleman, P. G., & O'Hanlon, A. (2008). Ageing and adaptation. In R. Woods, & L. Clare (Eds.), *Clinical psychology of ageing* (pp. 17–32). West Sussex, England: John Wiley & Sons.

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging, 32,* 243–258. doi:10.1037/pag0000168

Craik, F. I. M. (1986). A functional account of age differences in memory. In F. Klix, & H. Hagendorf (Eds.), *Human memory and cognitive capabilities: Mechanisms and performances* (pp. 409–422). Amsterdam: Elsevier Science Publishers, North-Holland.

Craik, F. I. M., & Rose, N. S. (2012). Memory encoding and aging: A neurocognitive perspective. *Neuroscience and Behavioral Reviews, 36,* 1729–1739. doi:10.1016/j.neubiorev.2011.11.007

*D'Argembeau, A., & van der Linden, M. (2004). Identity but not expression memory for unfamiliar faces is affected by ageing. *Memory, 12,* 644–654. doi:10.1080/09658210344000198

Deffenbacher, K. A., Bornstein, B. H., McGorty, E. K., & Penrod, S. D. (2008). Forgetting the once-seen face: Estimating the strength of an eyewitness's memory representation. *Journal of Experimental Psychology: Applied, 14,* 130–150. doi:10.1037/1076-898X.14.2.139

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

Ebner, N. C., & Johnson, M. K. (2009). Young and older emotional faces: Are there age group differences in expression identification and memory? *Emotion, 9,* 329–339. doi:10.1037/a0015179

Egger, M., Smith, G. D., 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

Faculty of Ophthalmologists, London. (1987). *Reading test types.* London: Society of Ophthalmologists.

*Firestone, A., Turk-Browne, N. B., & Ryan, J. D. (2007). Age-related deficits in face recognition are related to underlying changes in scanning behavior. *Aging, Neuropsychology, and Cognition, 14,* 594–607. doi:10.1080/13825580600899717

Flicker, C., Ferris, C. H., Crook, T., & Bartus, R. T. (1990). Age differences in the vulnerability of facial recognition memory to proactive interference. *Experimental Aging Research, 15,* 189–194.

*Fulton, A., & Bartlett, J. C. (1991). Young and old faces in young and old heads: The factor of age in face recognition. *Psychology and Aging, 6,* 623–630. doi:10.1037//0882-7974.6.4.623

Gallagher, C., Maguire, E. R., Mastrofski, S. D., & Reisig, M. D. (2001). *The public image of the police.* Final report to The International Association of Chiefs of Police by The Administration of Justice Program, George Mason University. Retrieved, 24 September 2015, from http://www.theiacp.org/PoliceServices/ProfessionalAssistance/ThePublicImageofthePolice/tabid/198/Default.aspx#ch2

Germine, L., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118,* 201–210. doi:10.1016/j.*cognition*.2010.11.002

Gilbert, J. G., Levee, R. F., & Catalano, F. L. (1974). *Guild memory test manual.* Newark, NJ: UNICO National Mental Health Research Foundation.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13,* 8–20. doi:10.3758/BF03198438

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York: Russell Sage Foundation.

Goodman-Delahunty, J. (2018). *Juror bias, demographic correlates, and verdict: Findings from a large Australian sample.* Unpublished manuscript. Charles Sturt University, Manly, Australia.

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4,* 330–337. doi:10.1016/S1364-6613(00)01519-9

Harrison, V., & Hole, G. J. (2009). Evidence for a contact-based explanation of the own-age bias in face recognition. *Psychonomic Bulletin & Review, 16,* 264–269. doi:10.3758/PBR.16.2.264

*He, Y., Ebner, N. C., & Johnson, M. K. (2011). What predicts the own-age bias in face recognition memory? *Social Cognition, 29,* 97–109. doi:10.1521/soco.2011.29.1.97

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445. doi:0.1037/1082-989X.9.4.426

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1,* 39–65. doi:10.1002/jrsm.5

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition, 21,* 1306–1336. doi:10.1080/13506285.2013.823140

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558. doi:10.1002/sim.1186

Hildebrandt, A., Wilhelm, O., Schmiedek, F., Herzmann, G., & Sommer, W. (2011). On the specificity of face cognition compared with general cognitive functioning across adult age. *Psychology and Aging*, *26*, 701–715. doi:10.1037/a0023056

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21,* 302–313. doi:10.1037/0278-7393.21.2.302

Hosch, H. (1994). Individual differences in personality and eyewitness identification. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 328–347). New York: Cambridge University Press.

Hugenberg, K., Miller, J., & Claypool, H. (2007). Categorization and individuation in the cross-race recognition deficit: Toward a solution to an insidious problem. *Journal of Experimental Social Psychology, 43,* 334–340. doi:10.1016/j.jesp.2006.02

Huh, T. J., Kramer, J. H., Gazzaley, A., & Delis, D. C. (2006). Response bias and aging on a recognition memory task. *Journal of the International Neuropsychological Society, 12,* 1–7. doi:10.1017/S1355617706060024

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd Ed.). Thousand Oaks, CA: Sage.

Johnston, R. A., & Edwards, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17,* 577–596. doi:10.1080/09658210902976969

Jones, A. M., & Brimbal, L. (2017). Lay perceptions of interrogation techniques: Identifying the role of Belief in a Just World and Right Wing Authoritarianism. *Journal of Investigative Psychology and Offender Profiling, 14,* 260–280. doi:10.1002/jip.1476

Kantner, J. & Lindsay, D. S. (2013). Response bias in recognition memory as a cognitive trait. *Memory & Cognition, 40,* 1163–1177. doi:10.3758/s13421–012–0226–0

Kocab, K., Martschuk, N., & Sporer, S. L. (2017). *Identification evidence by elderly eyewitnesses: A meta-analysis.* Unpublished manuscript, University of Giessen, Giessen, Germany.

*Lamont, A. C., Steward-Williams, S. & Podd, J. (2005) Face recognition and aging: Effects of target age and memory load. *Memory & Cognition, 33,* 1017–1024. doi:10.3758/BF03193209

Lipkus, I. (1991). The construction and preliminary validation of a global belief in a just world scale and the exploratory analysis of the multidimensional belief in a just world scale. *Personality and Individual Differences, 12,* 1171–1178. doi:10.1016/0191-8869(91)90081-L

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Lott, L. A., Haegerstrom-Portnoy, G., Schneck, M. E., & Brabyn, J. A. (2005). Face recognition in the elderly. *Optometry and Vision Science, 82,* 874–881. doi:10.1097/01.opx.0000180764.68737.91

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Malpass, R. S. (1990). An excursion into utilitarian analysis. *Behavior Science Research, 24,* 1–15. doi:10.1177/106939719002400101

Martschuk, N., Sporer, S. L., & Sauerland, M. (2018). *Confidence of older eyewitnesses: Is it diagnostic of identification accuracy?* Manuscript submitted for publication.

*Mason, S. E. (1986). Age and gender as factors in facial recognition and identification. *Experimental Aging Research, 12,* 151–154. doi:10.1080/03610738608259453

Maylor, E. A. (2005). Age-related changes in memory. In M. L. Johnson (Ed.), *The Cambridge handbook of age and ageing* (pp. 200–208). Cambridge UK: Cambridge University Press.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 734–760. doi:10.1037/0033-295X.105.4.734-760

Megreya, A. B. & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69,* 1175–1184. doi:10.3758/BF03193954

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in face memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7,* 1–35. doi:10.1037//1076-8971.7.1.3

Memon, A., Hope, L., Bartlett, J., & Bull, R. (2002). Eyewitness recognition errors: the effects of mugshot viewing and choosing in young and old adults. *Memory and Cognition, 30,* 1219–1227. doi:10.3758/BF03213404

Narby, D. J., Cutler, B. L., & Moran, G. (1993). A meta-analysis of the association between authoritarianism and jurors' perceptions of defendant culpability. *Journal of Applied Psychology, 78,* 34–42. doi:10.1037/0021-9010.78.1.34

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1179–1187. doi: 10.1037//0278-7393.26.5.1170

Naveh-Benjamin, M., Guez, J., Kilb, A., & Reedy, S. (2004). The associative memory deficit of older adults: Further support using face-name associations. *Psychology and Aging, 19,* 541–546. doi:10.1037/0882-7974.19.3.541

*Naveh-Benjamin, M., Shing, Y. L., Kilb, A., Werkle-Bergner, M., Lindenberger, U., & Li, S-C. (2009). Adult age differences in memory for name-face associations: The effects of intentional and incidental learning. *Memory, 17,* 220–232. doi:10.1080/09658210802222183

Old, S. R., & Naveh-Benjamin, M. (2008). Memory for people and their actions: Further evidence for an age-related associative deficit. *Psychology & Aging, 23,* 467–472. doi:10.1037/0882-7974.23.2.467

*Oloffson, M., & Bäckman, L. (1996). Influences of intentionality at encoding and retrieval on memory in adulthood and old age. *Aging Clinical and Experimental Research, 8,* 42–46. doi:10.1007/BF03340114

Orwin, R. G. (1983). A fail-safe *N* for effect size. *Journal of Educational Statistics, 8,* 157–159. doi:10.2307/1164923

O'Toole, A. J., Bartlett, J. C., & Abdi, H. (2000). A signal detection model applied to the stimulus: Understanding covariances in face recognition experiments in the context of face sampling distributions. *Visual Cognition, 7,* 436–463. doi:10.1080/135062800394603

Pigott, T. E. (2012). *Advances in meta-analysis.* New York: Springer.

Quattrone, G. A. & Jones, E. E. (1980). The perception of variability within ingroups and outgroups: Implications for the law of small numbers. *Journal of Personality and Social Psychology, 38,* 141–152.

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin, 138,* 146–174. doi:10.1037/a0025750

Rodin, M. J. (1987). Who is memorable to whom: A study of cognitive disregard. *Social Cognition, 5,* 144–165. doi:10.1521/soco.1987.5.2.144

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–641. doi:10.1037/0033-2909.86.3.638

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103,* 734–760. doi:10.1037/0033-295X.103.4.734

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103,* 403–428. doi:10.1037/0033-295X.103.3.403

*Savaskan, E., Müller, S. E., Böhringer, A., Philippsen, C., Müller-Spahn, F., & Schächinger, H. (2007). Age determines memory for face identity and expression. *Psychogeriatrics, 7,* 49–57. doi:10.1111/j.1479-8301.2007.00179.x

Schieber, F. (2006). Vision and aging. In J. E. Birren, & K. W. Schaie (Eds.), *The handbook of the psychology of aging* (pp. 129–161). Burlington, USA: Elvesier Academic Press.

Schmidt, F. L. (2017). Statistical and measurement pitfalls in the use of meta-regression in meta-analysis. *Career Development International, 22,* 469–476. doi:10.1108/CDI-08-2017-0136

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (Vol. 2, pp. 257–277). New York: Russell Sage Foundation.

Shapiro, P. N., & Penrod, S. (1986). A meta-analysis of facial identification studies. *Psychological Bulletin, 100,* 139–156. doi:10.1037/0033-2909.100.2.139

Short L. A., Semplonius T., Proietti V., & Mondloch C. J. (2014). Differential attentional allocation and subsequent recognition for young and older adult faces. *Visual Cognition, 22,* 1272–1295. doi:10.1080/13506285.2014.993007

*Smith, A. D., & Winograd, E. (1978). Adult age differences in remembering faces. *Developmental Psychology, 14,* 443–444. doi:10.1037//0012-1649.14.4.443

Sporer, S. L. (1991). Encoding strategies and the recognition of human faces. *Journal of Experimental Psychology: Human Learning, Memory, and Cognition, 17,* 323–333. doi:10.1037/0278-7393.17.2.323

Sporer, S. L. (1992). *Das Wiedererkennen von Gesichtern [Recognizing faces].* Weinheim: Beltz/Psychologie Verlags Union.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7,* 36–97. doi:10.1037//1076-8971.7.1.36

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.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75–98). West Sussex, England: John Wiley & Sons Ltd.

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:0895-4356/01/

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 6,* 661. doi:10.1037/h0054651

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (Vol. 2, pp. 435–452). New York: Russell Sage Foundation.

The Innocence Project. (2017, December 15). Help us put an end to wrongful convictions! Retrieved from http://www.innocenceproject.org/

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race on face recognition. *Quarterly Journal of Experimental Psychology, 43A,* 161–204. doi:10.1080/14640749108400966

Valentine, T., Chiroro, P., & Dixon, R. (1995). An account of the own-race bias and the contact hypothesis based on a ‘face space’ model of face recognition. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition: Explorations in face space* (pp. 64–94). London: Routledge.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavioral Research, 45,* 576–594. doi:10.3758/s13428-012-0261-6

Viechtbauer, W. (2010). Metafor: Meta-analysis package for R (Version 1.9-2). [Software]. Retrieved from http://www.metafor-project.org/

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition, 20,* 291–302.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale.* New York: Psychological Corporation.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25,* 1115–1125. doi:10.1177/01461672992512005

Wiese, H., Komes, J., & Schweinberger, S. R. (2012). Daily-life contact affects the own-age bias and neural correlates of face memory in elderly participants. *Neuropsychologia, 50,* 3496–3508. doi:10.1016/j.neuropsychologia.2012.09.022

Wiese, H., Wolff, N., Steffens, M. C., & Schweinberger, S. R. (2013). How experience shapes memory for faces: An event-related potential study on the own-age bias. *Biological Psychology, 94,* 369–379. doi:10.1016/j.biopsycho.2013.07.001

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

Wolff, N., Wiese, H., & Schweinberger, S. R. (2012). Face recognition memory across the adult life span: Event-related potential evidence from the own-age bias. *Psychology and Aging, 27,* 1066–1081. doi:10.1037/a0029112

Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta Psychologica, 114,* 101–114. doi:10.1016/S0001-6918(03)00052-0

Wright, D. B., & Stroud, J. S. (2002). Age differences in lineup identification accuracy: People are better with their own age. *Law and Human Behavior, 26,* 641–654. doi:10.1023/A:1020981501383

Zakrisson, I. (2005). Construction of a short version of the right-wing authoritarianism (RWA) scale. *Personality and Individual Differences, 39,* 863–872. doi:10.1016/j.paid.2005.02.026

CONFIDENCE OF OLDER EYEWITNESSES: IS IT DIAGNOSTIC OF IDENTIFICATION ACCURACY?

Eyewitness identification is an important determinant of trial outcomes, and mistaken identifications have frequently contributed to miscarriages of justice (Garrett, 2011). This is particularly true for in-court identifications, which are often presented with inflated confidence. Early reviews and meta-analyses of the confidence–accuracy relationship concluded that confidence only modestly predicts the accuracy of eyewitnesses identification (Bothwell, Deffenbacher, & Brigham, 1987; Wells & Murray, 1983). Consequently, many researchers advised against using expressed confidence as an indicator of accuracy (e.g., Penrod & Cutler, 1995). By contrast, when focusing on choosers of the suspect – only those who usually end up testifying in court against the defendant – the meta-analysis by Sporer, Penrod, Read, and Cutler (1995) revealed a *strong* confidence–accuracy correlation (r_{pb} weighted = .37; r_{pb} unweighted = .41).[8] However, Sporer et al. (1995) also noted a large heterogeneity across studies, ranging between r_{pb} = .10 and r_{pb} = .74. Subsequent studies found even higher confidence–accuracy correlations (e.g., Lindsay, Nilsen, & Read, 2000; Read, 1995). Therefore, it seems worth exploring under which conditions these relationships may be particularly high or not so high.

Recently, in several re-analyses of previous research, Wixted and colleagues concluded that confidence – if measured and recorded at the time of an initial eyewitness identification – was a reliable marker of identification accuracy (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted, Read, & Lindsay, 2016). However, this conclusion may only hold under "pristine" conditions (i.e., fair lineup instructions, fair lineups, etc.: Wixted & Wells, 2017). As noted by Palmer, Brewer, Weber, and Nagesh (2013), "it remains an empirical question as to whether there are conditions under which the confidence–accuracy relationship

---

[8] Note that point-biserial correlations of .10, .24, and .37, are equivalent to standardized mean difference *d* values of .20 ("small"), .50 ("medium"), and .80 ("large") effects (Cohen, 1988).

does collapse entirely" (p. 66). Here we draw attention to the possibility that even under "pristine" conditions moderators of the confidence–accuracy relationship do exist and may affect accuracy independently of confidence, thus yielding a dissociation that reduces the probative value of confidence. In particular, we focus on the age of eyewitnesses – from juveniles to the aged – as a potential moderator. While there is a robust effect of age on recognition performance (see the meta-analyses by Fitzgerald & Price, 2015; Kocab, Martschuk, & Sporer, 2018; Martschuk & Sporer, 2018; see also the review by Bartlett & Memon, 2007), much less is known about *confidence* of older witnesses, and in particular the confidence–accuracy relationship.

Metamemory Models of the Confidence–Accuracy Relationship

According to the direct-access view of metacognitive judgments, reported confidence can be understood as an indicator of the strength of a memory trace: The stronger the trace, the more likely a piece of information will be recalled or recognized. Followingly, in an identification decision confidence is expected to be higher when the witness experiences a strong sense of familiarity that the suspect in the lineup matches her or his image of the perpetrator. Signal detection theory models applied to eyewitness identification decisions still rely on this assumption but acknowledge that other types of information (e.g., feedback provided by a police officer) may also affect confidence (for a recent review, see Wixted & Wells, 2017).

An alternative meta-memorial explanation prevalent in the basic literature on memory is *cue utilization theory* (Koriat, 1997), which suggests that people use different types of cues to assign confidence judgments through analytic heuristics. However, metamemory studies have usually been conducted with paradigms that are quite different from those employed in eyewitness identification studies. Many studies have used word list or paired associate paradigms whereby words are either known or have varying degrees of associations with the stimuli-to-be-recalled or recognized. Somewhat closer to eyewitness studies are name-learning paradigms, which demonstrated monitoring accuracy for nouns as targets to be higher than for faces or names (Watier & Collin, 2011). In contrast, in eyewitness identification studies, the target is an unfamiliar person/face that has to be recognized among a list of distractors (fillers) that are usually preselected on the basis of some similarity to the target (as in the study presented here).

In studies on ease of learning, judgments of learning, or feeling of knowing, participants are asked **before** a memory test if particular items will be recalled or recognized (for review, see Dunlosky & Tauber, 2014). In eyewitness identification studies, this would correspond to the police asking a witness before an identification task, if he/she believes to be able to recognize the target in a lineup (referred to as pre-decision confidence vs post-decision confidence which is assessed after a lineup; e.g., Sporer, 1992). There is ample evidence that pre-decision confidence is unrelated to the accuracy of a later identification (see the meta-analysis by Cutler & Penrod, 1989). Hence, in more recent studies, researchers have no longer included pre-decision confidence as an assessment variable.

There is also a large body of research on metacognitive control and memory reporting (e.g. Goldsmith, Pansky, & Koriat, 2014; Koriat & Goldsmith, 1994, 1996), which has been fruitfully applied to eyewitness recall (e.g., Weber & Brewer, 2008). The major difference to identification studies is that in Koriat and colleagues' studies the focus is on quality vs quantity of recall, not on the accuracy of recognition (but see Perfect & Weber, 2012, for a noteworthy exception).

Thus, while we do not question the importance of these metamemory studies for eyewitness *recall*, we are reluctant to apply the underlying theories to face recognition or person identification paradigms. In person identification studies, there is (usually) only one unfamiliar target-to-be recognized which has been seen once, usually without instructions to commit the face to memory (i.e., an incidental learning paradigm).

When eyewitness researchers applied Koriat's (1997) cue utilization approach to an identification paradigm (Perfect & Weber, 2012), important differences emerged that further support our claim that metamemory theories developed in verbal learning paradigms may not yield comparable results. Whereas the order for "forced reports" and "free reports" did not matter in Koriat and Goldsmith's (1996, Experiment 2) study, for eyewitness identification decisions the order was crucial (Perfect & Weber, 2012). Also, in an identification study by Weber and Perfect (2012) free-report decisions were significantly more diagnostic than forced-report decisions, with no significant reduction in the quantity of correct identifications obtained. That is, there was a benefit to memory accuracy that was accompanied by a negligible cost to memory quantity, in an apparent contradiction of the memory control framework. These data indicate that a simple version of Koriat and Goldsmith's (1996) control framework may not apply to identification decisions when a "Don't know" option is provided. Unfortunately, most identification studies do not offer a "Don't know" option, thus, complicating the comparison between studies in the basic metamemory literature and the eyewitness domain even further (see also the recent discussion on the occurrence of positive, null, and negative confidence–accuracy relationships by Roediger & DeSoto, 2015).

In an attempt to test these rival views within an identification paradigm, Busey, Tunnicliff, Loftus, and Loftus (2000) provided evidence for both the direct access and cue utilization models. When they showed photographs with varying exposure time (Experiment 1) and different levels of illuminations (Experiment 2) the findings were consistent with the trace access memory provided. However, when illumination levels between the exposure and test were varied (Experiment 3), the findings were consistent with the cue utilization approach. Generally, Busey et al. (2000) showed a high confidence–accuracy relationship, unless the conditions were not optimal at the test. This is an example where confidence and accuracy rely on mainly the same information under some circumstances but may also yield a dissociation between confidence and accuracy under different conditions.

The Confidence–Accuracy Relationship across the Lifespan

The confidence–accuracy relationship has been well researched for over 40 years, yet to our knowledge only three studies investigated the issue over the lifespan in an eyewitness identification paradigm in a field setting (Palmer et al., 2013, age range 14–87; Sauer, Brewer, Zweck, & Weber, 2010, age range 15–85; Sauerland & Sporer, 2009, age range 15–84). However, in all three studies the analyses were conducted without separating the age groups. In Sauer et al. (2010) only 6% of the participants (*n* = 64) were over 60 years of age (Sauer, 2016, personal correspondence[9]). Palmer et al. (2013) conducted additional analyses

---

9 We greatly appreciate the authors for providing us with their age distribution data.

excluding participants younger than 18 years and older than 65 years. They found no differences in results, except for the influence of exposure time on resolution and of retention interval on calibration. When all participants were included, the resolution and calibration scores were significantly better following a longer exposure duration and a shorter retention interval. By contrast, when adolescents and older participants were excluded from the analyses, no effect of exposure time and retention interval was found. These analyses, however, did not specifically address age-related changes in the confidence–accuracy relationship. In Sauerland and Sporer (2009) more than 9% of the participants (10.8% of choosers) were over the age of 60.

Several laboratory studies that compared different age groups directly suggest age-related changes in the confidence–accuracy relationship: In a study with children, confidence was not diagnostic of identification accuracy (Keast, Brewer, & Wells, 2007), and in three studies the confidence of older adults was less diagnostic of accuracy than confidence of young adults (Searcy, Bartlett, & Memon, 1999; Searcy, Bartlett, Memon, & Swanson, 2001; Wright & Stroud, 2002). These results provide some evidence that the confidence–accuracy relationship varies by age, at least in laboratory settings.

A recently published online study compared eyewitness identifications of 890 young (18–30 years), 890 middle-aged (31–59) and 890 older adults (50–95 years) in an online experiment (Colloff, Wade, Wixted, & Maylor, 2017). Participants watched a 30-second video depicting a mock crime and were presented with a fair or unfair target-present or a target-absent lineup. Following their decision they rated their confidence. Analyses revealed an age-related decline in participants' ability to discriminate between guilty and innocent suspects, and to adjust their confidence accordingly using signal detection theory modeling. These results seem to suggest that the confidence–accuracy relationship is stable across the lifespan. However, in an online study the risk of a preselection bias is increased as participants are required to access and use a PC. Therefore, it is possible that a higher proportion of older people as opposed to young people might have not been able to participate (i.e., those who are not fit enough or confident to use a computer). By contrast, participants in the present study were citizens who were around a downtown shopping area during the day, and did not need to have access to a PC (as in Palmer, David, & Fleming, 2014, and Colloff et al., 2017).

Further, the major focus of the Colloff et al. study was on how best to construct fair lineups by modifying all lineup faces when the suspect has a striking mark on his face (e.g. a scar or a large bruising around the eye), using one single male target. As is evident from the literature on face recognition, adding lines to a face or blocking parts of a face disrupts holistic processing (e.g., Ellis, Davies, & Shepherd, 1978; Hancock, Bruce, & Burton, 2000; Richler, Cheung, & Gauthier, 2011; Richler, Mack, Gauthier, & Palmeri, 2009). It is an open question whether the results obtained by Colloff et al. (2017) regarding age effects would replicate with targets without any highly distinguishing marks and without editing lineup fillers to match this mark.

Taken together, there seems to be ample evidence that identification accuracy may decline at old age but we still do not know enough whether there is a parallel decline in the confidence–accuracy relationship. Here we present a re-analysis of the confidence–accuracy relationship from adolescence to old age using an incidental encoding task in a field experiment, originally published by Sauerland and Sporer (2009) without any consideration of participants' age. While the original study investigated identification decision processes independently of age separately for choosers and non-choosers, the present study focused on the influence of age on the confidence–accuracy relationship in choosers only. We refer to results from the earlier study only to the extent that they are contrasted with the present analysis.

Metamemorial Deficits of Older Eyewitnesses

Metamemory, defined as "knowledge and awareness about one's own memory and how memory works more generally" (Castel, McGillivray, & Friedman, 2012, p. 246), refers to both prospective and retrospective metacognitive judgments. The present research focused on retrospective cognitive judgments, namely the expressed confidence that an identification judgment was correct. Two reciprocal processes of metamemory have been suggested: (1) metacognitive *monitoring* and (2) metacognitive *control* (Nelson & Narens, 1990; Price, Mueller, Wetmore, & Neuschatz, 2014). Metacognitive monitoring evaluates the current state of retrieval and provides information about the learning status. Thus, monitoring consists of subjective perceptions about own introspections or mnemonic strength. Metacognitive control is a type of behavior that is used to optimize performance at a meta-level without direct access to the current state, representing an analytic factor of metacognitive judgments. Both monitoring and control are important for metacognitive judgments, and dissociations between behavior and beliefs may become more marked across the lifespan (Palmer et al., 2014).

Previous research demonstrated that while monitoring of well-learned information remained unaffected by aging, metacognitive monitoring of recently encountered information was more likely to be impaired with increasing age (Dodson, Bawa, & Krueger, 2007; Dodson & Krueger, 2006). For instance, Dodson et al. (2007) showed that older participants recognized statements as well as younger participants but made more source monitoring errors and displayed more high-confidence errors (Experiment 1). In Experiment 2, which used a cued-recall task with word lists also showed more high-confidence errors in older compared to younger adults. Perhaps more relevant to the eyewitness literature, older eyewitnesses also showed more high-confidence suggestibility errors than younger participants after watching a video (Dodson & Krueger, 2006). However, more recent studies using word lists containing related lures suggest that positive, null, and negative confidence–accuracy relationships may be observed as a function of the stimulus material and paradigm (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014). Applied to a lineup task, an innocent suspect strongly resembling the perpetrator may lead to high-confidence false identifications (Roediger & DeSoto, 2015, citing anecdotal case evidence from Buckhout, 1974). Whether this varies as a function of participant age remains an empirical question.

Metacognitive control – which is considered important in an identification task – is impaired in older age (Price et al., 2014). In other words, compared to young adults, older adults are less likely to effectively use available information when making decisions (Hertzog & Hultsch, 2000; Price et al., 2014) and have difficulties retrieving information (Shing, Werkle-Bergner, Li, & Lindenberger, 2009). Cross-sectional metacognitive studies showed that metacognitive efficiency improved during adolescence, reached a plateau during adulthood (age range 11–41 years; Weil et al., 2013), and declined in older adulthood (age range 18–84 years; Palmer et al., 2014). These studies together suggest that, over the lifespan, metacognitive efficiency resembles an inverted U: Adolescents and older adults have a stronger tendency towards overconfidence than young and middle-aged adults. However, most of this research was concerned with learning of word lists or semantic memory, not with face recognition or identification tasks. It has long been known in the memory literature that verbal and visual information, and faces in particular, are encoded and processed differently (e.g. Meissner, Sporer, & Schooler, 2007; Paivio, 1971; Tulving, 1985). Some of these differences may even be more pronounced in older adults (e.g. Jenkins, Myerson, Hale, & Fry, 1999). The present study extended these theoretical notions to the eyewitness identification task by analyzing the confidence–accuracy relationship as a function of age from adolescence to older adulthood.

Decision Time–Accuracy Relationship across the Lifespan

Theoretically, face recognition is considered a holistic process, which occurs rather automatically, even after very short exposures (Richler, Mack, Gauthier, & Palmeri, 2009). Self-reports of witnesses who claim that a face just "popped out" from a lineup and who are more likely to be correct than slow responders (e.g. Brewer & Weber, 2008; Sporer, 1992) support this notion. On the other hand, witnesses deliberating and engaging in a relative decision strategy spend more time on taking a decision and are more likely to be wrong (Sauerland & Sporer, 2007). Applied to lineup decisions of choosers, it can be predicted that decision times (response latencies) for hits are faster than for false alarms (Dunning & Stern, 1994; Sporer, 1994). By now, quite a few studies have shown that decision time – a more objective indicator than confidence – is associated with identification accuracy, that is, eyewitnesses made correct identifications faster than incorrect identifications (e.g. Brewer & Weber, 2008; Dunning & Perretta, 2002; Sporer, 1992, 1993, 1994). This is also in line with the dual-process model, which proposes that accurate decisions are made rapidly and automatically, and require no cognitive effort, while inaccurate decisions are time-consuming, conscious, and deliberate (Dunning & Perretta, 2002; Dunning & Stern, 1994). Similar tendencies have been shown for self-reported estimated decision times, namely that correct identifications were estimated by witnesses to be faster than false identifications (Sauerland & Sporer, 2009).

The question arises whether or not measured decision-times as well as self-reported decision times show the same stable relationship with identification accuracy across the lifespan. In other words, do older adults encode and retrieve faces differently and do they use different metacognitive strategies than younger adults? Sporer (2001) has proposed an in-group/out-group model which postulates that in-group faces, for example, of one's own ethnic group ("race"), gender, and age are processed relatively more holistically than faces of out-group members. Regarding age, a recent meta-analysis of face recognition studies comparing younger and older adults has found support for this prediction (Martschuk & Sporer, 2018). Considering that all studies reported above used only young targets in their research, the findings about decision times and their association with identification accuracy may be limited to young witnesses observing young targets.

In the light of the recent demographic developments which entail that more and more witnesses will be older it is an important question whether older witnesses observing young targets show the same pattern. To the extent that older adults do not engage in the type of holistic processing mentioned, or employ different metacognitive strategies than younger adults, this association may be weakened in older age. Some insight into lineup decision processes of different age groups were provided by Colloff et al. (2017, supplemental material) in their additional analyses. Particularly, older adults took longer to make a lineup decision (i.e., to identify or reject a lineup) and to render a confidence judgment than middle-aged and young adults. These analyses, however, were only conducted *across* choosers and non-choosers. It remains an open question whether these relationships will be stronger when focusing only on choosers who are particularly important in criminal cases.

Aims and Hypotheses

The aim of this study was to assess the confidence–accuracy relationship for different age groups across the lifespan. We predicted that the confidence–accuracy relationship deteriorates with increasing age. More specifically, we expected accuracy and confidence to dissociate across the lifespan, with older participants being relatively more overconfident, leading to poorer calibration (Hypothesis 1). To investigate the postdictive value of confidence as a function of age at different prior base rates of target-presence (Sauerland, Sagana, Sporer, & Wixted, 2018; Wells, Yang, & Smalarz, 2015; Wixted & Wells, 2017), we also conducted Bayesian analyses. While most lab studies employ a base rate of 50% target-present lineups and 50% target-absent lineups, presumably, base rates vary considerably across different legislations and police departments. Importantly, the postdictive value of confidence following a lineup decision varies as a factor of this prior base rate (Wixted & Wells, 2017), with higher target-presence base rates being associated with higher postdictive value compared to lower target-presence base rates (Hypothesis 2). In addition to replicating this finding, we expected the postdictive value of confidence to decrease as age increases. Finally, as proposed with the dual-process model of the decision process, we expected that correct identifications would be made faster and would be estimated to be faster than false identifications. To the extent that older participants rely less on holistic processing or differ in their metacognitive strategies, they may show lower decision time-accuracy associations than younger adults (Hypothesis 3). Considering the age-related decline of metacognitive monitoring and control, we assumed the self-reported decision time–accuracy associations to be similarly or even more impaired in older age (Hypothesis 4).

Method

Participants

This study comprised a re-analysis of data collected by Sauerland and Sporer (2009) with 720 citizens (50% female, 50% male) who agreed to participate in the experiment. Four hundred and thirty-six participants made a positive identification decision (called “choosers”), whereas 284 rejected the lineup.

In the present study only choosers (221 male, 215 female) were included, aged between 15 and 83 years (M = 37.63, SD = 15.71, Mdn = 34.5). Participants were categorized into six age groups: 15–20 years (n = 44); 21–30 (n = 149); 31–40 (n = 74); 41–50 (n = 58); 51–60 (n = 64); and 61–83 (n = 47). The proportions of choosers did not differ as a function of age group: 15–20 years: 60.3%; 21–30: 59.8%; 31–40: 59.2%; 41–50: 53.2%; 51–60: 66.0%; 61–83: 70.1%; $\chi^2(5, N = 720) = 5.39$, $p = .270$, Cramer's $V = .27$.

The majority of the participants were employees in the academic (16.3%) or non-academic field (38.6%), 27.5% were students, 7.6% were retired, 5.3% attended school, and 4.9% were unemployed.

Design

The study was a 2 (target-present vs absent) x 6 (age groups 1 to 6) between-participants design. Dependent measures were identification accuracy and post-decision confidence. We also report some results on decision times (in s) and self-reported decision times.

Procedure

Ten different targets aged between 20 and 37 years were used to fulfil the requirement of stimulus sampling (Wells & Windschitl, 1999). Targets recruited participants in pairs, while one had the role of the target and the other of the research confederate. Each target was instructed to approach male and female participants from different age groups, including old-age people, so targets would be distributed approximately equally across gender and age groups of participants. Chi-square analyses confirmed that targets were equally distributed among age groups (all participants: $\chi^2(5, N = 720) = 52.58$, $p = .204$, Cramer's $V = .270$; choosers: $\chi^2(5, n = 436) = 50.31$, $p = .271$, Cramer's $V = .34$).

Following a predetermined script, the target approached a pedestrian on the street and asked for directions to a certain location nearby. After the target walked away, a research confederate approached the pedestrian and explained the true purpose of the interaction (a face recognition study). Once participants agreed, they were randomly assigned to a target-present or a target-absent photo lineup. Unbiased lineup instructions were used, and the research confederate was blind to target presence. Immediately after the identification task participants provided their confidence on an 11-point rating scale (0% to 100%, at 10% intervals) and estimated the time they spent on making their decision on a paper-and-pencil answer sheet. The confederate inconspicuously stopped the time when the participant entered her or his response to the lineup task.

Lineups

Ten target-present and ten target-absent photo lineups were prepared for the study. Each lineup contained six frontal 9 x 13 cm photographs that were arranged in two rows of three pictures. The target or target replacement was always placed on Position 3. The lineups were constructed following a pilot study with 55 mock witnesses that determined which portrait photographs fit the general target description (Wells, Rydell, & Seelau, 1993). Those who received most selections were used as foils or replacements.

Analyses

The present study used different types of calibration analyses following suggestions to apply alternative analyses, which are more diagnostic of the confidence–accuracy relationship than point-biserial correlations (e.g., Juslin, Olsson, & Winman, 1996; Olsson, 2000; Weber & Brewer, 2003). Calibration curves plot confidence level against the proportion of witnesses who made correct decisions. However, as in the present study sample sizes for the separate age groups were too small for traditional calibration curves, we relied on graphic comparisons of mean accuracy and confidence levels using “confidence–accuracy characteristic” (CAC) analysis (Mickes, 2015), along with 95% confidence intervals, and on point-biserial correlations between confidence and accuracy. To assess curvi-linear relationships as a function of age, we calculated correlations for age and age-squared (age^2).

Following the categorizations used by Wixted and colleagues in their recent publications on low, medium and high confidence, we re-categorized the confidence values of the original 0% to 100% scale into low confidence (0–50%; n = 191), medium confidence (60–80%; n = 130), and high confidence (90–100%; n = 115).

Further statistical calibration analyses were included as follows. *Calibration* (*C*) is the extent to which an eyewitness’s confidence or certainty judgment (measured in %) corresponds with the probability that the decision is correct (from *0 = perfect calibration* to *1 = weak calibration*). *Over-/underconfidence* (*O/U*) describes participants’ tendency to be over- (positive score) or underconfident (negative score) in comparison to their accuracy. The *adjusted normalized resolution index* (*ANRI*) indicates the discrimination of confidence judgments (from *0 = no discrimination* to *1 = perfect discrimination*) and can be interpreted like the effect size η^2 (*eta*-squared). Hence, *ANRI* values corresponding to η^2 values of .010, .059, and .138 indicate small, moderate, and large effects (Cohen, 1988). It is important to consider different measures as good calibration does not imply good discrimination (Yaniv, Yates, & Smith, 1991). For a detailed description of the formulae used, see Brewer and Wells (2006), and Yaniv et al. (1991). Finally, the sensitivity measure *d'* was calculated from hits and false identifications, which describes discriminability between targets and non-targets (Macmillan & Creelman, 2005). To calculate *d'*, the false identification rate was divided by lineup size. Note, that the results for *d'* were derived from the entire data set, including non-choosers.

Results

Identification Accuracy and Confidence across Age Groups

Mean accuracy and mean confidence across the age groups.

To test Hypothesis 1 that the confidence–accuracy relationship would deteriorate with increasing age, we calculated how confidence and accuracy relate to each other as a function of age group. Overall, 51.6% of choosers made a correct identification, whereas 48.8% made a false identification. The mean confidence was $M = 63.07$ ($SD = 26.47$), ranging between 0% and 100% (none was 10% confident). Chi-square analyses revealed that the effect of age group on identification accuracy was non-significant, $\chi^2(5, 436) = 10.58$, $p = .060$, Cramer's $V = .16$. Notably, the performance measure *d'* showed a strong increase in discriminability performance from adolescence (*d'* = 1.48) to young adulthood (*d'* = 1.91), followed by a steady decrease until older adulthood (*d'* = 1.35) as shown in Table 1.

Correlational analyses showed that while age and confidence were not related with each other, $r(434) = -.01$, $p = .883$; age^2: $r(434) = -.02$, $p = .748$, the relationship between age and identification accuracy was curvi-linear, $r_{pb}(434) = -.09$, $p = .071$; age^2: $r_{pb}(434) = -.10$, $p = .039$, showing an increase in identification accuracy from adolescence to young adulthood, followed by a decrease until older adulthood. Figure 1 displays the mean accuracies in the six age groups and the associated levels of confidence. Mean confidence varied little across the six age groups while identification accuracy showed a small increase followed by a decrease with increasing age.

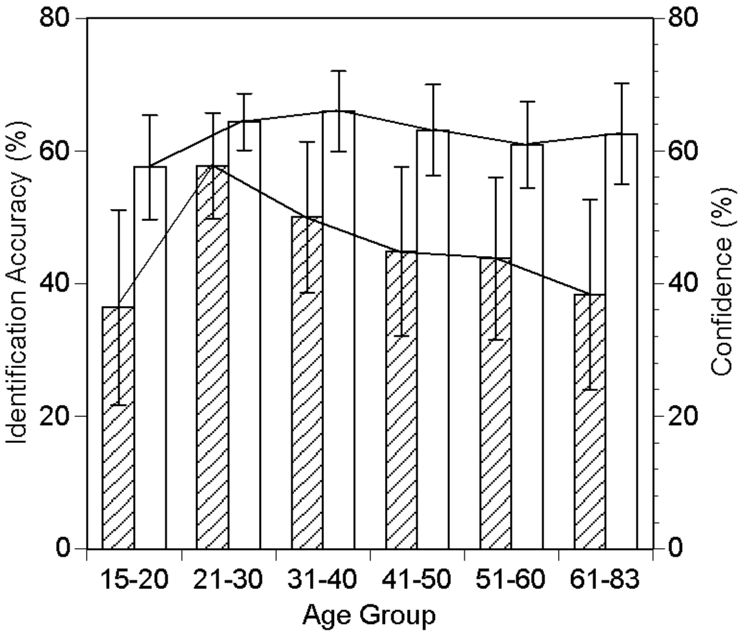


Figure 1. Mean confidence (%) and identification accuracy (%) of choosers as a function of age group. Error bars are 95% confidence intervals.

Table 1

Calibration (C), Over-/Underconfidence (O/U), Adjusted Normalized Resolution Index (ANRI), their Respective 5% Inferential Confidence Intervals (ICI), Recognition Performance (d'), and Point-biserial Correlations (r_{pb}) of Choosers as a Function of Age Group

Age group	***N***	**C [ICI]**	**O/U [ICI]**	**ANRI [ICI]**	***d'***	r_{pb}
15–20 years	44	.057	.211	.315	1.48	.56 **
		[.017, .098]	[.125, .297]	[.085, .546]		
21–30 years	149	.014	.066	.231	1.91	.45 **
		[.001, .027]	[.015, .117]	[.148, .315]		
31–40 years	74	.061	.159	.022	1.80	.27 *
		[.018, .104]	[.075, .239]	[.000, .088]		
41–50 years	58	.072	.188	.068	1.56	.18
		[.023, .122]	[.092, 284]	[.000, .178]		
51–60 years	64	.065	.173	.022	1.36	.12
		[.018, .111]	[.076, .270]	[.000, .094]		
61–83 years	47	.093	.242	.079	1.35	.34 *
		[.031, .155]	[.145, .339]	[.000, .211]		
All Choosers	436	.030	.147	.109	1.66	.34 **
		[.020, .044]	[.115, .180]	[.064, .153]		

*$p < .05$, **$p < .001$.

Note. Confidence values of the original 0% to 100% scale were re-categorized as follows: Low Confidence = 0–50 ($n = 191$); Medium Confidence = 60–80 ($n = 130$); High Confidence = 90–100 ($n = 115$). The point-biserial correlations r_{pb} are for accuracy correlated with the original 0–100% scale. Because ANRI can be interpreted as η^2, and η^2 is directly related to Cohen's *f*, cutoffs for small, moderate, and large ANRI values can be calculated using the .10, .25, and .40 cutoffs for *f*. The respective ANRI values are .010, .059, and .138.

Across all age groups, confidence was significantly associated with identification accuracy, $\chi^2(2, N = 436) = 49.96$, $p < .001$, Cramer's $V = .33$. However, the relationships were not consistent across age groups. With increasing age, confidence was less indicative of identification accuracy: The older the participants, the smaller the difference in confidence judgments between accurate and inaccurate responses. Figure 2 shows the regression lines of correct and incorrect responses and their associated 95% confidence bands, which crossed at the age of 55 years, indicating that adults were less likely to distinguish correct and incorrect identifications at the age of 55 years or more.

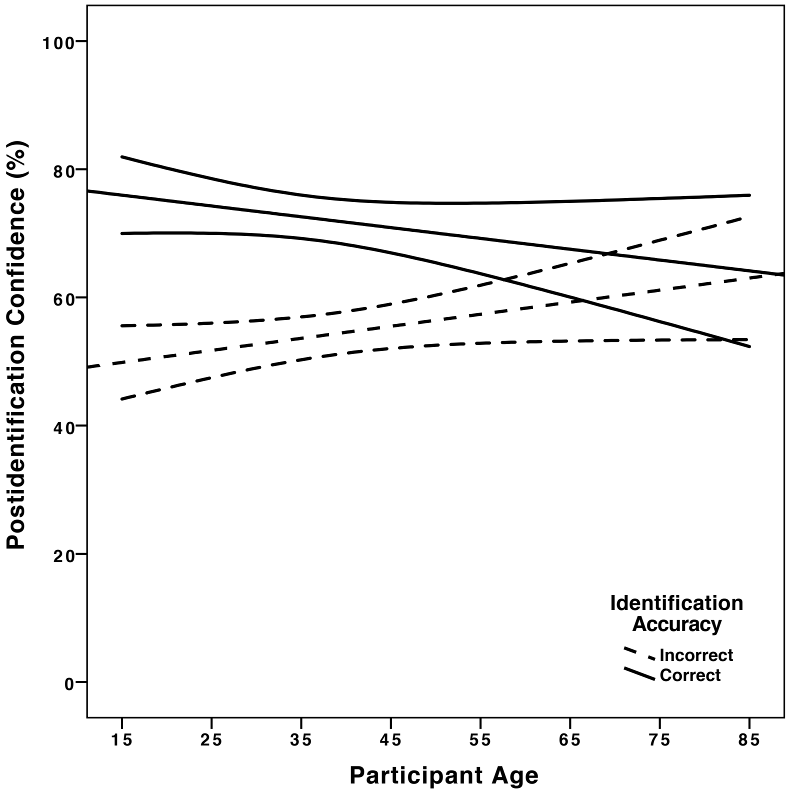


Figure 2. Confidence for correct and incorrect identification responses across the lifespan.

Confidence–accuracy characteristic analysis for each age group.

To further explore the confidence–accuracy relationship across the lifespan (Hypothesis 1), we conducted a CAC analysis, a graphic presentation which resembles the calibration curve. Figure 3 displays the mean identification accuracies for each age group with their respective 95% CIs at the three levels of confidence (low, medium, high). Values along the diagonal indicate well-calibrated witnesses, values above the line – underconfidence, and values below the diagonal – overconfidence[10]. As shown in Figure 3, high confidence was diagnostic of accuracy for only the two youngest age groups. Although low confidence was predictive of poor identification accuracy across all groups, with increasing age medium and high confidence became less indicative of identification accuracy. In particular, high confidence was diagnostic of accuracy in the 21–30 years age group only, and to some extent in the 15–20 years age group.

Calibration, over-/underconfidence and the discrimination of confidence judgments.

To test Hypothesis 1 statistically, in addition to the graphic representation, is to assess the statistics *Calibration, Over-/underconfidence, and Adjusted normalized resolution index* (Brewer & Wells, 2006). These indices were calculated by applying a bootstrap method (Efron, 1981; Efron & Gong, 1983) with n = 1000 replications to correct for a potential bias and to assess standard errors (SE) for the calculation of inferential confidence intervals (ICI; Tryon, 2001). The bootstrap analyses were conducted for each age group separately to calculate separate SE for them (Efron & Gong, 1983). The reduction factor E, defined as "the ratio of the SE of the difference between the two groups to the sum of the standard errors of both groups" (Tryon, 2001, p. 375), was averaged across all pairwise comparisons between the age groups. Non-overlapping ICIs showed significant differences between the age groups at the α = .05 level. A similar approach was applied by Palmer, Brewer, and Weber (2010) who tested the confidence–accuracy relationship of witnesses following post-identification feedback.

---

[10] Error bars are two-sided 95% confidence intervals calculated for proportions using an algorithm and spreadsheet developed by Newcombe (1998). We are grateful to Colin E. Tredoux for making us aware of this publication and sending us the spreadsheet.

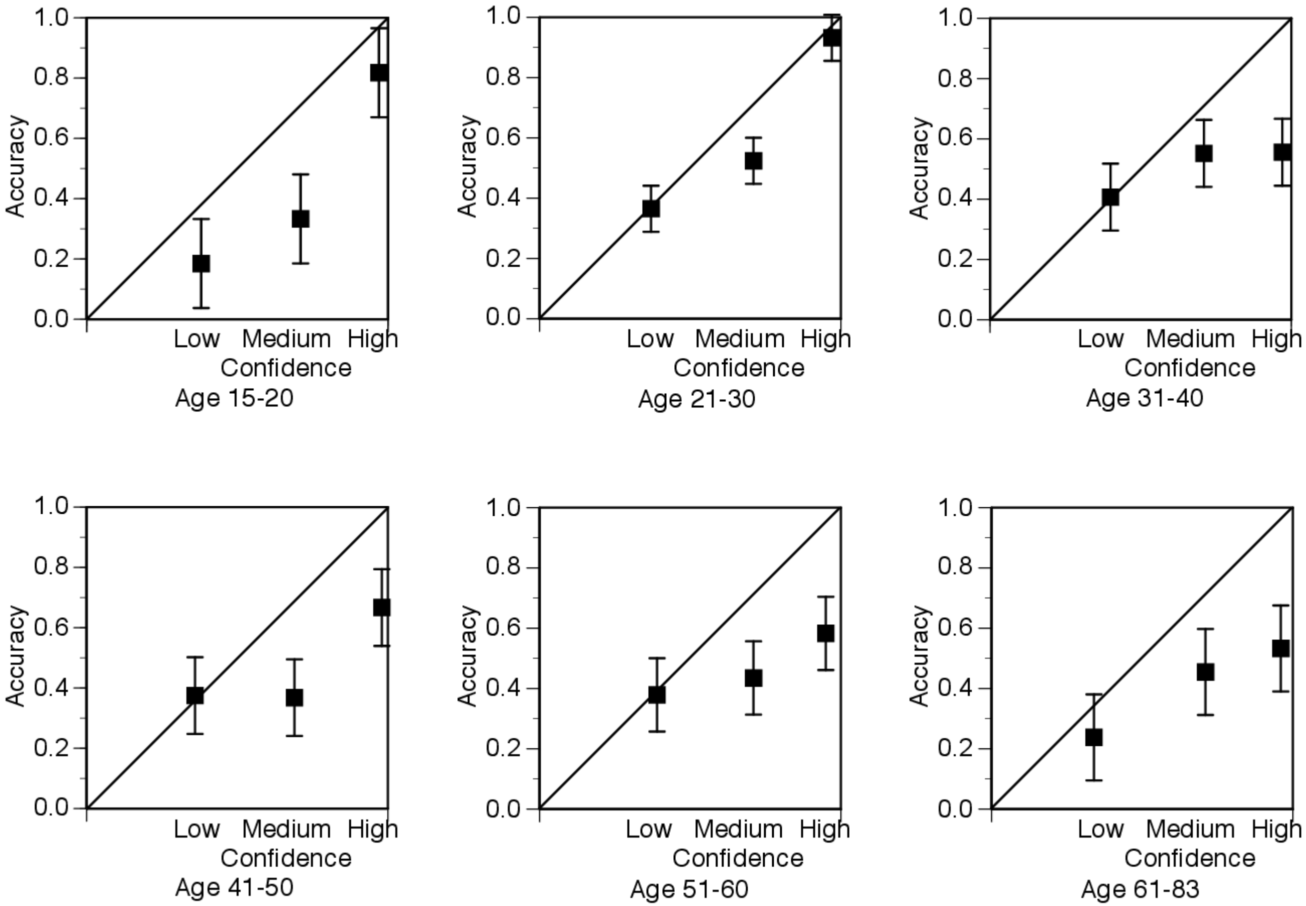


Figure 3. Confidence–accuracy characteristic analysis for choosers as a function of age group for low (0–50%), medium (60–80%), and highly (90–100%) confident witnesses. Error bars are 95% confidence intervals. Labels on the X-axis are placed under the means of the respective confidence range.

Results for calibration, over-/underconfidence and the *ANRI*, and their respective 5% ICIs are displayed in Table 1. These indices support the aforementioned findings: Although all age groups were overconfident for low and medium confidence (see Figure 2), calibration analyses revealed differences between the groups.

Young adults (21–30 years old) were well calibrated (*C* = .01, 5% ICI [.00, .03]); they showed, on average, very little overconfidence (*O/U* = .07, 5% ICI [.02, .12]), and their discrimination was above the average of the entire sample (*ANRI* = .23, 5% ICI [.15, .32]). By contrast, with increasing age, average confidence–accuracy correlations, calibration and discrimination decreased whereas overconfidence increased. In fact, adults older than 60 years had significantly weaker calibration and significantly higher overconfidence (*C* = .09, 5% ICI [.03, .16]; *OU* = .24, 5% ICI [.15, .34]) than adults aged between 21 and 30 years. Notably, the results for the youngest age group (15–20 years) varied across the measures: Although they showed the highest discrimination and correlation of all age groups (*ANRI* = .32, 5% ICI [.09, .55], r_{pb} = .56), their calibration was weaker than the calibration of 21 to 30 year olds (15–20 years: *C* = .06, 5% ICI [.12, .10]; 21–30 years; *C* = .01, 5% ICI [.00, .03]), and their degree of overconfidence was significantly higher than that of young adults, thus comparable to that of old-age participants (15–20 years: *O/U* = .21, 5% ICI [.13, .30]; 61–83 years: *O/U* = .24, 5% ICI [.15, .34]).

Interestingly, all age groups older than 31 years showed weak to no discrimination as indicated by the inferential confidence intervals that included zero (see Table 1). Furthermore, discrimination of young adults aged from 21 to 30 years was significantly higher than the discrimination of 31 to 40 year old and 51 to 60 year old adults (21–30 years: *ANRI* = .23, 5% ICI [.15, .32]; 31–40 years: *ANRI* = .02, 5% ICI [.00, .09]; 51–60 years: *ANRI* = .02, 5% ICI [.00, .09]).

Base Rates of Target-Presence

Next, we addressed the possible concern that the moderation of the confidence–accuracy relationship by age may hold only for certain base rates or prior probabilities that the suspect is the perpetrator (Hypothesis 2). Figure 4 maps the probability that a suspect identification was accurate (i.e., that the suspect is the perpetrator) across all possible base rate values from 0% (none of the lineups included a guilty suspect) to 100% (all lineups included a guilty suspect; see Wixted & Wells, 2017). One curve was created for each age group. The identity line shows where the data would fall if an identification was non-diagnostic.

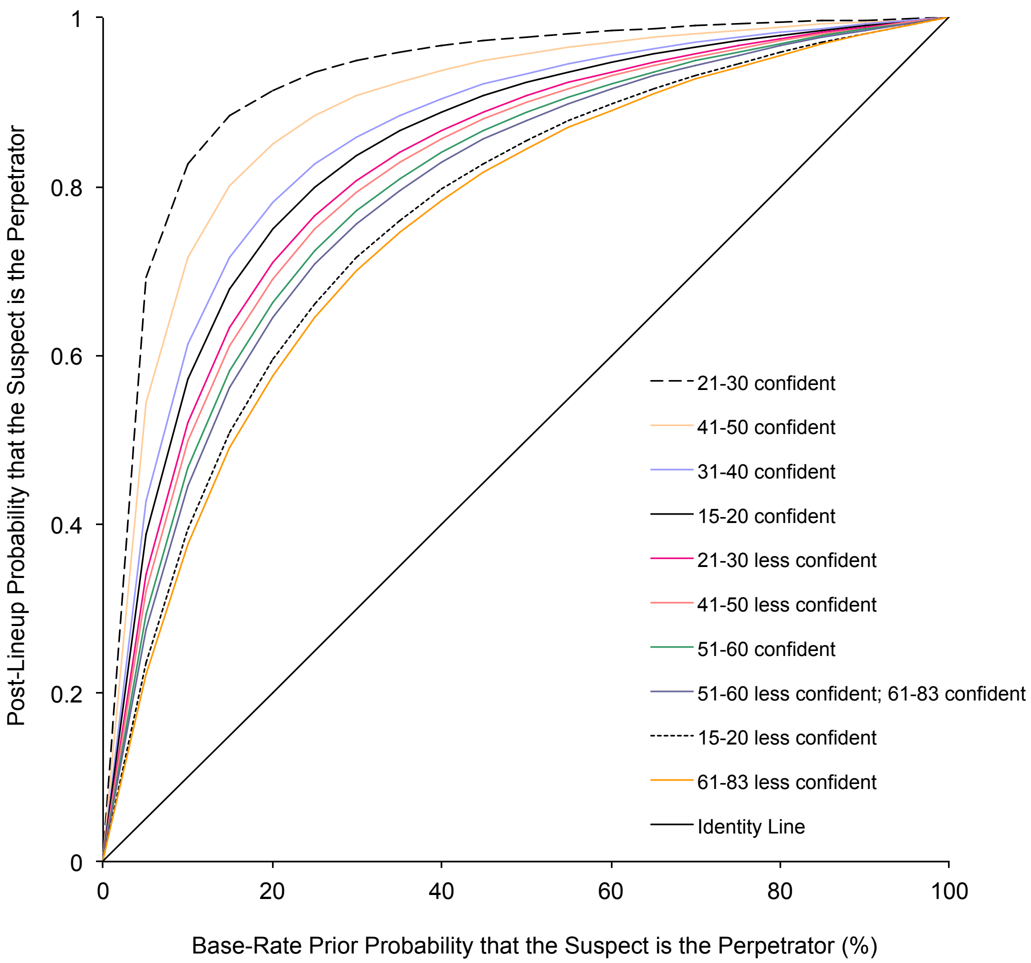


Figure 4. Post-lineup probability that the suspect is the perpetrator as a function of the base rate of target-presence for different age groups.

The Figure 4 shows that all curves are above the identity line, demonstrating that identifications were diagnostic of guilt. Generally, as expected, the heights of the curves for confident and younger participants are far above those of the curves for less confident or older participants. Strikingly, less confident decisions made by 21–30 year olds were still more diagnostic of guilt than highly confident decisions made by 51–83 year olds. In all, the Figure suggests that the moderating effect of age on the confidence–accuracy relationship prevails across different prior base rates of target-presence.

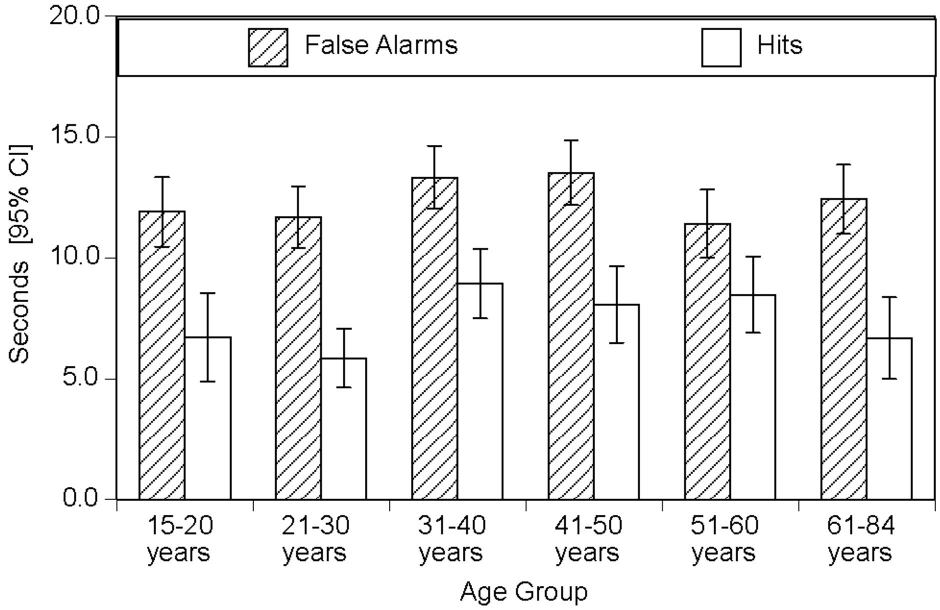
Observed and Self-reported Decision Time and Accuracy across the Lifespan

Observed decision times and self-estimated decision times showed positive skewness. Followingly, inferential analyses were conducted with log-transformed data (log-base 10) to approximate normal distributions. Means, SDs, and 95% CIs are reported for back-transformed values.

Analyses revealed that participants' age was not correlated with observed decision time, $r(434) = .08$, $p = .060$ (age^2: $r(434) = .08$, $p = .112$). However, there was a small positive correlation between age and self-estimated decision time: The older the participants, the longer the self-estimated time to identify a person from the lineup, $r(434) = .16$, $p = .001$ (age^2: $r(434) = .14$, $p = .006$).

As predicted by Hypothesis 3, the decision time was negatively correlated with the identification accuracy, $r_{pb}(434) = -.30$, $p < .001$, indicating that participants made a correct identification faster ($M = 7.05$ seconds, $SD = 2.62$, CI [6.07, 8.17]) than those who made a false identification ($M = 12.27$ seconds, $SD = 2.27$, 95% CI [10.86, 13.87]; $t(413) = 6.46$, $p < .001$). This was constant across all age groups, as shown in Figure 5a. By contrast, although the self-estimated decision time was generally shorter for correct identifications ($M = 7.35$ seconds, $SD = 3.02$, 95% CI [6.19, 8.72]) than for false identifications ($M = 11.59$ seconds, $SD = 2.75$, 95% CI [9.97, 13.48]; $t(434) = 4.48$, $p < .001$), this varied by age, supporting Hypothesis 4. Particularly adults aged between 31 and 40 years and older than 50 years overestimated their decision time when making a correct identification, and adolescents overestimated their decision time for both correct and false identifications, as shown in Figure 5b.

(a) Decision time



(b) Self-reported decision time

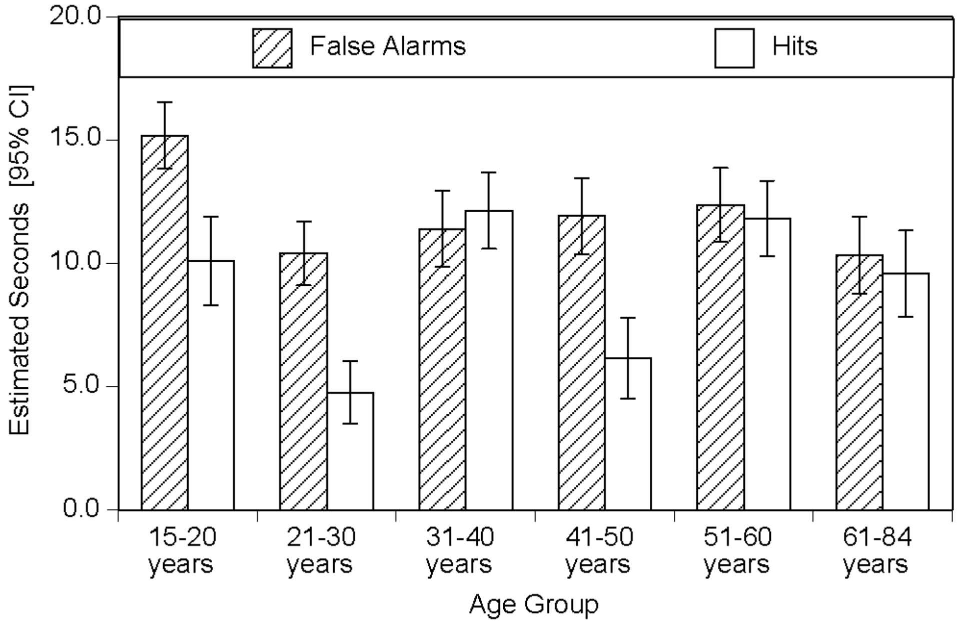


Figure 5. Decision time (a) and self-reported decision time (b) for correct and false identifications as a function of age group.

Combined Postdictive Value of Age, Confidence and Decision Time

Logistic regression analyses were conducted to combine the postdictors age, observed decision time, self-reported decision time, and postdecision confidence. The aim was to assess their respective associations with identification accuracy while controlling for the other postdictors. Adolescents were excluded from these analyses to control for curvi-linearity, which is difficult to detect in regression analyses.

First, a logistic regression analysis was conducted with the postdictors age, observed decision time, and accuracy. The full model was statistically significant, $\chi^2(3, 392) = 64.80$, $p < .001$, and explained 15.2% (Cox & Snell R^2) and 20.3% (Nagelkerke R^2) of the variance, respectively. Participants' age and decision time were negatively associated with identification accuracy while confidence was positively associated with it, as shown in Table 2a. The postdictive value was the strongest for decision time while controlling for age: Participants were four times more likely to correctly identify a person when the identification was fast (*OR* = 0.26, 95% CI [0.15, 0.46]; reciprocal *OR* = 3.85, 95% CI [2.16, 6.85], $p < .001$). Similarly, participants with higher confidence were almost twice as likely to correctly identify a person than those with lower confidence (*OR* = 1.93, 95% CI [1.46, 2.54], $p < .001$). Age remained a significant postdictor of identification accuracy showing a small but significant effect: The older the participants, the less likely they were to correctly identify the target person (*OR* = 0.98, 95% CI [0.97, 0.99], $p = .018$).

Separate analyses were conducted to assess the postdictive value of self-reported decision time on identification accuracy while controlling for age and confidence. The model was statistically significant, χ^2 (3, 392) = 46.39, $p < .001$, and explained 11.2% (Cox & Snell R^2) and 14.9% (Nagelkerke R^2) of the variance, respectively. Self-reported decision time was negatively associated with identification confidence, showing a small but significant effect (*OR* = 0.61, 95% CI [0.38, 0.98]; reciprocal *OR* = 1.63, 95% CI [1.02, 2.61], $p = .041$). Age and confidence remained significant postdictors with little variability from the model specified above. The full model details are in Table 2b.

Notably, when observed and self-reported decision times were combined in the same logistic regression model analyses, self-reported decision time did not contribute to the improvement of the model in Table 2a, that is, it did not explain any additional variance in the model (adding 0.0% to the Cox and Snell R^2, and 0.0% to Nagelke R^2, respectively).

Table 2

Logistic Regression of (a) Age, Observed Decision Time and Confidence on Identification Accuracy, and (b) Age, Self-reported Decision Time and Confidence on Identification Accuracy

	B	*SE*	*Wald's* χ^2 (1)	*p*	*OR*	*95% CI OR*
(a)						
Constant	0.789	0.499	2.497	.114	2.201	
Age	-0.018	0.007	5.572	.018	0.983	[0.968, 0.997]
Observed decision time	-1.346	0.294	20.952	<.001	0.260	[0.146, 0.463]
Confidence	0.656	0.140	21.898	<.001	1.928	[1.464, 2.538]
(b)						
Constant	-0.124	0.459	0.073	.787	0.883	
Age	-0.017	0.007	5.572	.018	0.983	[0.969, 0.997]
Self-reported decision time	-0.490	0.240	4.171	.041	0.613	[0.383, 0.980]
Confidence	0.688	0.140	24.089	<.001	1.990	[1.512, 2.620]

Discussion

For more than three decades, studies have assessed the postdictive value of eyewitness confidence following a suspect identification. While earlier meta-analyses concluded that confidence is an unreliable indicator of identification accuracy (Bothwell et al., 1987), a focus on choosers led to more optimistic conclusions (Sporer et al., 1995). More recent re-analyses of previously published data further emphasized the postdictive value of confidence, particularly for high confidence levels and under “pristine” identification conditions (Wixted et al., 2015; Wixted & Wells, 2017). These analyses, however, did not consider some potentially important moderators like race of target and witness (Wright, Boyd, & Tredoux, 2001), or old age of a witness. The present study aimed to fill this gap by analyzing the confidence–accuracy relationship as a function of witness age.

Our re-analyses of a large-scale field study (Sauerland & Sporer, 2009) revealed that age was a potentially important moderator of the confidence–accuracy relationship. As expected, the confidence–accuracy relationship was strong for younger participants (15–20 years) and young adults (21–30 years) and declined as age increased, confirming Hypothesis 1. This difference prevailed across different prior base rates of target-presence (Hypothesis 2), although the differences between age groups were strongest for small to moderate prior base rates of target-presence. Further, correct identifications were observed to be faster regardless of participants' age. However, the relationship between self-reported decision time and accuracy varied by age, showing little discrimination between correct and false identifications in older age (Hypothesis 3).

Similar to the studies re-analyzed in Wixted et al. (2015), the present study was an experimentally controlled field study. The identification instructions were fair (i.e. the participants were informed that the person might or might not be present in the lineup, and the interviewer was blind to target-presence and position in the lineup). Thus, our study fulfilled what Wixted and Wells (2017) described as pristine conditions. Moreover, participants did not know that they were part of a study until after the target walked away (unlike e.g. in Palmer et al., 2013). The latter factor is important for the legal system, thus enhances the ecological validity of our findings as in many criminal cases eyewitnesses do not anticipate that they will have to identify the person while they witness an event unless it is obvious from the beginning that they witness a crime. In support of this point, prior research showed that identification performance was poorer under incidental rather than intentional encoding conditions (Read, Lindsay, & Nicholls, 1998; Sporer, 1991).

Figure 3 shows that only the youngest two age groups were reasonably well calibrated. Particularly the results of young adults (21–30 years old) are in line with the findings by Wixted et al. (2015), namely that high confidence is indicative of identification accuracy (93.2% correct identification at high confidence level). This age group performed well and above the average of the entire sample on measures of calibration, over-/underconfidence, and the discrimination indices ANRI and the SDT performance measure *d'* (see Table 1). By contrast, with increasing age, confidence–accuracy correlations and calibration decreased whereas overconfidence increased. Moreover, ANRI discrimination was nearly zero for participants who were older than 30 years suggesting that the age-related decline may begin much earlier than at old age. Discrimination between targets and non-targets (*d'*) decreased with increasing age as has been shown across many studies (see the recent meta-analysis by Martschuk & Sporer, 2018). Taken together, age had a robust effect on every measure of identification performance while confidence remained unchanged across the lifespan.

The analyses in the present study differed from previous studies in the following aspects. Instead of averaging across large data sets we separated the age groups and showed a curvi-linear relationship between age and the postdictive value of the confidence–accuracy relationship. Further, the present study focused on choosers – those who would end up testifying in court. Although Key et al. (2015) analyzed the confidence–accuracy relationship between young, middle-aged and older adults, their results were not directly comparable to ours because the point-biserial correlations they reported were not separately calculated for choosers and non-choosers. As already pointed out in the introduction, the study by Colloff et al. (2017) differed in many ways from the current work. In particular the online methodology may have led to a self-selection bias in the older participant groups. Further, many analyses were conducted across choosers and non-choosers in Colloff et al. (2017) thus not making them directly comparable.

Decision time has often been found to be negatively associated with confidence and identification accuracy among choosers (e.g., Brewer & Weber, 2008; Sporer, 1992; Sauerland & Sporer, 2007, 2009). As postulated by the dual-process model, correct decisions are automatic and require little cognitive effort whereas incorrect decisions follow a conscious and time-consuming process (Dunning & Perretta, 2002; Dunning & Stern, 1994). The present findings confirmed that shorter decision times were associated with correct identifications, regardless of age. Moreover, the objectively measured decision time was a much stronger postdictor of identification accuracy (reciprocal $OR = 3.85$) than confidence ($OR = 1.93$). By contrast, self-reported decision time did not improve the postdictive value of observed decision time. It was, however, a reliable but less strong postdictor (reciprocal *OR* = 1.63) when it was substituted for measured decision time. In real cases, where decision times may not have been measured, self-reported decision time may be informative (but perhaps not with older witnesses over 50).

This further supports our theoretical assumptions that metacognitive processes vary by age. These findings demonstrate the importance of recording the entire identification procedure on video to not only capture the expressed confidence but also to be able to assess decision-time at the time of the original identification (Sporer, 1992, 1994). Confidence and self-reported decision times assessed retrospectively at later interviews or even in court may not only be affected by metacognitive deficits as a function of delay but, more troublesome, distorted by feedback effects (Steblay, Wells, & Douglass, 2014).

Limitations and Implications

The present study faced the following limitations. The participants in this study were not witnesses of a crime; instead, they were instructed to identify a person who approached them to ask for directions. Hence, the participants did not experience the level of stress that people might experience when they witness a crime. Nevertheless, the present findings are comparable to many real-life criminal cases, as the participants did not know that the person who approached them was a target until after the person walked away. Witnesses of crimes frequently do not anticipate that they will have to identify a person. Further, the scenario in the present study resembles a police call-out for people who saw or spoke to a person at a particular time and place, with the difference of a much shorter retention interval than in most criminal cases.

A further limitation of our study was the relatively small sample size per age group. Calibration curves require at least 200 participants or multiple judgments by each participant per condition to produce reliable results (Weber & Brewer, 2003). However, averaging across large data sets can obscure the results when moderators are not considered. To overcome this problem and approximate the calibration curve, we plotted the relationship between confidence and accuracy at fewer levels of confidence (low, medium, high). In addition, we applied different types of statistical analyses (as proposed Brewer & Wells, 2006). To obtain more reliable estimates of inferential confidence intervals using bootstrap analyses were conducted. This allowed the examination of the confidence–accuracy relationship from different perspectives. Our results (particularly the youngest age group) demonstrated the necessity of applying different types of analyses that yielded different coefficients that are not redundant. The calibration graph of adolescents showed robust calibration and a discrimination index that was the highest of all age groups. Nevertheless, their identification accuracy was as low (see Figure 3 above) and their overconfidence score was as high as that of people aged 61 years and over, attributable to the high confidence level despite low identification accuracy.

Further research is needed to assess the postdictive value of confidence of other samples of older people as well as of children, teenagers, or vulnerable witnesses. In addition, procedures should be explored that can improve the metacognitive processes at retrieval. For instance, Thomas, Gordon, and Bulevich (2014) suggested that metacognitive processes could be improved with supportive instructions that guide conscious processing of retrieved information, which results in more stringent metamemorial processes.

Finally, the targets in the present study were young to middle-aged adults. Previous research repeatedly demonstrated a robust effect of an own-age bias in eyewitness identification and face recognition (Kocab et al., 2018; Martschuk & Sporer, 2018; Rhodes & Anastasi, 2012). Thus it is possible that the age effect decreased if older targets were likewise included. Future research should assess whether this effect holds for older targets.

Conclusion

The results demonstrated that objective decision time had a greater postdictive value of identification accuracy than confidence and that the confidence–accuracy relationship depended on witness age. Whereas the high confidence expressed by young adults was indeed a reliable indicator of identification accuracy, the increased overconfidence of adolescents, middle-aged and older people, implies that we should take age and perhaps also the difference in age between the target and witness into account when postdicting accuracy from confidence.

Using age as a moderator, this study showed that factors which negatively influence identification accuracy do not necessarily affect post-decision confidence in the same way, leading to a decrease in the confidence–accuracy relationship. Thus, when evaluating eyewitness identification and post-decision confidence, decision-makers should be aware of factors that moderate identification accuracy, such as the cross-race or the old-age effect. Other factors might yield similar patterns and should be explored further. Finally, to better assess witness responses and how their postdictive value is evaluated by factfinders (Kaminski & Sporer, 2017), a lineup decision should be videotaped, which also provides information on decision times (response latencies), as suggested by Sporer (1992; Sporer et al., 1995) and many other researchers thereafter (e.g., Wixted & Wells, 2017). If, however, objective decision time is not available, the self-reported decision time together with confidence can assist in the evaluation of the likelihood of a correct identification, at least for young adults.

References

Bartlett, J. C., & Memon, A. (2007). Eyewitness memory in young and older eyewitnesses. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology*, *Vol. 2. Memory for people* (pp. 309–338). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). An exploratory study of personality differences in eyewitness memory. *Journal of Applied Psychology*, *72*, 691–695. doi:0021-9010/87

Brewer, N., & Weber, N. (2008). Eyewitness confidence and latency: Indices of memory processes not just markers of accuracy. *Applied Cognitive Psychology*, *22*, 827–840. doi:10.1002/acp.1486

Brewer, N., & Wells, G. (2006). The confidence–accuracy relation in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*, 11–30. doi:10.1037/a0031602

Buckhout, R. (1974). Eyewitness testimony. *Scientific American*, *231*, 23–31. doi:10.1038/scientificamerican1274-23

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence–accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*, 26–48. doi:10.3758/BF03210724

Castel, A. D., McGillivray, S., & Friedman, M. C. (2012). Meta-memory and memory efficiency in older adults: Learning about the benefits of priority processing and value-directed remembering. In M. Naveh-Benjamin & N. Ohta (Eds.), *Memory and aging: Current issues and future directions* (pp. 245–270). New York: Psychology Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge Academic.

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging*, *32*, 243–258. doi:10.1037/pag0000168

Cutler, B. L., & Penrod, S. T. (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

DeSoto, K. A., & Roediger, H. L. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science*, *25*, 781–788. doi:10.1177/0956797613516149

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence effors: A misrecollection account. *Psychology & Aging*, *22*, 122–133. doi:10.1037/0882-7974.22.1.122

Dodson, C. S., & Krueger, L. E. (2006). I misremember it well: Why older adults are unreliable eyewitnesses. *Psychological Bulletin & Review*, *13*, 770–775. doi:10.3758/BF03193995

Dunlosky, J., & Tauber, S. K. (2014). Understanding people's metacognitive judgments: An isomechanism framework and its implications for applied and theoretical research. In T. Perfect & D. S. Lindsay (Eds.), *Sage handbook of applied memory* (pp. 444–464). London, England: Sage Publications.

Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10-to 12-second rule for distinguishing accurate from inaccurate positive identification. *Journal of Applied Psychology, 87*, 951–962. doi:10.1037/0021-9010.87.5.951

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, *67*(5), 818–835. doi:10.1037/0022-3514.67.5.818

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap, and other methods. *Biometrica*, *68*, 589–599. doi:10.1093/biomet/68.3.589

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37, 36–48.* doi:10.1080/00031305.1983.10483087

Ellis, H. D., Davies, G. M., Shepherd, J. W. (1978). Remembering pictures of real and "unreal" faces. *British Journal of Psychology, 69*, 467–474.

Fitzgerald, R. J., & Price, H. L. (2015). Eyewitness identification across the life span: A meta-analysis of age differences. *Psychological Bulletin*, *141*, 1228–1265. doi:10.1037/bul0000013

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge: Harvard University Press.

Goldsmith, M., Pansky, A., & Koriat, A. (2014). Metacognitive control of memory reporting. In T. Perfect & D. S. Lindsay (Eds.), *Sage handbook of applied memory* (pp. 481–500). London, England: Sage Publications.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337. doi:10.1016/S1364-6613(00)01519-9

Hertzog, C., & Hultsch, D. F. (2000). Metacognition in adulthood and old age. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp. 417–466). Mahwah, NJ: Lawrence Erlbaum.

Jenkins, L., Myerson, J., Hale, S., & Fry, A. F. (1999). Individual and developmental differences in working memory across the life span. *Psychonomic Bulletin & Review*, *6*, 28–40. doi:10.3758/BF03210810

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

Kaminski, K. S., & Sporer, S. L. (2017). Discriminating between correct and incorrect eyewitness identifications: The use of appropriate cues. *Journal of Experimental Psychology: Applied*, *23*, 59–70. doi:10.1037/xap0000110

Keast, A., Brewer, N., & Wells, G. L. (2007). Children's metacognitive judgments in an eyewitness identification task. *Journal of Experimental Child Psychology*, *97*, 286–314. doi:10.1016/*j*.jecp.2007.01.007

Key, K. N., Cash, D. K., Neuschatz, J. S., Price, J., Wetmore, S. A., & Gronlund, S. D. (2015). Age differences (or lack thereof) in discriminability for lineups and showups. *Psychology, Crime & Law*, *21*, 871–889. doi:10.1080/1068316X.2015.1054387

Kocab, K., Martschuk, N., & Sporer, S. L. (2018). *Identification evidence by elderly witnesses: A meta-analysis*. Unpublished manuscript.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370. doi:10.1037/0096-3445.126.4.349

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology. General*, *123*, 297–315. doi:10.1037/0096-3445.123.3.297

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517. doi:10.1037/0033-295x.103.3.490

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 & Human Behavior*, *24*, 685–697. doi:10.1023/A:1005504320565

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

Martschuk, N., & Sporer, S. L. (2018). Memory of faces in old age: A meta-analysis. *Psychology & Aging*, *33*, 904–923. doi:10.1037/pag0000282

Meissner, C. A., Sporer, S. L., & Schooler, J. W. (2007). Person descriptions as eyewitness evidence. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology*, *Vol. 2. Memory for people* (pp. 3–34). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*, 93–102. doi:10.1016/j.jarmac.2015.01.003

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 125–173). New York: Academic Press.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17*, 857–872. doi:0.1002/(SICI)1097-0258(19980430)

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

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston.

Palmer, M. A., Brewer, N., & Weber, N. (2010). Postidentification feedback affects subsequent eyewitness identification performance. *Journal of Experimental Psychology: Applied*, *16*, 387–398. doi:10.1037/a0021034

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

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, *28*, 151–160. doi:10.1016/j.concog.2014.06.007

Penrod, S. & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, & Law*, *1*, 817–865. doi:10.1037%2F1076-8971.1.4.817

Perfect, T. J., & Weber, N. (2012). How should witnesses regulate the accuracy of their identification decisions: One step forward, two steps back? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1810–1818. doi:10.1037/a0028461

Price, J., Mueller, M. L., Wetmore, S., & Neuschatz, J. (2014). Eyewitness memory and metamemory in older adults. In M. Toglia, D. Ross, J. Pozzulo, & E. Pica (Eds.), *The elderly witness in court* (pp. 167–191). New York: Psychology Press.

Read, J. D. (1995). The availability heuristic in person identification: The sometimes misleading consequences of enhanced contextual information. *Applied Cognitive Psychology*, *9*, 91–121. doi:10.1002/acp.2350090202

Read, J. D., Lindsay, D. S., & Nicholls, T. (1998). The relationship between accuracy and confidence in eyewitness identification studies: Is the conclusion changing? In C. P. Thompson, D. Herrmann, J. D. Read, D. Bruce, D. Payne, & M. P. Toglia (Eds.), *Eyewitness memory: Theoretical and applied perspectives* (pp. 107–130). Mahwah, NJ: Erlbaum.

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin*, *138*, 146–174. doi:10.1037/a0025750

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological Science*, *22*, 464–471. doi:10.1177/0956797611401753

Richler, J. J., Mack, M. L., Gauthier, I., & Palmeri, T. J. (2009). Holistic processing of faces happens at a glance. *Vision Research*, *49*, 2856–2861. doi:10.1016/j.visres.2009.08.025

Roediger, H. L., & DeSoto, K. A. (2014). Confidence in memory: Assessing positive and negative correlations. *Memory*, *22*, 76–91. doi:10.1080/09658211.2013.795974

Roediger, H. L., & DeSoto, K. A. (2015). Understanding the relation between confidence and accuracy in reports from memory. In D. S. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. Roediger (Eds.), *Remembering: Attributions, processes, and control in human memory* (pp. 347–367). New York, NY: Psychology Press.

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence–accuracy relationship for eyewitness identification. *Law & Human Behavior*, *34*, 337–347. doi:10.1007/s10979-009-9192-x

Sauerland, M., Sagana, A., Sporer, S. L. & Wixted, J. T. (2018). Decision time and confidence predict choosers’ identification performance in photographic showups. *PLoS ONE, 13,* e019041. doi:10.1371/journal.pone.0190416

Sauerland, M., & Sporer, S. L. (2007). Post-decision confidence, decision time, and self-reported decision processes as postdictors of identification accuracy. *Psychology, Crime, & Law*, *13*, 611–625. doi:10.1080/10683160701264561

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*, 46–62. doi:10.1037/a0014560

Searcy, J. H., Bartlett, J. C., & Memon, A. (1999). Age differences in accuracy and choosing in eyewitness identification and race recognition. *Memory & Cognition*, *27*, 538–552. doi:10.3758/BF03211547

Searcy, J. H., Bartlett, J. C., Memon, A., & Swanson, K. (2001). Aging and lineup performance at long retention intervals: Effects of metamemory and context reinstatement. *Journal of Applied Psychology, 86*, 207–214. doi:10.1037/0021-9010.86.2.207

Shing, Y. L., Werkle-Bergner, M., Li, S.-C., & Lindenberger, U. (2009). Committing memory errors with high confidence: Older adults do but children don't. *Memory*, *17*, 168–179. doi:10.1080/09658210802190596

Sporer, S. L. (1991). Deep–deeper–deepest? Encoding strategies and the recognition of human faces. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *17*, 323–333. doi:10.1037/0278-7393.17.2.323

Sporer, S. L. (1992). Post-dicting eyewitness identification accuracy: Confidence, decision-times and person descriptions among choosers and non-choosers. *European Journal of Social Psychology*, *20*, 157–180. doi:10.1002/ejsp.2420220205

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology*, *78,* 22–33. doi:10.1037/0021-9010.78.1.22

Sporer, S. L. (1994). Decision times and eyewitness identification accuracy in simultaneous and sequential lineups. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 300–327). New York, NY, US: Cambridge University Press. doi:10.1017/CBO9780511759192.015

Sporer, S. L. (2001). The cross-race bias: Beyond recognition of faces in the laboratory. *Psychology, Public Policy, and Law*, *7*, 170–200. doi:10.1037/1076-8971.7.1.170

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

Thomas, A. K., Gordon, L. T., & Bulevich, J. B. (2014). Uniting theory to empirical evidence: How to understand memory of the elderly witness. In M. Toglia, D. Ross, J. Pozzulo, & E. Pica (Eds.), *The elderly witness in court* (pp. 308–335). New York: Psychology Press.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386. doi:10.1037/1082-989X.6.4.371

Tulving, E. (1985). How many memory systems are there? *American Psychologist*, *40*, 385–398. doi:10.1037/0003-066X.40.4.385

Watier, N. N., & Collin, C. A. (2011). Metamemory for faces, names, and common nouns. *Acta Psychologica*, *138*, 143–154. doi:10.1016/j.actpsy.2011.05.017

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. (2008). Eyewitness recall: Regulation of grain size and the role of confidence. *Journal of Experimental Psychology: Applied*, *14*, 50–60. doi:10.1037/1076-898x.14.1.50

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., ... Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*, 264–271. doi:10.1016/j.concog.2013.01.004

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., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology*, *78*, 835–844. doi:10.1037/0021-9010.78.5.835

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115–1125. doi:10.1177/01461672992512005

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior*, *39*, 99–122. doi:10.1037/lhb0000125

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*, 515–526. doi:10.1037/a0039510

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence–accuracy relationship. *Journal of Applied Research in Memory and Cognition*, *4*, 329–334. doi:10.1016/j.jarmac.2016.04.006

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

Wright, D. B., Boyd, C. E., & Tredoux, C. G. (2001). A field study of own-race bias in South Africa and England. *Psychology, Public Policy, & Law*, *7*, 119–133. doi:10.1037/1076-8971.7.1.119

Wright, D. B., & Stroud, J. (2002). Age differences in lineup identification accuracy: People are better with their own age. *Law & Human Behavior*, *26*, 641–654. doi:10.1023/A:1020981501383

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

MOCK JURORS' AWARENESS OF AGE-RELATED CHANGES IN MEMORY AND COGNITIVE CAPACITY

In the course of the last century life expectancy increased. An implication for criminal justice is that older people are more and more at risk to become victims of crimes. Crime statistics show that the victimization rate of older people in proportion to young people varies by crime. For instance, older people are at a higher risk to become victims of robbery or handbag snatching than victims of homicide, and these crimes are more likely to be perpetrated by strangers (James & Graycar, 2000; Walker, 2018). We used such a scenario as the focus of the present study to investigate how mock jurors perceive older victims as witnesses and use their testimony in a robbery trial.

In a jury trial it is for jurors to decide, based on the presented evidence, as to whether the defendant is guilty or not. Typically, criminal (and other jury) trials contain contradictory accounts of events (Bornstein & Greene, 2011). In cases in which the only witness is the victim, jurors are faced with the question whether the victim was willing *and* able to tell the truth. In other words, they have to decide whether the victim provided an honest, truthful (i.e., credible in the sense of lack of *deception*) *and* accurate, reliable account (i.e., credible in the sense of lack of *error*; Sporer, 2008). These two constructs are not exclusive and need to be considered simultaneously when assessing witness credibility. This complexity is also mirrored in jury instructions, as in their decision the jury has the discretion to accept or reject the whole or a part of the evidence (Judicial Commission of New South Wales, 2016; Queensland Courts, 2016).

According to the *conceptual model* of juror decision-making (Devine & Caughlin, 2014), case characteristics (e.g., victim characteristics, crime type), juror characteristics (e.g., gender, experiences, legal trust), and defendant characteristics (e.g., race, criminal history) are associated with inferences made about the case, and ultimately with judgments of guilt. In all, jurors are "active information processors" (Bornstein & Greene, 2011, p. 65) whose decision reflect not only evidence and legal guidelines (i.e., legal factors) but also juror stereotypes and attitudes (i.e., extra-legal factors) which in turn may affect case perceptions. The present study investigated the extent to which victim age and crime severity as case characteristics, and legal trust as a juror characteristic contribute to juror decisions. Defendant characteristics were held constant in this study.

Victim Age as a Case Characteristic

Stereotypes and attitudes are not only held about defendants but also about victims and witnesses. Here, we investigate perceptions of older people in a trial. Considering age-related reductions in the efficiency of the sensory system (Schieber, 2006) and age-related changes of brain structure (Coleman & O'Hanlon, 2008), older people have more difficulties to process and remember newly learned information. This applies to memory for faces (see the meta-analysis by Martschuk & Sporer, 2018), eyewitness recall of events (see the overview by Aizpura, Migueles, & Garcia-Bajos, 2014), and metacognitive judgments (e.g., Dodson & Krueger, 2006; Martschuk, Sporer, & Sauerland, 2019). For instance, despite the age-related decline in identification performance, expressed confidence that the identification was correct remained relatively unchanged across the lifespan (Martschuk et al., 2019). A meta-analysis of studies of stereotypes about older people showed that older people were perceived as less competent than young people, but the effect decreased when more detailed information about the target was provided than age alone (Kite, Stockdale, Whitley, & Johnson, 2005).

Studies that investigated perceptions of older witnesses (as opposed to older people in general) revealed conflicting patterns. While some studies found that older witnesses were perceived more positively than young or middle-aged adults (e.g., Allison, Brimacombe, Hunter, & Kadlec, 2006; Kwong See, Hoffmann, & Wood, 2001; Ross, Dunning, Toglia, & Ceci, 1990), other studies showed the opposite effect (e.g., Neal, Christiansen, Bornstein, & Robicheaux, 2012). Factors, such as questioning type, speech style, or demographic background were associated with perceived witness credibility (Brimacombe, Jung, Garrioch, & Allison, 2003; Nunez, McCoy, Clark, & Shaw, 1999). Further factors were the content or narrative style of a given testimony. For instance, a study investigated the quantity and quality of eyewitness statements regarding a staged crime video under direct and cross-examination of young (18–25 years), middle-aged (30–44 years) and older adults (65–85 years) and their consequences on perceived credibility (Brimacombe, Quinton, Nance, & Garrioch, 1997).

Older adults were indeed less accurate under direct and cross-examination and provided fewer details of the offender than young and middle-aged adults. Consequently, when another group of participants evaluated these videotaped testimonies, they perceived older people as less credible (perceived description accuracy, perceived confidence and perceived competence) but also as more honest than younger adults.

Altogether, these findings reflect the complex construct of witness credibility, which consists of "witness competence, confidence, convincingness, accuracy, honesty, witness's observation of the event, memory for the event, witness's level of cognitive functioning, and witness suggestibility" (Pittman, Toglia, Leone, Mueller-Johnson, 2014, p. 368), mirroring both elements of credibility: honesty and accuracy (Sporer, 2008). It includes the individual's credibility (i.e., *believability*, often based on witness demeanor and attributed character) and evidentiary credibility (i.e., *plausibility* of the facts provided; Andrewartha, 2014).

The present study tested both types of credibility as a function of witness age (young vs. older adult) and crime severity (injury vs. no injury) in a robbery case scenario. Furthermore, this study considered the extent to which mock jurors' beliefs about memory changes in old age were associated with their decisions. Due to the generally known age-related memory decline and the decreased identification performance associated with it, we hypothesized that mock jurors will be less likely to rely upon the testimony of a 75 years old victim than a 25 years old victim (Hypothesis 1a). Following, mock jurors will perceive the prosecution case weaker (Hypothesis 1b) and the defense case stronger when the victim was 75 years old (Hypothesis 1c).

Crime Severity as a Case Characteristic

Conflicting arguments exist as to whether and how crime severity influences juror decisions. According to the *severity-leniency hypothesis* (Kerr, 1978), jurors are less willing to convict when the crime is more severe. This hypothesis predicts that in order to avoid the risk of convicting an innocent person, jurors are more reluctant to convict in more serious cases (which would result in more severe sentences), thus require more evidence of guilt. Studies that investigated the influence of charge severity supported the severity-leniency hypothesis inasmuch as jurors were more likely to convict the defendant when the charge was more severe (McComas & Noll, 1974, as cited in Kerr, 1978). This study, however, did not control for strength of evidence between the charges.

Another study showed support for the severity-leniency hypothesis only when evidence was more ambiguous in a more severe crime, that is, when the conduct was the same but more evidence was considered necessary to prove guilt in a first-degree murder vs. second-degree murder vs. manslaughter case (Freedman, Krismer, MacDonald, & Cunningham, 1994, Experiment 1). However, when Freedman et al. (1994, Experiments 2 to 4) controlled for the strength of evidence, that is, when proof was provided that the crime was committed, and mock jurors had to decide whether the defendant was the perpetrator, the association between crime severity (robbery vs. robbery and assault vs. robbery and aggravated assault) and verdict disappeared. In other words, mock jurors were as likely to convict when the crime was robbery as they were when the robbery was followed by assault or aggravated assault, when evidence strength was the same.

A different theoretical perspective postulates that evidence of a serious offence may evoke emotional arousal (Bright & Goodman-Delahunty, 2004). As a consequence, jurors' perception of the defendant's guilt is elevated by their negative emotional state. Referring to the potential prejudicial impact of gruesome evidence, the Australian Law Reform Commission (1985) cautioned that "the fact-finder may use the evidence to make a decision on an improper, perhaps emotional, basis, [...] logically unconnected with the issues in the case" (pp. 351–352). Put another way, jurors may engage in a biased interpretation of emotionally laden information (e.g., verbal gruesome evidence) which in turn may increase conviction rates (Bright & Goodman-Delahunty, 2004). A recent meta-analysis confirmed that the presence of visual (photographic) gruesome evidence increased conviction rates in other studies (Grady, Reiser, Garcian, Koeu, & Scurich, 2018).

The same principle applies to the effect of crime severity and outcome severity, such that jurors confronted with evidence of a violent crime experience strong negative emotions. Accordingly, jurors may lower their threshold to convict or accord undue weight to the presented evidence when faced with the prospect of erroneously acquitting a violent perpetrator who then would return to the community (Devine, 2012).

Some evidence for the influence of crime severity on mock jurors' emotions and verdicts was gathered in a study that compared responses to life-threatening conduct vs. less threatening conduct (Goodman-Delahunty, Martschuk, & Ockenden, 2016). Mock jurors expressed more negative emotions and were four times more likely to convict the defendant when the crime was more severe. Notably, the effect disappeared when the defendant was additionally charged with a terrorist crime, presumably because the charge severity outshone or outweighed the conduct severity when terrorism was mentioned. The complex interplay of crime type (e.g., shoplifting, rape, murder, child abuse), type of evidence (none, witness, physical evidence, DNA evidence) and criminal history (none, prior conviction) was also investigated by Pearson et al. (2018). After controlling for evidence type, crime severity was associated with more guilty verdicts and higher punishment ratings.

A similar pattern to those observed in jurors' judgments of guilt emerged in studies on sentencing decisions. Mock jurors were more punitive (i.e., awarded longer sentences) when the victim reported more severe injuries compared to mild injuries in cases of burglary and robbery (Nadler & Rose, 2003). Mock judges were also more punitive following unintended consequences of the defendant's actions that involved the killing vs. the injury of a third party (Sporer & Goodman-Delahunty, 2009).

Taken together, the foregoing research supports the theoretical notion that mock jurors are less conservative in their decisions and hence more likely to convict when the crime involved is more emotionally arousing and/or considered more severe. Accordingly, we hypothesized that mock jurors will be more likely to believe the complainant when she was injured during the robbery than when she was not (Hypothesis 2a). This should be reflected in the perceptions of a stronger prosecution case (Hypothesis 2b) and a weaker defense case (Hypothesis 2c).

Juror Attitudes as Juror Characteristics

In addition to case characteristics, the present study investigated attitudes as juror characteristics in legal decisions. The association between juror characteristics, juror attitudes, and their decisions has been the focus of research since at least the 1950s (e.g., Adorno, Levinson, Frenkel-Brunswik, & Sanford, 1950). While some researchers investigated personality traits, others assessed legally relevant attitudes (for an overview, see Lecci & Myers, 2008). Studies repeatedly showed that juror attitudes have an important role in their legal decisions (see the meta-analysis by Devine & Caughlin, 2014), particularly when trial evidence is ambiguous.

As Sherif and Sherif (1969) posited long ago in the context of social perception, the more ambiguous a situation (i.e., the stimulus to be judged) becomes, the greater is the contribution of internal factors (e.g., an individual's past experiences, social attitudes and temporary expectations) on decisions (see also Sporer & Goodman-Delahunty, 2009). In the context of criminal trials, the more conflicting or equivocal the evidence, the stronger is the contribution of juror characteristics and attitudes on the assessment of the presented evidence, and consequently on verdict. Considering that the prosecution case and the defense case are typically contradicting (Bornstein & Greene, 2011), jurors' internal factors are likely to contribute to the evaluation of the presented evidence, at least to some extent. This was shown by a study that tested the predictive value of mock jurors' pretrial attitudes on pre- and post-deliberation verdicts in a simulated armed robbery trial (Lecci & Myers, 2009). Analyses revealed that mock jurors' confidence in the justice system and conviction proneness were associated with their verdicts.

Based on this reasoning we derived the following additional hypotheses: Mock jurors' attitudes towards the justice system and towards ageing will be associated with the perceived likelihood that the identification of the perpetrator was correct (Hypothesis 3a), the probability of guilt (Hypothesis 3b), and the guilty verdict (Hypothesis 3c). Furthermore, we hypothesized that, after controlling for mock jurors' attitudes, perceived victim credibility and perceptions of the prosecution and defense case would be associated with the perceived likelihood that the identification was correct (Hypothesis 4a), the probability of guilt (Hypothesis 4b), and the guilty verdict (Hypothesis 4c).

Method

Participants

After screening,[11] 205 participants (120 women; 84 men; 1 unknown), aged between 18 and 85 years (*M* = 31.24, *SD* = 10.95) were included in the study. About half of the participants were U.K. citizens (52.7%), 31.7% were U.S. citizens, 12.7% Canadians, 4.9% Australians, and 2.0% New Zealand citizens. Four percent were dual citizens. Most participants were White (79.2%), 11.4% Asian, 3.5% Hispanic or Latino, 3.0% Black, and 3.0% were mixed. More than half of the participants had a university degree (Bachelor's Degree 39.7%, Master's Degree 9.8%, Doctorate 3.9%), 21.6% had a Diploma or a trade certificate, and 25.0% had completed high school or less. One-fifth of the participants reported that they had been called to jury duty (21.6), and 6.9% had served on a jury trial.

Stimulus Material

Trial scenario.

The trial was a summary of a robbery trial (ca. 800 words), which included general instructions, opening of the prosecution and the defense case, evidence-in-chief and cross-examination of the victim and the investigating officer. The participants were informed that the defendant did not testify following his lawyer's advice. All case facts were held constant, except for victim age and crime severity. See Appendix A.

Questionnaire.

The questionnaire included items that measured guilt both as a binary variable (*guilty* vs. *not guilty*) and the probability of guilt (on an 11-point rating scale from 0–100%), the likelihood of correct identification (11-point rating scale from 0–100%), and the strength of

---

[11] Seven participants were excluded from the study, because they (a) took less than one minute to read the trial scenario (*n* = 4), (b) did not complete the study (*n* = 2), or (c) responses indicated misunderstanding of questions (*n* = 1). One participant did not provide demographic information but completed the experimental tasks and was, therefore, not excluded from the study.

the prosecution and the defense case (7-point rating scale; *1 = strongly disagree; 7 = strongly agree*). The questionnaire is in Appendix B.

Perceived victim credibility was assessed using the Eyewitness Credibility Scale with the subscales *Victim Believability* (11 items; e.g., the victim accurately remembered the event; Cronbach's $\alpha = .89$) and perceived *Victim Lenient Response Bias* (3 items; e.g., the victim wanted to have someone to blame; Cronbach's $\alpha = .73$). Participants indicated their agreement on a 7-point rating scale from *1 = strongly disagree* to *7 = strongly agree*. The inter-correlation between *Victim Believability* and *Victim Lenient Response Bias* was $r = -.24$, $p < .05$. The results of the principal component analysis are in Appendix C.

Participants' attitudes towards the justice system were assessed using the Pretrial Juror Attitude Questionnaire (PJAQ; Lecci, & Myers, 2008, 2009) subscales *System Confidence* (6 items, e.g., A suspect who runs from police, probably committed a crime; Cronbach's $\alpha = .70$[12]), *Conviction Proneness* (5 items; e.g., Criminals should be caught and convicted by "any means necessary"; Cronbach's $\alpha = .63$), and *Social Justice* (4 items; e.g., Wealthy individuals are almost never convicted of their crimes; Cronbach's $\alpha = .66$). *System Confidence* and *Conviction Proneness* were highly correlated with each other $r = .58$, $p < .001$, while they were not correlated with *Social Justice*. *System Confidence* and *Conviction Proneness* were combined (hereafter referred to as the *Pretrial Juror Bias*) to one composite measure yielding a higher degree of reliability (Cronbach's $\alpha = .78$) than either subscale separately.

Participants' beliefs about memory changes in old age were assessed using the subscale *Stereotype of Capable-Incapable*[13] of the Ageism Scale (10 items; e.g., with old age, people

---

[12] Cronbach's $\alpha \geq .6$ is interpreted as questionable, Cronbach's $\alpha \geq .7$ as acceptable, Cronbach's $\alpha \geq .8$ as good, and Cronbach's $\alpha \geq .9$ as excellent (George & Mallery, 2003).

[13] As noted by Kite and Wagner (2002), some measures about ageing do not necessarily reflect age-related bias; instead, they reflect accurate perceptions of ageing. The subscale used here included items that were in line with research findings that memory and cognitive abilities decline in older age (e.g., Coleman & O'Hanlon, 2008; Park et al., 2002).

are inclined to become forgetful; Cronbach's α = .85; Braithwaite, Lynd-Stephenson, & Pigram, 1993). For each item, participants indicated their agreement on a 7-point rating scale from *1 = strongly disagree* to *7 = strongly agree*. Item order was randomized within each scale to control for order effects.

Research Design

The study was a randomized 2 x 2 between-participants design with victim age (25 years vs. 75 years) and crime severity (no injury vs. sprained ankle) as the manipulated variables. The four groups were (1) 25-years old victim, no injury ($n = 51$); (2) 25-years old victim, sprained ankle ($n = 54$); (3) 75-years old victim, no injury ($n = 54$); and (4) 75-years old victim, sprained ankle ($n = 47$).

Procedure

The study was approved by the Local Ethics Committee at the Justus-Liebig-University of Giessen (approval number: 2017-0009). Participants were recruited via Prolific, an online crowdsourcing platform (*www.prolific.ac*), and participated online.

Participants were randomly assigned to one of the experimental groups. They read the trial scenario and answered a series of questions about the trial and the victim. Further, they indicated their agreement to attitudinal items and provided demographic information. Following, they were debriefed and reimbursed for their time (£ 2 = approximately US $ 2.60). Participation was about 10–15 minutes ($M = 11.55$, $SD = 4.70$).

Pilot study.

A pilot study with 40 participants revealed that the manipulation of victim age was successful. When the victim was young, participants were more likely to convict the defendant (57.1% guilty) and perceive that he committed the crime ($M = 74.3\%$, $SD = 23.64$) than when the victim was older (19.2% guilty, $\chi^2(1, 40) = 5.96$, $p = .015$, phi = -.386, $OR = 5.6$; probability of guilt: $M = 51.2\%$, $SD = 24.39$, $F(1, 40) = 7.05$, $p = .012$, $g_u = 0.94$). Further, participants were more likely to agree that a young victim made a correct identification ($M = 75.7\%$, $SD = 21.74$) than her older counterpart ($M = 59.6\%$, $SD = 23.06$, $F(1, 40) = 4.43$, $p = .042$, $g_u = 0.70$). Crime severity was not associated with any of the dependent measures ($p > .05$).

Perceived victim credibility consisted of eighteen items. Analyses revealed the presence of multicollinearity for several items ($r > .80$) and a very low corrected item-total correlation (*CITC*) for one item (*CITC* = .08). Following, redundant items were removed or rephrased and the item with low *CITC* was removed from the questionnaire. The final questionnaire consisted of 14 items (see Appendix B).

Results

In the following, we first present the results of 2 x 2 between-participants analyses of variance (ANOVAs) of victim age and crime severity on perceived victim credibility and on mock jurors' perceptions of the case strength for the prosecution and for the defense as dependent variables. In addition, we report a paired-sample *t*-test to assess whether the prosecution or the defense case was stronger. Next, we report two hierarchical multiple regression analyses to assess the predictive value of mock jurors' demographics, general attitudes, perceived victim credibility, and case strength for the prosecution and the defense on (a) the perceived likelihood that the identification was correct, and (b) the probability of guilt.

Finally, we report results on verdict. The reported effect sizes are the corrected standardized mean difference Hedges g_u, Pearson correlations r, and point-biserial correlations r_{pb} (Lipsey & Wilson, 2001). In addition, semi-partial correlations (*sr*) are reported with multiple regression analyses to assess the unique contribution of each variable to the regression model (Tabachnick & Fidell, 2013). Effect sizes are interpreted as small for Pearson correlations and point-biserial correlations of .10 and standardized mean differences of 0.20; moderate for Pearson correlations of .30 and point-biserial correlations of .24 and a standardized mean difference of 0.50; and large for Pearson correlations of .50 and point-biserial correlations of .37 and a standardized mean difference of .80 (Cohen, 1988).

Perceived Victim Credibility

Perceived victim credibility was measured using two factors, Victim Believability ($M = 4.55$, $SD = 1.01$) and Victim Lenient Response Bias ($M = 4.07$, $SD = 1.37$). Table 1 shows the means and standard deviations of credibility measures, convincingness of the prosecution and the defense case, and the perceived probability of guilt by victim age and injury presence.

Two-way ANOVAs showed no main effects of victim age and crime severity on perceived Victim Believability, $F(1, 201) = 1.45$, $p = .231$, $g_u = 0.20$, 95% CI [-0.08, 0.47], and $F(1, 201) = 0.05$, $p = .832$, $g_u = -0.05$, 95% CI [-0.33, 0.22], respectively, nor on perceived Victim Lenient Response Bias, $F(1, 201) = 0.51$, $p = .477$, $g_u = 0.18$, 95% CI [-0.10, 0.45], and $F(1, 201) = 3.70$, $p = .056$, $g_u = 0.19$, 95% CI [-0.46, 0.08], respectively. The interaction between victim age and crime severity was significant for perceived Victim Believability, $F(1, 201) = 4.52$, $p = .035$, $\eta_p^2 = .022$, but not for perceived Victim Lenient Response Bias, $F(1, 201) = 0.02$, $p = .877$, $\eta_p^2 = .000$. Post-hoc analyses using a Tukey test showed no age effect in absence of injury, $g_u = -0.13$, 95% CI [-0.51, 0.25]. By contrast, when the victim reported she was injured during the robbery, mock jurors believed the 25 years old victim more than the 75 years old victim, $g_u = 0.45$, 95% CI [0.06, 0.85]. Figure 1 shows the interaction effect between victim age and crime severity on perceived victim credibility.

Table 1

Victim Credibility, Case Strength, Probability of Guilt, and Verdict as a Function of Victim Age and Crime Severity

	Young Victim	Older Victim	
	M (SD)	*M (SD)*	Effect size g_u [*95% CI*]
Victim Believability	4.63 (1.02)	4.47 (1.00)	0.16 [-0.12, 0.43]
No injury	4.49 (1.04)	4.62 (0.91)	-0.13 [-0.51, 0.25]
Injury	4.76 (0.98)	4.29 (1.07)	**0.45 [0.06, 0.85]**
Victim Response Bias	3.99 (1.45)	4.15 (1.29)	-0.11 [-0.38, 0.16]
No injury	4.20 (1.51)	4.30 (1.29)	-0.08 [-0.46, 0.31]
Injury	3.80 (1.37)	3.96 (1.28)	-0.13 [-0.51, 0.27]
Prosecution Case Strength	3.12 (1.70)	2.80 (1.63)	0.19 [-0.08, 0.46]
No injury	2.78 (1.54)	3.13 (1.61)	-0.22 [-0.60, 0.16]
Injury	3.43 (1.79)	2.43 (1.58)	**0.59 [0.19, 0.98]**
Defence Case Strength	4.52 (1.67)	4.95 (1.64)	-0.26 [-0.53, 0.02]
No injury	4.92 (1.44)	5.02 (1.42)	-0.07 [-0.45, 0.31]
Injury	4.13 (1.80)	4.87 (1.87)	**-0.40 [-0.79, -0.01]**
Likelihood of a correct ID	60.39 (20.19)	56.44 (19.58)	0.20 [-0.08, 0.47]
No injury	57.65 (21.22)	58.15 (19.04)	-0.03 [-0.41, 0.36]
Injury	63.02 (18.97)	57.47 (20.20)	0.28 [-0.11, 0.67]
Probability of Guilt	55.39 (21.49)	51.68 (20.15)	0.18 [-0.10, 0.45]
No injury	51.76 (22.42)	51.48 (21.14)	0.01 [-0.37, 0.39]
Injury	58.87 (20.16)	51.91 (19.18)	0.35 [-0.04, 0.74]
Guilty verdict	22.1%	14.9%	*OR* = 1.63 [0.79, 3.34]
No injury	17.6%	13.0%	*OR* = 1.44 [0.49, 4.20]
Injury	26.4%	17.0%	*OR* = 1.75 [0.66, 4.64]

Note. Bold values indicate a significant age effect.

Figure 1. Perceived witness credibility as a function of victim age and crime severity. Error bars are 95% confidence intervals.

Prosecution vs. Defense Case Strength

Overall, mock jurors perceived the defense case stronger, $M = 4.73$, $SD = 1.67$, than the prosecution case, $M = 2.96$, $SD = 1.67$, $t(204) = 9.19$, $p < .001$. Further, the defense and the prosecution case strength were negatively correlated with each other, $r = -.37$, $p < .001$. In other words, mock jurors were more likely to agree that the defense created reasonable doubt about the accused's guilt than they were to agree that the prosecution case was convincing.

A two-way ANOVA showed no main effects of victim age and crime severity on prosecution case strength, $F(1, 201) = 2.09$, $p = .149$, $g_u = 0.19$, 95% CI [-0.08, 0.46], and $F(1, 201) = 0.01$, $p = .906$, $g_u = 0.00$, 95% CI [-0.27, 0.27], respectively. However, their interaction was significant, $F(1, 201) = 8.72$, $p = .004$, $\eta_p^2 = .042$, as shown in Figure 2a. Post hoc analyses using a Tukey test showed that mock jurors perceived the prosecution case stronger when the injured victim was 25 years old than 75 years old, $g_u = 0.58$, 95% CI [0.19, 0.98]. No age effect was found in the absence of injury, $g_u = -0.22$, 95% CI [-0.60, 0.16].

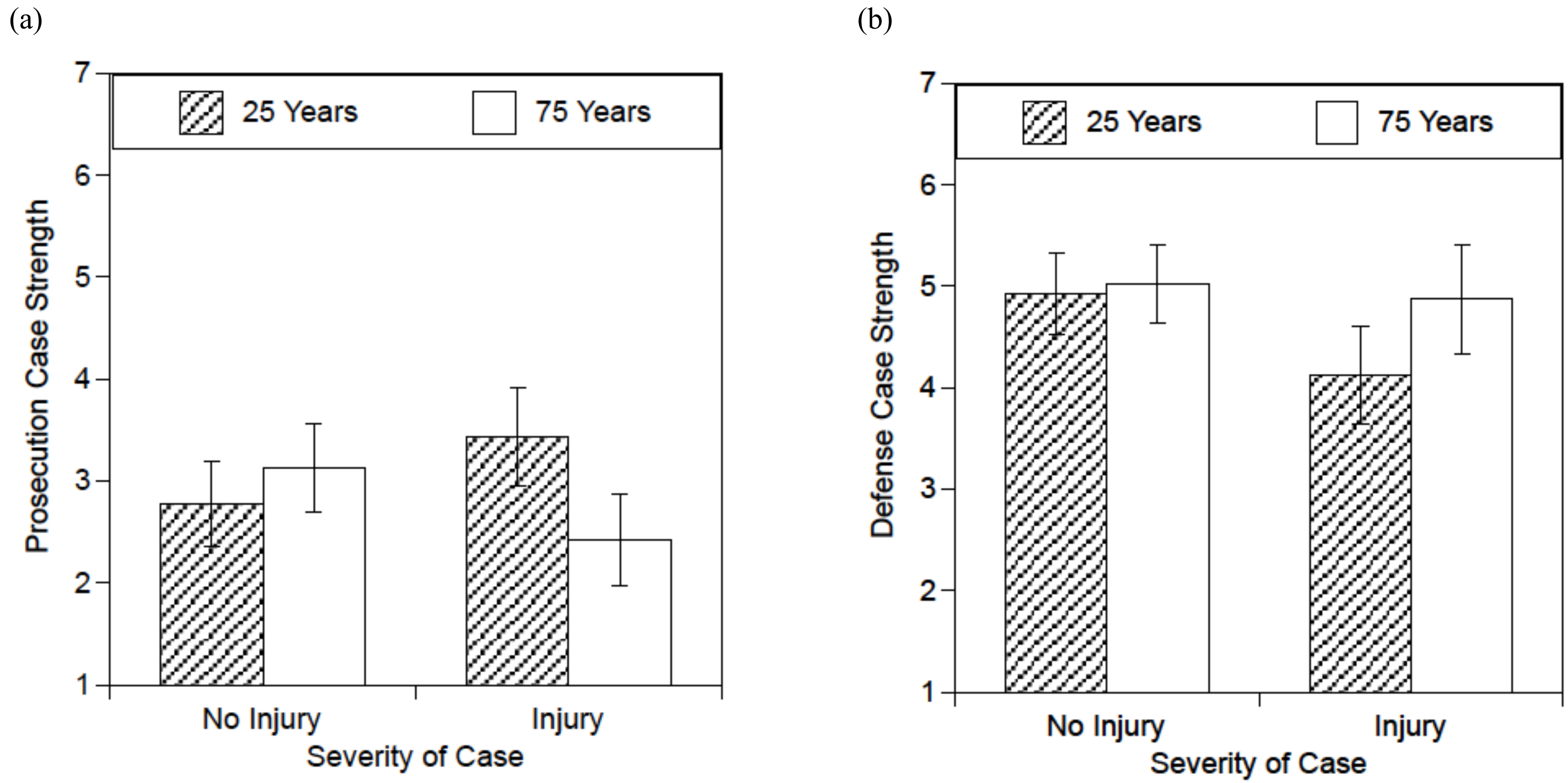


Figure 2. Perceived case strength for (a) the prosecution and (b) the defense. Error bars are 95% confidence intervals.

Further, for the young victim mock jurors perceived the prosecution case stronger in the presence than in the absence of an injury, $g_u = 0.39$, 95% CI [0.00, 0.77]. The effect was reversed for the older victim, $g_u = -0.43$, 95% CI [-0.83, -0.04].

A two-way ANOVA for defense case strength showed a main effect of crime severity, $F(1, 201) = 4.16$, $p = .043$, $\eta_p^2 = .020$. Mock jurors perceived the defense case stronger when the victim was not injured during the robbery, $g_u = 0.30$, 95% CI [0.02, 0.57]. The effect of victim age and the interaction between victim age and crime severity were not significant, $F(1, 201) = 3.33$, $p = .069$, $g_u = -0.26$, 95% CI [-0.53, 0.02], and $F(1, 201) = 1.97$, $p = .162$, $\eta_p^2 = .010$, respectively. Mock jurors' perceptions of the defense case strength are shown in Figure 2b.

Combined Predictive Value of Attitudes, Credibility, and Case Strength

In response to Hypotheses 4a and 4b, hierarchical multiple regression analyses were conducted to assess the combined predictive value of mock jurors' demographics and attitudes (Step 1), perceived victim credibility (Step 2), and perceptions of case strength (Step 3) on (1) the perceived likelihood of the identification to be correct, and (2) probability of guilt. Victim age and crime severity were excluded from the regression models, as they were not associated with the dependent measures. (ANOVAs revealed that victim age, crime severity, and their interaction were not associated with the perceived likelihood of a correct identification, $p > .10$, or with probability of guilt, $p > .10$.)

Table 2 shows correlations between mock jurors' pretrial attitudes, case evaluations and decisions. Pretrial Juror Bias was positively correlated with the likelihood that the identification was correct, $r(203) = .26$, $p < .001$, the probability of guilt , $r(203) = .20$, $p = .002$, prosecution case strength, $r(203) = .29$, $p <$ 001, and Victim Believability, $r(203) = .15$, $p = .017$, and negatively correlated with defense case strength, $r(203) = -.17$, $p = .007$. In other words, the higher mock jurors' confidence in the justice system and proneness to convict, the more likely they were to believe the victim and decide in favor of the prosecution. Social Justice was not correlated with any measure, all rs(203) < .10, thus was excluded from regression analyses.

Table 2

Correlations between independent and dependent variables

	M	*SD*	Victim Believability	Victim Lenient Response Bias	Prosecution Case Strength	Defense Case Strength	Likelihood of Correct Identification	Probability of Guilt	Verdict
Victim age: 25 vs. 75 years	—	—	-.08	.06	-.09	.13	-.10	-.09	-.09
Crime severity: not injured vs. injured	—	—	-.01	-.14	-.00	-.15	.03	.10	.09
Gender: male (1) vs. female (2)	—	—	.21	-.10	.26	-.12	.26	.26	.12
Age	31.34	11.01	.32	-.04	.15	-.06	.17	.18	.06
Pretrial Juror Bias	3.66	0.98	.15	-.02	.29	-.17	.26	.20	.17
Social Justice	4.41	0.97	.01	.02	.03	.06	.06	.07	.09
Beliefs about Memory Changes in Old Age	4.58	0.87	-.27	.15	-.24	.06	-.24	-.21	-.06
M	—	—	4.55	4.07	4.73	2.96	58.44	53.56	—
SD	—	—	1.01	1.37	1.67	1.67	19.94	20.87	—

Note. Point-biserial correlations r_{pb} are reported for victim age, crime severity, gender, and verdict. All other correlations are Pearson's *r*.

Absolute values of any $r \geq .12$ are significant at $p < .05$; $r \geq .16$ are significant at $p < .01$, $r \geq .23$ are significant at $p < .001$.

The subscale Stereotype of Capable-Incapable was negatively correlated with the perceived Victim Believability, $r(203) = -.27$, $p < .001$, and positively correlated with the perceived Victim Lenient Response Bias, $r(203) = .15$, $p = .014$. Further, the subscale Stereotype of Capable-Incapable was negatively correlated with the likelihood that the identification was considered correct, $r(203) = -.24$, $p < .001$, the probability of guilt, $r(203) = -.21$, $p = .001$, and case strength for the prosecution, $r(203) = -.24$, $p < .001$. That is, an increased degree of beliefs that memory changes in older age was associated with decisions in favor of the defense.

Participants' age was not correlated with Pretrial Juror Bias, $r(202) = .07$, $p = .322$, while it was negatively correlated with the subscale Stereotype of Capable-Incapable, $r(202) = -.37$, $p < .001$, indicating that Ageism stereotypes decreased with increasing age.

Likelihood of a correct identification.

Table 3 shows hierarchical regression results for the likelihood that the identification was considered correct. The model and model changes were significant at each step, indicating a significant contribution to the model when variables were added to the model.

At Step 1, participants' gender but not age was associated with the likelihood of a correct identification. Particularly, women were more likely to perceive that the identification was correct than were men, $\beta = .15$, $p = .033$, $sr = 14$. Further, the greater Pretrial Juror Bias, $\beta = .25$, $p < .001$, $sr = .24$, and the lower the Stereotypes of Capable-Incapable, $\beta = -.21$, $p < .001$, $sr = -.19$, the higher was the perceived likelihood that the identification was correct. The model explained 16.0% of the variance, $F(4, 199) = 9.45$, $p < .001$.

When victim credibility was added to the model at Step 2, Pretrial Juror Bias was the only predictor from Step 1 to remain significant, $\beta = .19$, $p = .001$, $sr = .18$. Victim Believability had the highest predictive value in the model, $\beta = .41$, $p < .001$, $sr = .37$. In other words, the higher the perceived Victim Believability, the more likely were mock jurors to assume that the victim made a correct identification. Further, the higher the perceived Victim Lenient Response Bias, the less likely they were to agree that the identification was correct, $\beta = -.21$, $p < .001$, $sr = -.20$. The model at Step 2 was significant, $F(6, 197) = 20.33$, $p < .001$, and explained 38.2% of the variance.

Table 3

Hierarchical Multiple Regression for the Likelihood of a Correct Identification

Predictors	*B*	*Beta*	*t*	*p*	*sr*	R^2
Step 1						.160
Participant Gender	6.05	.15	2.15	.033	.14	
Participant Age	0.07	.04	0.55	.580	.04	
Pretrial Juror Bias	5.03	.25	3.69	<.001	.24	
Beliefs about Memory Changes in Old Age	-4.71	-.21	-2.86	.005	-.19	
Step 2						.382
Participant Gender	4.07	.10	1.67	.097	.09	
Participant Age	-0.08	-.05	-0.73	.469	-.04	
Pretrial Juror Bias	3.82	.19	3.23	.001	.18	
Beliefs about Memory Changes in Old Age	-2.38	-.10	-1.64	.102	-.09	
Victim Believability	8.09	.41	6.51	<.001	.37	
Victim Response Bias	-3.07	-.21	-3.59	<.001	-.20	
Step 3						.499
Participant Gender	2.40	.06	1.08	.282	.06	
Participant Age	-0.02	-.01	-0.20	.846	-.01	
Pretrial Juror Bias	1.83	.09	1.64	.102	.08	
Beliefs about Memory Changes in Old Age	-1.32	-.06	-1.00	.320	-.05	
Victim Believability	4.02	.20	3.13	.002	.16	
Victim Response Bias	-2.21	-.15	-2.79	.006	-.14	
Prosecution Case Strength	4.83	.40	5.98	<.001	.30	
Defence Case Strength	-0.99	-.08	-1.49	.139	-.08	
Regression Constant	35.66		3.819	.002		

Note. Model at Step 1: $F(4, 199) = 9.45$, $p < .001$; Model at Step 2: $F(6, 197) = 20.33$, $p < .001$; Model at Step 3: $F(8, 195) = 24.25$, $p < .001$.

The full model was also statistically significant, $F(6, 198) = 32.42$, $p < .001$, and explained 49.9% of the variance. Mock jurors' perception of the prosecution case strength had the largest predictive value regarding the likelihood that the identification was correct, $\beta = .40$, $p < .001$, $sr = .30$, while perceived defense strength was not associated with the perceived likelihood of a correct identification, $\beta = -.08$, $p = .139$, $sr = -.08$. Perceived Victim Believability and the perceived Victim Lenient Response Bias remained significant at Step 3, as shown in Table 3.

Probability of guilt.

The regression model change was significant at each step, indicating that each pair of variables added a significant contribution to the respective model. The hierarchical regression results for probability of guilt are presented in Table 4.

At Step 1, participant gender, Pretrial Juror Bias, and Stereotypes of Capable-Incapable were associated with the probability of guilt. Female participants and those with stronger Pretrial Juror Biases assigned higher probability of guilt, $\beta = .17$, $p = .017$, $sr = .16$, and $\beta = .19$, $p = .007$, $sr = .18$. Participants who were more likely to agree that memory changes with increasing age assigned lower probability of guilt, $\beta = -.16$, $p = .034$, $sr = -.14$. The model explained 13.0% of the variance in probability of guilt, $F(4, 199) = 7.41$, $p < .001$.

At Step 2, perceived Victim Believability and Victim Lenient Response Bias were both associated with the probability of guilt, $F(6, 197) = 21.88$, $p < .001$, explaining 40.0% of the variance. The more mock jurors believed the victim, the higher was the probability of guilt, $\beta = .42$, $p < .001$, $sr = .37$. Further, the more mock jurors believed that the victim had a more Lenient Response Bias, the lower was the perceived probability of guilt, $\beta = -.27$, $p < .001$, $sr = -.26$.

Table 4

Hierarchical Multiple Regression for Probability of Guilt

Predictors	*B*	*Beta*	*t*	*p*	*sr*	R^2
Step 1						.130
Participant Gender	7.21	.17	2.41	.017	.16	
Participant Age	0.13	.07	0.97	.333	.06	
Pretrial Juror Bias	3.94	.19	2.72	.007	.18	
Beliefs about Memory Changes in Old Age	-3.74	-.16	-2.14	.034	-.14	
Step 2						.400
Participant Gender	4.94	.12	1.97	.050	.11	
Participant Age	-0.03	-.02	-0.24	.808	-.01	
Pretrial Juror Bias	2.62	.12	2.15	.033	.12	
Beliefs about Memory Changes in Old Age	-1.04	-.04	-0.07	.486	-.04	
Victim Believability	8.64	.42	6.57	<.001	.37	
Victim Response Bias	-4.13	-.27	-4.70	<.001	-.26	
Step 3						.538
Participant Gender	3.05	.07	1.37	.173	.07	
Participant Age	0.04	.02	0.40	.689	.02	
Pretrial Juror Bias	0.35	.02	0.32	.753	.02	
Beliefs about Memory Changes in Old Age	0.14	.01	0.10	.919	.01	
Victim Believability	4.05	.20	3.15	.002	.15	
Victim Response Bias	-3.12	-.21	-2.94	<.001	-.19	
Prosecution Case Strength	5.36	.43	6.63	<.001	.32	
Defence Case Strength	-1.35	-.11	-2.02	.045	-.10	
Regression constant	30.37		2.71	.007		

Note. Model at Step 1: $F(4, 199) = 7.41$, $p < .001$; Model at Step 2: $F(6, 197) = 21.88$, $p < .001$; Model at Step 3: $F(8, 195) = 28.39$, $p < .001$.

The full regression model with all eight predictors was significant, $F(8, 195) = 28.39$, $p < .001$, and explained 53.8% of the variance. The stronger the perceived prosecution case the higher was the perceived probability of guilt, $\beta = .43$, $p < .001$, $sr = .32$. Further, the stronger the perception that the defense created reasonable doubt the lower the probability of guilt, $\beta = -.11$, $p = .045$, $sr = -.10$. Notably, mock jurors' gender, Pretrial Juror Biases and Stereotypes of Capable-Incapable did no longer have any predictive value of probability of guilt in the full regression model, $p > .10$, as shown in Table 4. Inspection of the semi-partial correlations indicated that mock jurors perception of prosecution strength had the greatest predictive value of probability of guilt, $sr = .32$, followed by perceived Victim Lenient Response Bias, $sr = -.19$, and Victim Believability, $sr = .15$.

Verdict

The vast majority of mock jurors acquitted the defendant (81.5% guilty, 18.5% not guilty). The conviction rate was irrespective of age of the victim, χ^2 (1, 205) = 1.79, $p = .181$, *phi* = -.093, *OR* = 1.63, 95% CI [0.79, 3.34], or crime severity, χ^2 (1, 205) = 1.55, $p = .213$, *phi* = .087, *OR* = 1.57, 95% CI [0.77, 3.20]. Table 2 shows the proportions of guilty verdicts by victim age and crime severity.

Correlational analyses revealed that with increased Pretrial Juror Bias mock jurors were more likely to convict the defendant, $r_{pb} = .18$ (see Table 1). Further, verdict was positively correlated with perceived Victim Believability, $r_{pb} = .31$, and negatively correlated with perceived Victim Lenient Response Bias, $r_{pb} = -.23$. Finally, verdict was positively correlated with the perceived likelihood that the identification was correct, $r_{pb} = .46$, the probability of guilt, $r_{pb} = .50$, and the prosecution case strength, $r_{pb} = .52$. The cell sizes were too small for guilty verdicts ($n < 10$) to conduct further inferential statistical analyses.

Discussion

Age-related decline in the sensory system (Schieber, 2006), cognitive functioning and memory (Coleman & O'Hanlon, 2008; Park et al., 2002) have been well documented. Yet studies showed conflicting results about how jurors perceived older people compared to young people in regards to their credibility and reliability as witnesses (e.g., Kwong See et al., 2001; Neal et al., 2012; Ross et al., 1990). The differences may be attributed to characteristics other than victim age alone, such as crime type or juror characteristics, as proposed by the conceptual model of juror decision-making that linked defendant, case, and juror characteristics to their verdict (Devine & Caughlin, 2014). This model implies that both legal and extra-legal factors drive the decision process (Sporer & Goodman-Delahunty, 2009). The present study aimed to investigate the effect of manipulated victim age and crime severity as legal factors and juror characteristics as extra-legal factors on mock jurors' decisions. First, we examined whether mock jurors were aware of age-related memory changes and whether they believed more the testimony of the young than the older victim. Second, we investigated whether crime severity influenced mock jurors' decisions. Third, we investigated the extent to which mock juror characteristics were associated with their decisions.

Unlike many other studies that only measured verdict (Lundrigan, Dhami, & Müller-Johnson, 2018) the present study investigated mock jurors' decision process from different perspectives. Perceptions of the victim were assessed using questions about credibility in the sense of deception *and* in the sense of error (Sporer, 2008). Analyses showed that these two constructs were not independent of each other, as reflected in the subscale Victim Believability that included items about both honesty *and* correct memory of the victim. The subscale Victim Lenient Response Bias assessed the perceived victim's proneness to choose anyone from the lineup, which was moderately correlated with the subscale Victim Believability (r = -.24). Finally, to gain insight into different aspects of the decision process beyond the binary variable verdict, the present study included jurors' perceptions of the likelihood that the victim correctly identified the offender and of the probability of guilt. Both measures showed a strong correlation with the binary verdict, r_{pb} = .46, and r_{pb} = .50, respectively.

Mock Jurors' Awareness of Memory Changes in Old Age

Mock jurors' awareness of memory changes in old age was tested via their beliefs about cognitive capacity in older age and by manipulating victim age. Analyses showed that mock jurors expressed a moderate degree of agreement to memory changes in older age (i.e., mean values above the midpoint of the scale), when they were asked about their beliefs about cognitive capacity in old age. These findings suggest that participants tended to associate old age with some losses in cognitive capacity, confirming previous research findings (Braithwaite et al., 1993). This is discussed in more detail further below.

The manipulation of victim age was successful only to some extent. Compared to a young victim, mock jurors perceived the older victim as neither more credible (cf. Allison et al., 2006; Ross et al., 1990) nor did they doubt their testimony more than that of the young victim (cf. Neal et al., 2012). Further, no age effect was found for the perceived Victim Lenient Response Bias, indicating that mock jurors were unaware that older people may have a more liberal response criterion than young people (see the meta-analysis by Martschuk & Sporer, 2018). However, when crime severity was taken into account, an age effect was found for the severe crime condition such that mock jurors believed more the young than the older victim. Consequently, compared to the young victim mock jurors perceived the prosecution case weaker and the defense case stronger in the severe case involving an older victim. Thus, although both victims delivered the same testimony, mock jurors seemed to be not as persuaded by the older prosecution witness as they were by the young prosecution witness. Still, these perceptions were not reflected in the perceived likelihood that the identification was correct, probability of guilt, or verdict.

Further, no age effect was found for perceptions of the strength of the prosecution, or the defense case, failing to support Hypotheses 1b and 1c. These unexpected findings may be explained by the mode of trial presentation, namely identical trial summaries without any visual cues of ageing or differences in speech styles consistent with ageing (Pittman et al., 2014). Some support for this explanation was provided in a series of experiments by Brimacombe et al. (1997). They showed that older witnesses provided weaker testimonies than young witnesses (i.e., less accurate, more frequent use of negative qualifiers, higher expressed hesitance). As a consequence, mock jurors perceived testimonies of older witnesses as less credible, irrespective of whether they knew the age of the witness or not. Thus, speech style influenced perceived credibility more than the actual witness age.

Influence of Crime Severity on Mock Jurors' Decisions

Crime severity in the form of victim injury was the second case characteristic manipulated in the present study. Although injury should not be considered when evaluating the strength of the evidence, if crime is committed in a brutal manner, mock jurors interpret the evidence in a biased way by putting more weight to the emotionally laden information (Bright & Goodman-Delahunty, 2004). Following, we hypothesized that mock jurors would believe the victim more (Hypothesis 2a) and perceive the prosecution case stronger when the crime was more severe (Hypothesis 2b). The present findings confirmed this theoretical notion for the young victim only: Mock jurors perceived the prosecution case of a more severe crime stronger than of a less severe crime. However, the effect was reversed for the older victim such that mock jurors perceived the prosecution case weaker when the crime was severe compared to a less severe crime. These findings indicate that mock jurors might have seen the injury as a further proof that the crime happened for the young victim, while they had doubts about the case when an older victim was injured.

The present study further showed that crime severity influenced perceptions of the defense case: Mock jurors were more likely to agree that the defense created reasonable doubt of the defendant's guilt in the less severe case, confirming Hypothesis 2c. Put another way, victim injury weakened the defense case—a finding that was previously shown in cases of sexual abuse and child sexual abuse (e.g., Blackwell & Seymour, 2014; Fitzgerald, 2006; Gray-Eurom, Seaberg, & Wears, 2002).

Finally, the case of a more severe crime may evoke an emotional reaction, which in turn would result in more guilty verdicts. A number of studies demonstrated that crime severity was associated with more guilty verdicts (e.g., Bright & Goodman-Delahunty, 2004; Goodman-Delahunty et al., 2016; Pearson et al., 2018) and more punitive sentencing decisions (Nadler & Rose, 2003; Sporer & Goodman-Delahunty, 2009). The present results revealed very low conviction rates (despite a perceived probability of guilt above 50%), independent of crime severity. These findings were in line with a study that used a similar trial scenario, namely a robbery trial with varying degrees of violence against the victim (Freedman et al., 1994). Taken together, more research is needed to further investigate whether and how crime severity influences perceptions of a victim, particularly of an older age, and ultimately verdict.

Juror Attitudes and Their Decisions

The extra-legal factors investigated in the present study were juror characteristics (i.e., age, gender), their trust in the legal system and perceptions of older people. Analyses revealed that participant gender but not age were associated with the perceived likelihood of a correct identification and probability of guilt. These findings were in line with previous research showing that female mock jurors were more likely to render guilt judgments than men (see the meta-analysis by Devine & Caughlin, 2014).

Further, the present findings showed that, without any case-specific information, mock jurors' evaluation of the evidence (i.e., the likelihood that the identification is correct) and their ratings of the probability of guilt were associated with their Pretrial Juror Bias (i.e., System Confidence and Conviction Proneness) and with their beliefs about ageing (Stereotype of Capable-Incapable). Specifically, the more confident mock jurors were in the legal system and the more prone they were to convict the more likely they were to decide in favour of the prosecution. These findings support previous research findings that jurors' trust in the legal system was associated with guilt judgments (Devine & Caughlin, 2014; Lecci & Myers, 2009).

Finally, mock jurors holding stronger beliefs about memory changes in old age were more sceptical of the victim's testimony regardless of victim age, identification accuracy and probability of guilt. In all, they were more skeptical of the prosecution case and less likely to render guilt judgments. These findings also indicated that mock jurors holding stronger beliefs about memory changes in old age were in general more skeptical of memory performance of the victim. A large body of research showed that juror characteristics are only moderately correlated with their decisions (see review by Devine, 2012) and explained only about 10-15% of the variance in their decision, while the presented evidence had a stronger impact on juror decisions (Saks & Spellman, 2016). For instance, a study that tested juror decisions across multiples types of trials showed that perceived case strength was associated with higher probability of guilt (Pearson et al., 2018). The present results showed the same pattern: Mock juror characteristics explained together 13% of the variance in probability of guilt and 16% of the variance in the likelihood of a correct identification. Once case-specific facts were considered, influence of juror characteristics and general views became less important. Instead, mock jurors' perceptions of the victim and their evaluation of the prosecution case were much more predictive of the perceived likelihood that the identification was correct and the probability of guilt than juror characteristics, as assumed in Hypotheses 4a and 4b. These findings lend further support to Sherif and Sherif's (1969) Gestalt theoretical postulate that internal factors affect judgments only when the stimulus configuration is rather ambiguous (see also Sporer & Goodman-Delahunty, 2009).

Unfortunately, the cell sizes for guilty verdicts were too small to conduct logistic regression analyses to test Hypothesis 4c that after controlling for juror characteristics perceptions of the prosecution and the defense case would be associated with verdict. Nevertheless, correlational analyses confirmed that perceptions of the victim and of the prosecution case strength were associated with verdict. A reason for the low conviction rate may be that neither witnesses other than the victim nor forensic evidence were presented supporting the prosecution case—factors that influence perceptions of case strength and ultimately guilty judgments, as shown in a study by Pearson et al. (2018). Following, mock jurors applied the principle of in dubio pro reo in this case and acquitted the defendant.

Limitations

The following methodological issues must be considered when drawing conclusions from the findings of the present study. The mode of trial presentation was a written (simulated) trial summary as opposed to a live evidence presentation at court. Mock jurors were aware that their decision did not have any consequences for the defendant or the community. This may have limited the extent to which mock juror decisions resemble those in real cases. Further, mock jurors made their decision individually and not in groups, which limits the ecological validity of decisions due to the lack of mutual influence of other jury members and discussions that arise during jury deliberations. Notably, a previous study showed similarities in the decision processes between individual and group decisions in the control group (i.e., in absence of educative information provided by an expert; Goodman-Delahunty & Martschuk, 2017).

Finally, participants were jury-eligible citizens and not actual jurors. Compared to empaneled jurors, participants in the present study were more educated (i.e., a higher proportion of participants held a university degree) and the proportion of female participants was higher (cf. Goodman-Delahunty et al., 2007; Trimboli, 2008). However, explicit attempts were made to include potential jurors; for instance, only participants from countries in which jury trials are held were eligible. Moreover, more than 20% of participants reported that they had been previously called to jury duty and one-third of them reported they had served in a jury trial.

Implications for Research and Practice

The fact that victim age did not influence mock jurors' decisions in the present study indicated that mock jurors were unaware of memory changes in old age. Notably, the findings that mock jurors were skeptical of memory in older age, and were also more skeptical of victim testimony in general, suggested that mock jurors applied their knowledge about memory (or lack thereof) when making decisions. Following this, in cases in which memory of an older witness might be an issue, mock jurors would benefit from educative information about a memory decline in older age. Similar proposals involving witness and victim memory have already been made (e.g., Conway, 2012; Doyle, 2012), and professional psychological associations (e.g., The British Psychological Society Research Board, 2010; Wells et al., 2018) prepared empirical guidelines to inform legal decision makers about different aspects of memory. In jury trials, educative information can be provided in the form of judicial instructions or expert evidence. However, more research is needed to assess which mode of presentation is more effective to inform decision makers about memory.

Further research is needed to test whether the present findings hold following a simulated video trial and group deliberations. Analyzing jury deliberations will assist to assess the reasoning process and the extent to which, for instance, victim age is discussed as part of evidence.

Conclusion

In summation, our findings demonstrated the importance of including moderators, such as age, when evaluating juror decisions. The theoretical notion that mock jurors evaluated the case more in favor of the prosecution when the crime was more severe was only confirmed for the young victim, while the effect was reversed for the older victim. Further, the data showed that mock jurors were only moderately aware of age-related changes in memory. While mock juror characteristics were associated with their decisions to some extent, case-specific facts were much more predictive of the evaluation of evidence and judgments of the defendant’s guilt.

References

Adorno, T. W., Levinson, D. J., Frenkel-Brunswik, E., & Sanford, R. N. (1950). *The authoritarian personality.* New York: Harper & Row.

Aizpura, A., Migueles, M., & Garcia-Bajos, E. (2014). Accuracy of eyewitness memory for events in young and older adults. In M. P. Toglia, D. F. Ross, J. Pozzulo, & E. Pica (Eds.), *The elderly eyewitness in court* (pp. 210–231). New York: Psychology Press.

Allison, M., Brimacombe, C. A., Hunter, M. A., & Kadlec, H. (2006). Young and older eyewitnesses' use of narrative features in testimony. *Discourse Processes, 41,* 289–314. doi:10.1207/s15326950dp4103_3

Andrewartha, A. M. (2014). *An analysis of the jury's role in detecting lies, assessing credibility, and decision making*. Thesis submitted for the Degree of Doctor of Philosophy in Monash University, Melbourne, Australia

Australian Law Reform Commission. (1985). *Evidence* (Vol. 26). Canberra: Australian Government Publishing Service.

Blackwell, S., & Seymour, F. (2014). Prediction of jury verdicts in child sexual assault trials. *Psychiatry, Psychology and Law, 21,* 567–576. doi:10.1080/13218719.2013.856278

Bornstein, B. H., & Greene, E. (2011). Jury decision making: Implications for and from psychology. *Current Directions in Psychological Science, 20,* 63–67. doi:10.1177/0963721410397282

Braithwaite, V., Lynd-Stevenson, R., & Pigram, D. (1993). An empirical study of ageism: From polemics to scientific utility. *Australian Psychologist, 28,* 9–15. doi:10.1080/00050069308258857

Bright, D. A., & Goodman-Delahunty, J. (2004). The influence of gruesome verbal evidence on mock juror verdicts. *Psychiatry, Psychology and Law, 11,* 154–166, doi:10.1375/pplt.2004.11.1.154

Brimacombe, C. A., Jung, S., Garrioch, L., & Allison, M. (2003). Perceptions of older adult eyewitnesses: Will you believe me when I'm 64? *Law and Human Behavior, 27,* 507–

522. doi:10.1023/A:1025486006769

Brimacombe, C. A., Quinton, N., Nance, N., & Garrioch, L. (1997). Is age irrelevant? Perceptions of young and old adult eyewitnesses. *Law and Human Behavior, 21,* 619–634. doi:10.1023/A:1024808730667

The British Psychological Society Research Board. (2010). *Guidelines on memory and the law: Recommendations from the scientific study of human memory*. Leicester, UK: The British Psychological Society.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Coleman, P. G., & O'Hanlon, A. (2008). Ageing and adaptation. In R. Woods, & L. Clare (Eds.), *Clinical psychology of ageing* (pp. 17–32). West Sussex, England: John Wiley & Sons.

Conway, M. A. (2012). Ten things the law and others should know about human memory. In L. Nadel & W. P. Sinnott-Armstrong (eds.), *Memory and law* (pp. 359–372). Oxford University Press.

Devine, D. J. (2012). *Jury decision making: The state of the science.* New York: New York University Press.

Devine, D. J., & Caughlin, D. E. (2014). Do they matter? A meta-analytic investigation of individual characteristics and guilt judgments. *Psychology, Public Policy, and Law, 20,* 109–134. doi:10.1037/law0000006

Dodson, C. S., & Krueger, L. E. (2006). I misremember it well: Why older adults are unreliable eyewitnesses. *Psychological Bulletin and Review*, *13*, 770–775. doi:10.3758/BF03193995

Doyle, J. M. (2012). Ready for the psychologists: Learning from eyewitness errors. *Court Review: The Journal of the American Judges Association, 48,* 4–12.

Fitzgerald, J. (2006). The attrition of sexual offences from the New South Wales criminal justice system. *Contemporary Issues in Crime and Justice* (Vol. 92). Sydney, Australia: NSW Bureau of Crime Statistics and Research.

Freedman, J. L., Krismer, K., MacDonald, J. E., & Cunningham, J. A. (1994). Severity of penalty, seriousness of the charge, and mock jurors' verdicts. *Law and Human Behavior, 18,* 189–202. doi:10.1007/BF01499015

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Goodman-Delahunty, J., Brewer, N., Clough, J., Horan, J., Ogloff, J., Tait, D., & Pratley, J. (2007). Practices, policies and procedures that influence juror satisfaction in Australia. *Research and Public Policy Series, 87.* Australian Institute of Criminology.

Goodman-Delahunty, J., & Martschuk, N. (2017). Using pretest-posttest research designs to enhance jury decisions. *Newcastle Law Review, 12,* 86–103.

Goodman-Delahunty, J., Martschuk, N., & Ockenden, E. (2016). Effects of terrorist charges and threatening conduct on mock jurors' decisions. *Psychiatry, Psychology and Law, 23,* 696–708. doi:10.1080/13218719.2015.1120247

Grady, R. H., Reiser, L., Garcia, R. J., Koeu, C., & Scurich, N. (2018). Impact of gruesome photographic evidence on legal decisions: A meta-analysis. *Psychiatry, Psychology and Law, 25,* 503–521. doi:10.1080/13218719.2018.1440468

Gray-Eurom, K., Seaberg, D. C., & Wears, R. L. (2002). The prosecution of sexual assault cases: Correlation with forensic evidence. *Annals of Emergency Medicine, 39,* 39–46. doi:10.1067/mem.2002.118013

James, M., & Graycar, A. (2000). *Preventing crime against older Australians.* Canberra, ACT: Australian Institute of Criminology Research and Public Policy Series 32.

Judicial Commission of New South Wales. (2016). *Criminal Trial Courts Benchbook: Trial procedure.* Retrieved from https://www.judcom.nsw.gov.au/publications/benchbks/criminal/the_jury.html

Kerr, N. L. (1978). Severity of prescribed penalty and mock jurors' verdicts. *Journal of Personality and Social Psychology, 36,* 1431–1442. doi:10.1037/0022-3514.36.12.1431

Kite, M. E., Stockdale, G., Whitley, B., & Johnson, B. (2005). Attitudes towards younger and older adults: An updated meta-analytic review. *Journal of Social Issues, 61,* 241–266. doi:10.1111/j.1540-4560.2005.00404.x

Kite, M. J., & Wagner, L. S. (2002). Attitudes towards older adults. In Nelson, T. D. (Ed.), *Ageism: Stereotyping and prejudice against older persons.* Cambridge, Massachusetts: The MIT Press.

Kwong See, S. T., Hoffman, H. G., & Wood, T. (2001). Perceptions of an elderly eyewitness: Is the older eyewitness believable? *Psychology and Aging, 16,* 346–350. doi:10.1037/0882-7974.16.2.346

Lecci, L., & Myers, B. (2008). Individual differences in attitudes relevant to juror decision making: Development and validation of the Pretrial Juror Attitude Questionnaire (PJAQ). *Journal of Applied Social Psychology, 38,* 2010–2038. doi:10.1111/j.1559-1816.2008.00378.x

Lecci, L., & Myers, B. (2009). Predicting guilt judgments and verdict change using a measure of pretrial bias in a videotaped mock trial with deliberating jurors. *Psychology, Crime, & Law, 15,* 619–634. doi:10.1080/10683160802477757

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Lundrigan, S., Dhami, M. K., & Müller-Johnson, K. (2018). A re-examination of the acquittal biasing effect of offence seriousness. *Psychiatry, Psychology and Law, 25,* 769–778. doi:10.1080/13218719.2018.1478334

Martschuk, N., & Sporer, S. L. (2018). Memory of faces in old age: A meta-analysis. *Psychology & Aging, 33,* 904–923. doi:10.1037/pag0000282

Martschuk, N., Sporer, S. L., & Sauerland, M. (2019). Confidence of older eyewitnesses: Is it diagnostic of identification accuracy? *Open Psychology, 1*, 132–151. doi:10.1515/psych-2018-0010

McComas, W. C., & Noll, M. E. (1974). Effects of seriousness of charge and punishment the judgments of simulated jurors. *The Psychological Record, 24,* 545–547. doi:10.1007/BF03394276

Nadler, J., & Rose, M. R. (2003). Victim impact testimony and the psychology of punishment. *Cornell Law Review, 88,* 419–456. doi:10.2139/ssrn.377521

Neal, T. M. S., Christiansen, A., Bornstein, B. H., & Robicheaux, T. R. (2012). The effects of mock jurors' beliefs about eyewitness performance on trial judgments. *Psychology,* *Crime, & Law, 18,* 49–64. doi:10.1080/1068316X.2011.587815

Nunez, N., McCoy, M. L., Clark, H. L., & Shaw, L. A. (1999). The testimony of elderly victim/witnesses and their impact on juror decisions: The importance of examining multiple stereotypes. *Law and Human Behavior, 23,* 413–423. doi:10.1023/A:1022308014652

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging, 17,* 299–320. doi:10.1037//0882-7974.17.2.299

Pearson, J. M., Law, J. R., Skene, J. A. G., Beskind, D. H., Vidmar, N., Ball, D. A., … Skene, J. H. P. (2018). Modelling the effects of crime type and evidence on judgments about guilt. *Nature Human Behaviour, 2,* 856–866. doi:10.1038/s41562-018-0451-z

Pittman, A. E., Toglia, M. P., Leone, C. T., & Mueller-Johnson, A. U. (2014). Testimony by the elderly in the eyes of the jury: The impact of juror characteristics. In M. P. Toglia, D. F. Ross, J. Pozzulo & E. Pica (Eds.), *The elderly eyewitness in court* (pp. 363–382). New York: Psychology Press.

Queensland Courts. (2016). *Supreme and District Courts Benchbook: 24 general summing up directions.* Retrieved from http://www.courts.qld.gov.au/__data/assets/pdf_file/0018/86022/sd-bb-24-summing-up-general.pdf

Ross, D. F., Dunning, D., Toglia, M. P., & Ceci, S. (1990). The child in the eyes of the jury: Assessing mock jurors' perceptions of the child witness. *Law and Human Behavior, 14,* 5–23. doi:10.1007/BF01055786

Saks, M., J., & Spellman, B. A. (2016). *The psychological foundations of evidence law.* New York and London: New York University Press.

Schieber, F. (2006). Vision and aging. In J. E. Birren, & K. W. Schaie (Eds.), *The handbook of the psychology of aging* (pp. 129–161). Burlington, USA: Elvesier Academic Press.

Sherif, M., & Sherif, C. W. (1969). *Social psychology.* New York: Harper & Row.

Sporer, S. L. (2008). Lessons from the origins of eyewitness testimony research in Europe. *Applied Cognitive Psychology, 22*, 737–757. doi:10.1002/acp.1479

Sporer, S. L., & Goodman-Delahunty, J. (2009). Disparities in sentencing decisions. In M. E. Oswald, S. Bieneck, & J. Hupfeld-Heinemann (Eds.), *Social psychology of punishment of crime* (pp. 379–401). New York, NY: John Wiley & Sons Ltd.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Trimboli, L. (2008). Juror understanding of judicial instructions in criminal trials. *Contemporary Issues in Crime and Justice* (Vol. 119). Sydney: NSW Bureau of Crime Statistics and Research.

Walker, S. (2018). *Crimes against older people in Victoria.* Melbourne, VIC: Crime Statistics Agency. Retrieved from https://www.crimestatistics.vic.gov.au/research-and-evaluation/publications/victims/crimes-against-older-people-in-victoria

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C., & Wixted, J. T. (2018). *Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence.* Report to the American Psychology/Law Society and Division 41 of the American Psychological Association.

Appendix A. Trial scenario

General instructions

Suppose you have been selected as a member of the jury in the case of R v Mitchell. Please read the following case summary. You will be asked your verdict in a series of questions at the end.

The complainant in this case is Ms Megan Riley, a 25- [75-] year-old woman. She alleges that, on the evening of 15th April 2017, the defendant, Mr Blake Mitchell, robbed her in a parking garage. He grabbed her handbag with all her belongings. [He violently took her handbag with all her belongings and injured her.]

Mr Mitchell has pleaded "not guilty" to the charge, after he had declared to the police that he had never seen Ms Riley and that he had never been in that garage.

The prosecution case relies upon the testimony of the complainant, Ms Riley, and the testimony of Detective Sergeant Malcolm Grey, who led the investigation of the case, interviewed Ms Riley and conducted a photo lineup with her.

The Defence argues that the allegations were unsubstantiated and that Ms Riley made a false identification. The defendant Mr Mitchell followed his lawyer's advice not to testify at trial.

<u>The first prosecution witness is Ms Riley, the robbed victim.</u>
On direct examination, Ms Riley testified that she drove into a public parking garage to park her car at about 7pm to attend a meeting. As she got out of the car, a man appearing to be in his late twenties approached her and informed that she parked in a reserved spot. Ms Riley reassured him that she parked in a parking lot reserved for customers, and started to walk away. The man blocked her way saying, "Give me your wallet—and the car keys". As she refused to give her belongings to him, he snatched her handbag away from her and ran away

with her handbag. [As she refused to give her belongings to him, he violently grabbed her arm. He snatched her handbag away from her, pushed her over and ran away with her handbag. Following, she fell and sprained her ankle. As a result she had to attend the hospital where she was given a brace and crutches to support and reduce weight on the injured ankle.] Ms Riley testified that the handbag contained her wallet, her phone, her keys, and some personal items. In the wallet she had her credit card, her driving licence, some membership cards, and approximately $ 80 in cash.

Ms Riley described the offender as a white man in his late twenties. He was wearing a black leather jacket, blue jeans, and white sneakers. She told the court that he had brown eyes and short brown hair, and was about 6.1 inches tall with an athletic build. Further, Ms Riley reported that she identified the offender when she was presented a photo lineup at the police station.

On cross-examination, Ms Riley testified that she could not remember what the offender was wearing under the jacket. The Defence counsel challenged this, because at the police station she told he was wearing a dark shirt. Further, Ms Riley was asked why she reported to the police that the offender was wearing light-coloured sneakers, while at trial she told he was wearing white sneakers. She could not explain this.

The Defence counsel stated that Ms Riley falsely identified Mr Mitchell as the attacker and added, “You would have identified any person that was presented to you, just to have someone to blame”. Ms Riley responded with “This is not true”.

<u>The second prosecution witness was Detective Sergeant Grey.</u>

On direct examination, the Detective Sergeant Grey reported that he had been investigating the case from the beginning. He reported that Ms Riley had come to the police the same evening, told him what happened and gave a detailed description of the attacker. The police received reports that the description of the defendant matched Mr Mitchell and that he was seen in the area in the evening of the 15th April 2017.

The Detective testified that when they arrested Mr Mitchell, he denied any wrongdoing. Following the arrest Ms Riley was called and presented a photo lineup including Mr Mitchell and five foils. The five foils were members of the police and matched the description of the person Ms Riley described. Mr Mitchell was placed as number 3 in the lineup. Ms Riley identified Mr Mitchell as the attacker within one minute, saying, “I am certain that person number 3 was the attacker”.

On cross-examination, the Defence counsel claimed the description Ms Riley gave fit an average man, and it could have been any man in his late twenties. Further, the Defence council outlined that some details of Ms Riley’s description of the robber were conflicting. The Defence council argued that Ms Riley confused Mr Mitchell with another person and asked the detective how likely it was that Ms Riley could have seen Mr Mitchell on the street and mistakenly held him as the person who robbed her. The Detective said that it could be possible, however, in this case all the investigation led to the defendant Mr Mitchell.

Final Instructions

In this trial, the Crown must prove each element of the three counts against the accused. You must be satisfied that the Crown has proved each element, beyond reasonable doubt, before you can convict the accused of any one particular count. This high standard of proof means that the accused does not have to prove any fact or issue that is in dispute. It is not for the accused to prove his innocence but for the Crown to prove his guilt in relation to each offence and to prove it beyond reasonable doubt.

If the Crown fails to prove any part of any of the counts beyond reasonable doubt, the accused is not guilty of that count.

Appendix B. Posttrial questionnaire

Your final decision:		
Do you find beyond reasonable doubt that Mr Mitchell is guilty of robbery?	☐ Yes	☐ No

How likely is it that (tick the appropriate number)…											
Ms Riley made a correct identification?	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Mr Mitchell committed the crime?	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Please indicate your agreement to the following statements.							
Ms Riley…	**Strongly disagree**						**Strongly agree**
accurately remembered the event.	1	2	3	4	5	6	7
identified the correct man in the police lineup.	1	2	3	4	5	6	7
was cautious when she selected somebody from the lineup.	1	2	3	4	5	6	7
wanted to choose someone from the lineup to help the police.	1	2	3	4	5	6	7
wanted to have someone to blame.	1	2	3	4	5	6	7
would have identified anyone from the lineup for justice to be served.	1	2	3	4	5	6	7
gave a consistent account of the event.	1	2	3	4	5	6	7
gave a plausible account of the event.	1	2	3	4	5	6	7
was confident when she identified the robber at the lineup procedure.	1	2	3	4	5	6	7
was confident when she testified at trial.	1	2	3	4	5	6	7
was telling the truth about the event.	1	2	3	4	5	6	7
was credible.	1	2	3	4	5	6	7
was unbiased.	1	2	3	4	5	6	7
was honest.	1	2	3	4	5	6	7
was trustworthy.	1	2	3	4	5	6	7

Please indicate your agreement to the following statements.							
	Strongly disagree						**Strongly agree**
The prosecution case was sufficient to convince me that the accused was guilty.	1	2	3	4	5	6	7
The defence case was sufficient to create reasonable doubt about the guilt of the defendant.	1	2	3	4	5	6	7

Please tick the box with the best response:	
Ms Riley was robbed	
☐	on the street.
☐	in a parking garage.
☐	in a shopping centre.
The robber blocked Ms Riley's way	
☐	with a gun.
☐	with a knife.
☐	none of the above.
The Detective Mr Mitchell testified that	
☐	Ms Riley gave a detailed description of the attacker.
☐	the police received reports that Mr Mitchell was seen in the parking garage.
☐	all of the above.
Ms Riley was	
☐	25 years old.
☐	45 years old.
☐	75 years old.
Ms Riley identified Mr Mitchell in a	
☐	showup.
☐	photo lineup.
☐	live lineup.
Which statement is <u>false</u>? Ms Riley testified the offender was wearing	
☐	blue jeans.
☐	grey sneakers.
☐	a black leather jacket.
The robber took Ms Riley's handbag, incl. her phone, her wallet and	
☐	$40 in cash.
☐	$80 in cash.
☐	$150 in cash.
Was Ms Riley injured during the robbery?	
☐	Yes, she sprained her hand.
☐	Yes, she sprained her ankle.
☐	No, she did not report any injury.

Please indicate your agreement to the following statements.							
	Strongly disagree						**Strongly agree**
A suspect who runs from police, probably committed a crime.	1	2	3	4	5	6	7
A defendant should be found guilty if 11 out of 12 jurors vote guilty.	1	2	3	4	5	6	7
Too often jurors hesitate to convict someone who is guilty out of pure sympathy.	1	2	3	4	5	6	7
Out of every 100 people brought to trial, at least 75 are guilty of the crime with which they are charged.	1	2	3	4	5	6	7
For serious crimes like murder, a defendant should be found guilty so long as there is a 90% chance that he committed the crime.	1	2	3	4	5	6	7
Generally, the police make an arrest only when they are sure about who committed the crime.	1	2	3	4	5	6	7
Extenuating circumstances should not be considered—if a person commits a crime, that person should be punished.	1	2	3	4	5	6	7
If the defendant committed a victimless crime like gambling or possession of marijuana, that person should never be convicted.	1	2	3	4	5	6	7
Criminals should be caught and convicted by “any means necessary.”	1	2	3	4	5	6	7
A prior record of conviction is the best indicator of a person’s guilt in the present case.	1	2	3	4	5	6	7
Wealthy individuals are almost never convicted of their crimes.	1	2	3	4	5	6	7
When it is the suspect’s word against the police officer’s, I believe the police.	1	2	3	4	5	6	7
A Black man on trial with a predominantly White jury will always be found guilty.	1	2	3	4	5	6	7
If a witness refuses to take a lie detector test, it is because he/she is hiding something.	1	2	3	4	5	6	7
Famous people are often considered to be “above the law.”	1	2	3	4	5	6	7

Please indicate your agreement to the following statements.							
	Strongly disagree						**Strongly agree**
Older people are as capable as ever of concentrating on any given task.	1	2	3	4	5	6	7
Older people lose the ability to pay attention to detail.	1	2	3	4	5	6	7
An older person tends to become tired and fatigued very easily.	1	2	3	4	5	6	7
With old age, people are inclined to become forgetful.	1	2	3	4	5	6	7
Older people can pay great attention to detail in many tasks.	1	2	3	4	5	6	7
As we grow older, we become slower in getting things done.	1	2	3	4	5	6	7
Older people are not able to concentrate as well as they could in their youth.	1	2	3	4	5	6	7
Older people are quite capable of performing tasks that require effort and stamina.	1	2	3	4	5	6	7
Older people have no difficulty in getting things done quickly.	1	2	3	4	5	6	7
A person's memory is not adversely affected by increasing age.	1	2	3	4	5	6	7

About yourself:		
Please indicate your gender.	☐ Male ☐ Female	**What is your age?** ________ years
What is your citizenship?	____________________________	
What is your ethnic background? (select all that apply)	☐ Aboriginal or Torres State Islander ☐ American Indian or Alaska Native ☐ Asian	☐ Black or African-American ☐ Hispanic or Latino ☐ White ☐ Other, __________
What is your highest level of education?	☐ Less than high school ☐ High school (12 years) ☐ Trade certificate or equivalent ☐ Diploma or equivalent	☐ Bachelor's degree ☐ Master's degree ☐ Doctoral degree and above ☐ Other, __________
Please indicate your occupation	☐ Unemployed ☐ Retiree or pensioner ☐ Student ☐ Home duties ☐ Casual employee ☐ Self-employed	☐ Labourer or related worker ☐ Tradesperson ☐ Administrative or clerical worker ☐ Professional
Have you ever been called for jury duty?	☐ Yes	☐ No
Have you ever served as a juror on a jury trial?	☐ Yes	☐ No

Appendix C. Component Structure of the Perceived Victim Credibility Scale

All credibility items elicited the full range of possible responses on the 7-point rating scale, indicating a wide range of perceptions of victim credibility. Table C1 presents item means, standard deviations and inter-item correlations.

A principal components analysis (PCA) was conducted on 14 items that measured perceived witness credibility. Initial analyses revealed suitability of data for PCA: (a) Inspection of the correlation matrix revealed that all items were between $r(205) = .33$ and $r(205) = .78$ (see Table C1), and, therefore, between the recommended values of $r = .3$ and $r = .8$ (Field, 2013); (b) the Kaiser-Myer-Olkin measure was KMO = .89 and all KMO values for individual items were greater than KMO = .68 and above the recommended value of .6 (Kaiser, 1974; Kaiser & Rice, 1974); (c) the Bartlett's Test of Sphericity reached statistical significance, Chi-squared (91) = 1347.39, $p < .001$.

Multiple methods were used to assess the number of components. Inspection of the eigenvalues revealed three components that exceeded 1, and explained 41.8%, 13.0%, and 7.6% of the variance, respectively. The screeplot revealed one point of reflection after the second factor, and a second point of reflection after the third factor. The Monte Carlo PCA for Parallel Analysis (Watkins, 2000) with randomly generated data matrix using 1000 replications supported the two-component structure.

Solutions for two and three components were examined using varimax (orthogonal) rotation of the component-loading matrix. The component loadings in Table C2 measured the relationship between the components and the items indicating the overlapping variance among components, which are interpreted as follows: $\lambda \geq .71$: 50% overlapping variance, $\lambda \geq .63$: 40% overlapping variance, $\lambda \geq 0$ .55: 30% overlapping variance, $\lambda \geq 0$ .45: 20% overlapping variance, $\lambda \geq .32$: 10% overlapping variance (Comrey & Lee, 1992; as cited in Tabachnick & Fidell, 2013). This was the case for each item for both models.

Table C1

Perceived Credibility Items, Means, Standard Deviations and Inter-item Correlations (N = 205)

The victim	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13
(1) accurately remembered the event.	4.07	1.42													
(2) was cautious when she selected somebody from the lineup.	3.62	1.68	.38												
(3) wanted to choose someone from the lineup to help the police.	4.41	1.70	-.02	-.10											
(4) wanted to have someone to blame.	4.40	1.72	-.20	-.07	.50										
(5) would have identified anyone from the lineup for justice to be served.	3.40	1.67	-.25	-.14	.38	.55									
(6) gave a consistent account of the event.	3.75	1.56	.64	.35	-.03	-.18	-.27								
(7) was plausible.	5.17	1.37	.53	.18	.06	-.08	-.27	.55							
(8) was confident when she identified the robber at the lineup procedure.	5.30	1.50	.35	.12	-.10	-.25	-.27	.35	.44						
(9) was confident when she testified at trial.	4.54	1.39	.42	.15	-.15	-.21	-.21	.40	.42	.52					
(10) was truthful.	5.31	1.36	.52	.26	.02	-.15	-.35	.44	.66	.35	.37				
(11) was credible.	4.55	1.39	.61	.37	-.09	-.22	-.35	.55	.62	.43	.45	.62			
(12) was unbiased.	3.75	1.60	.41	.21	-.03	-.15	-.17	.41	.34	.24	.25	.32	.36		
(13) was honest.	5.19	1.36	.49	.26	.00	-.13	-.33	.52	.65	.43	.39	.78	.61	.34	
(14) was trustworthy.	4.78	1.33	.60	.35	-.01	-.11	-.30	.55	.57	.36	.40	.67	.68	.41	.70

Note: Higher mean scores denote greater agreement to the items. Possible scores: 1–7. Correlations $r > .11$ are significant ($p < .05$).

Table C2

Summary of Principal Component Analysis Results for Perceived Victim Credibility (N = 205)

	2 Components		3 Components		
The victim…	1	2	1	2	3
was trustworthy.	**0.83**	-0.04	0.59	**0.59**	-0.01
was honest.	**0.82**	-0.05	**0.74**	0.40	-0.01
was credible.	**0.81**	-0.17	**0.61**	0.55	-0.14
was truthful.	**0.81**	-0.05	**0.70**	0.42	-0.01
was plausible.	**0.80**	0.01	**0.77**	0.33	0.06
accurately remembered the event.	**0.76**	-0.10	0.47	**0.64**	-0.08
gave a consistent account of the event.	**0.73**	-0.11	0.46	**0.61**	-0.09
was confident when she testified at trial.	**0.56**	-0.25	**0.67**	0.08	-0.21
was confident when she identified the robber at the lineup procedure.	**0.53**	-0.28	**0.72**	-0.03	-0.23
was unbiased.	**0.52**	-0.08	0.27	**0.50**	-0.07
was cautious when she selected somebody from the lineup.	**0.42**	-0.07	-0.11	**0.81**	-0.10
wanted to have someone to blame.	-0.11	**0.85**	-0.13	-0.06	**0.84**
wanted to choose someone from the lineup to help the police.	0.08	**0.81**	0.09	-0.04	**0.81**
would have identified anyone from the lineup for justice to be served.	-0.31	**0.71**	-0.30	-0.17	**0.70**
Eigenvalues	5.85	1.83	5.85	1.83	1.07
% of variance	41.80	13.04	41.80	13.04	7.65
Cronbach's α	.89	.73	.86	.78	.73

Note. Loadings on the respective component appear in bold.

Inspection of the three-component solution revealed a complex structure. Six of 14 items showed at least 20% overlapping variance on two components. Additional analyses showed that Components 1 and 2 were highly correlated, $r(205) = .71, p < .001$. The correlation between Component 1 and 3 was $r(205) = -.27, p < .001$, and between Component 2 and 3 $r(205) = -.23, p = .001$.

By contrast, the two-component matrix revealed a simple structure, with both components containing a number of strong loadings and items loading substantially on only one of the components. The inter-correlation between Component 1 and Component 2 was $r(205) = -.27, p < .05$.

The two-component solution was preferred, because Components 1 and 2 in the three-component solution were statistically and semantically related, indicating that they measured the same construct. Component 1 contained 11 items that measured *Victim Believability.* Component 2 contained three items that measured *Response Bias*. The reliabilities were Cronbach's $\alpha = .89$, and Cronbach's $\alpha = .73$, respectively, and above the acceptable value of .7.

References

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ Lawrence Erlbaum.

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). Los Angeles, USA: Sage Publications Ltd.

Kaiser, H.F. (1974). An index of factorial simplicity. *Psychometrika, 39,* 31-36.

Kaiser, H. F. and Rice, J. (1974) Little Jiffy, Mark IV. *Educational and Psychological Measurement, 34,* 111-117.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson, Boston.

Watkins, M. W. (2000). *Monte Carlo PCA for parallel analysis* (Computer Software). State College, PA Ed & Psych Associates.

GENERAL DISCUSSION

This dissertation reports three studies that investigated reliability and perceived credibility of older eyewitnesses on the basis of the integrative model for evaluating eyewitness testimony (Sporer, 2008). Studies 1 and 2 examined the reliability of older eyewitnesses at the information-processing level. In Study 1, a meta-analysis of studies of face recognition in young versus old age showed a strong age effect on almost all measures: Older people had fewer hits and more false alarms, and used a more liberal response criterion than young people. Study 2 revealed that identification accuracy decreased with increasing age in a large scale field study. Further, Study 2 shed light on metamemorial processes across the lifespan, showing a dissociation of the relationship between identification performance and confidence, and an increase in error rates across different calibration indices in older age. Finally, findings in Study 3 suggested that "memory judges", in this case mock jurors, were aware of age-related memory changes only to some extent. In the following, the main results of each of the three studies are discussed. For a more detailed discussion, see the respective discussion sections in the three studies.

The Information-Processing Level: Memory for Faces in Old Age

The present dissertation investigated memory for faces in older age at the information-processing level using two research paradigms: face recognition and eyewitness identification. In Study 1, a meta-analysis examined age effects on face recognition memory by including ten hypothesis tests across multiple dependent measures and targets of different age groups (faces of young, mixed-age, and old persons). Consistent with age-related sensorial (Schieber, 2006), neurological (Coleman & O'Hanlon, 2008), and cognitive changes (Salthouse, 1996), the meta-analysis revealed medium to strong age effects (Cohen, 1988) across all recognition performance measures. That is, older adults were not only less likely to correctly recognize previously seen faces, they were also more likely to incorrectly choose a new face as previously seen. Notably, the meta-analysis included only studies with participants that did not report any cognitive deficits or psychiatric impairments. Considering that approximately one-quarter of the older population has cognitive impairments at least to some extent (e.g., Di Carlo et al., 2000; Graham et al., 1997) and that cognitive decline is more pronounced at higher age of the older population (Park, O'Connell, & Thompson, 2003), the meta-analytic findings probably underestimate the age effect in more representative populations.

The meta-analysis was calculated using different methods to corroborate the results. Analyses were conducted via hand-calculated Excel spread sheets and the SPSS Macros by Lipsey and Wilson (2001), and the *R* package "metafor" by Viechtbauer (2010). Each of the methods revealed identical findings for the fixed effects models and for the random effects model using the Excel spread sheet and the SPSS Macros (see Appendix). When the random effects model was applied in *R*, using the Hedges estimator (Hedges & Olkin, 1985) to estimate heterogeneity, the results were virtually identical. Only in three meta-analyses (hits and recognition performance for older faces, and response criterion for young and mixed faces) the results in *R* yielded slightly different results to the Excel spread sheet and the SPSS Macros, presumably due to estimating via the REML in *R* rather than the methods of moments. In these cases, we relied on the results of the Excel spread sheet and the SPSS Macros. Yet overall the findings demonstrated the robustness of the methods employed and ultimately of the results.

Study 2 was a re-analysis of the field study data on eyewitness identification by Sauerland and Sporer (2009). Results showed that older people were less likely to make a correct lineup decision in an eyewitness identification paradigm, regardless of target presence. Moreover, analyses revealed a culvi-linear relationship between identification accuracy and participant age: The identification accuracy increased from adolescence to young age and then steadily decreased until old age, replicating research findings on metacognitive efficiency outside of the identification paradigm (Palmer, David, & Fleming, 2014; Weil et al., 2013). Remarkably, Study 2 was conducted in a field setting using ten young to middle-aged targets, and participants did not know that they were part of an eyewitness identification study until after the target person was out of sight. Thus, these findings are comparable to many real-life criminal cases and police investigations of eyewitnesses, and not specific to one particular target. By and large, the findings in Studies 1 and 2 were in line with meta-analytical results in the eyewitness identification paradigm, which demonstrated an advantage of young adults over older adults in lineup decisions (Erickson, Lampinen, & Moore, 2015; Kocab, Martschuk, & Sporer, 2019).

The meta-analysis also considered some important moderators that may influence memory at the information-processing level. One important factor was the own-age bias. Social-psychological approaches proposed that people are better at recognizing in-group than out-group faces (Sporer, 2001), because out-group faces appear rather “homogeneous” (out-group homogeneity effect: Quattrone & Jones, 1980). Results from Study 1 supported the theoretical notion that age effects were smaller for old than young targets for hits, false alarms and recognition performance. For hits, the effect even became negative. Results were in line with findings by Rhodes and Anastasi (2012) who demonstrated an own-age bias across the lifespan from childhood to old age. Following, the results suggest that fact finders need to consider eyewitness age in relation to target age when evaluating eyewitness testimony.

Finally, meta-regression analyses in Study 1 showed that retention interval was a significant moderator for hits: The longer the retention interval the larger was the difference between young and older adults. No other moderators were associated with hits, and none, not even retention interval, were associated with false alarms. These different results between hits and false alarms were also reflected when these dependent measures were plotted against each other for young and old age groups separately. Young adults showed a negative relationship between hits and false alarms, the so-called “mirror effect” (Glanzer & Adams, 1985), and replicated previous meta-analytical findings on the own-race bias with predominantly young participants (Meissner & Brigham, 2001). By contrast, the relationship between hits and false alarms was non-existent or even positive for older adults. Thus, these findings suggest that recognition and forgetting over time operate differently in old compared to young age, and that forgetting is more pronounced in older age. Unfortunately, the mirror effect could not be tested statistically, because the primary studies did not report intercorrelations between hits and false alarms. Thus, a statistical comparison of the mirror effect of old and young participants in future studies is desirable.

The Metamemory Level: Confidence-Accuracy Relationship in Old Age

At the metamemory level, Study 2 of the present dissertation tested the effectiveness of metacognitive strategies across the lifespan of participants who made a positive identification (i.e., choosers from an experimentally controlled field study by Sauerland & Sporer, 2009). Different types of calibration indices and Bayesian analyses were applied to assess metacognitive judgments of different age groups from adolescence to old age.

Previous research findings suggested that confidence of choosers was strongly correlated with identification accuracy (see the meta-analysis by Sporer, Penrod, Read, & Cutler, 1995). Recent re-analyses by Wixted and colleagues of a series of studies drew even stronger conclusions that postidentification confidence was a reliable indicator of identification accuracy (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted, Read, & Lindsay, 2016; Wixted & Wells, 2017). Findings from Study 2 confirmed that confidence is a reliable postdictor of identification accuracy for *young* adults between 21 and 30 years of age—the age group that typically constitutes the majority of research participants. This age group was reasonably well calibrated and performed above average on all statistical measures of calibration. Further, high confidence in this age group was more diagnostic of guilt across different prior base rates of target-presence than in any other age group.

However, general metamemory research shows that metacognitive monitoring of newly learned information is impaired in older age (Dodson, Bawa, & Krueger, 2007; Dodson & Krueger, 2006), and that older people display high confidence errors (Dodson et al., 2007). Laboratory eyewitness identification studies showed the same pattern (Searcy, Bartlett, & Memon, 1999; Searcy, Bartlett, Memon, & Swanson, 2001; Wright & Stroud, 2002). Study 2 replicated these findings in the field: Older adults expressed high confidence despite the comparably low identification accuracy. The findings indicated that the usefulness of confidence to differentiate between correct and incorrect positive identification judgments seemed to be limited for participants older than 55 years. Bayesian analyses of different base-rates showed that even high confident decisions of older participants were less diagnostic of guilt than less confident decisions of young adults. Moreover, discrimination indices suggested that the age-related decline may already begin from the age of 30 years.

The present dissertation also considered decision time as an objective metacognitive measure of identification accuracy. In line with the dual-process model that correct decisions are fast and automatic, while inaccurate decisions are time-consuming and conscious (Dunning & Perretta, 2002; Dunning & Stern, 1994), shorter decision times were strongly associated with correct identifications in Study 2. In particular, shorter decisions were four times more likely to be associated with correct identifications than longer decisions, while high confidence was twice as likely to be associated with correct identifications as lower confidence.

Self-reported decision time was, as confidence, a less strong postdictor than the objectively measured decision time, and among older participants was not indicative of identification accuracy. This finding indicated, again, that metacognitive processes are impaired in older age. Notably, participant age remained a significant postdictor of identification accuracy when confidence and decision time (or self-reported decision time) were considered: The older the participants the less likely was the identification correct. Overall, these findings, together with the Study 1 findings, suggest that fact finders should consider eyewitness age when evaluating eyewitness identifications and the speed and confidence associated with it.

The Judgmental Level: Perceived Credibility of Older Eyewitnesses

As discussed above, Studies 1 and 2 showed that decisions made at the information-processing and the metamemory level were impaired in old age. Study 3 aimed to investigate how fact finders perceive older eyewitnesses at the judgmental level, which also known as the interpersonal reality monitoring level (Sporer, 2008). In particular, Study 3 investigated jurors' evaluations of eyewitness memory—those who make the final decision of a defendant's guilt. Therefore, different mock trial scenarios were presented, experimentally manipulating crime severity and eyewitness age—in this case of a victim as opposed to a bystander witness—as legal factors; juror personal and attitudinal characteristics were assessed as extra-legal factors in the decision process.

In accordance with a Gestalt theoretical postulate by Sherif and Sherif (1969) juror characteristics were associated with mock jurors' evaluation of eyewitness identification and probability of guilt. Specifically, mock jurors who were confident in the legal system were more likely to believe the prosecution case. Further, mock jurors who were more skeptical of age-related memory changes were in general more skeptical of both the older and the young victim. However, once case-specific facts were considered, juror characteristics became less important. Instead, mock juror perceptions of the eyewitness, and of the prosecution and the defense case were more important. Similar finding were reported by Pearson et al. (2018) across different trial types, namely that perceived case strength was associated with higher probability of guilt (see also reviews by Devine, 2012; Saks & Spellman, 2016).

Findings regarding mock juror judgments in Study 3 showed no overall effect of manipulated eyewitness age. While Studies 1 and 2 demonstrated that older witnesses were less likely to be correct when recognizing faces, mock jurors' perceived witness believability did not vary as a function of eyewitness age in Study 3. Further, despite the findings in Studies 1 and 2 that older adults are more prone to have a liberal response criterion, no age effect was found on mock jurors' perceived lenient response bias in Study 3. However, an age effect was found when the crime was severe, that is, mock jurors perceived the older eyewitness to be less believable than the young eyewitness in that condition. Similar findings were shown in another study with a case of a convenience store robbery (Neal, Christiansen, Bornstein, & Robicheaux, 2012).

Yet victim age did not influence perceived likelihood that the identification was correct nor culpability ratings. This was somewhat surprising, considering that mock jurors showed a moderate degree of agreement when they were directly asked about changes in memory and cognitive capacity in old age, using the subscale Stereotype Capable-Incapable (Braithwaite, Lynd-Stephenson, & Pigram, 1993). A stronger discrepancy between perceptions of cognitive capacity and credibility was found in a study by Ross, Dunning, Toglia, and Ceci (1990): Mock jurors perceived older witnesses as more credible than their young counterparts, despite the belief that cognitive capacity was more likely to be impaired in older age (see also findings by Allison, Brimacombe, Hunter, & Kadles, 2006; Kwong See, Hoffman, & Wood, 2001). Taken together, these findings indicate that despite some awareness that memory is impaired in older age, mock jurors do not seem to take eyewitness age into account when evaluating eyewitness identifications and rendering guilt judgments of a defendant. Notably, in Study 3 mock jurors were largely unconvinced of the defendant's guilt, as more than 80% of mock jurors acquitted the defendant, probably because no forensic or other evidence or independent witnesses were provided to support the prosecution case (cf. Pearson et al., 2018).

Conclusion

To conclude, the present dissertation demonstrated that memory and metamemory are more fallible in older age than in young age, beginning approximately from the age of 50 years. Therefore, fact finders, that is, police investigators, prosecutors, judges, and jurors, need to consider the age of the witness, together with decision time and expressed confidence when they evaluate identification judgments. A video recording of the lineup decision is recommended, so fact finders can form their own view of the identification decision process.

The present dissertation also showed that mock jurors as fact finders were aware of age-related memory changes only to some extent but that they did not consider them when they evaluated the evidence of older witnesses in a case scenario. Thus, jurors would benefit from educative information in cases involving older eyewitnesses in the form of expert evidence or jury instructions. Further research is required to assess the influence of group deliberations on the decision process when older witnesses are called at the witness stand, as well as how other fact finders within the justice system evaluate identification evidence of older eyewitnesses.

References

Allison, M., Brimacombe, C. A., Hunter, M. A., & Kadlec, H. (2006). Young and older eyewitnesses' use of narrative features in testimony. *Discourse Processes, 41,* 289–314. doi:10.1207/s15326950dp4103_3

Braithwaite, V., Lynd-Stevenson, R., & Pigram, D. (1993). An empirical study of ageism: From polemics to scientific utility. *Australian Psychologist, 28,* 9–15. doi:10.1080/00050069308258857

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Coleman, P. G., & O'Hanlon, A. (2008). Ageing and adaptation. In R. Woods, & L. Clare (Eds.), *Clinical psychology of ageing* (pp. 17–32). West Sussex, England: John Wiley & Sons.

Devine, D. J. (2012). *Jury decision making: The state of the science.* New York: New York University Press.

Di Carlo, A., Baldereschi, M., Amaducci, L., Maggi, S., Grigoletto, F., Scarlato, G., & Inzitari, D. (2000). Cognitive impairment without dementia in older people: Prevalence, vascular risk factors, impact on disability. The Italian longitudinal study on aging. *Journal of the American Geriatric Society, 48,* 775–782. doi:10.1111/j.1532-5415.2000.tb04752.x

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence effors: A misrecollection account. *Psychology & Aging*, *22*, 122–133. doi:10.1037/0882-7974.22.1.122

Dodson, C. S., & Krueger, L. E. (2006). I misremember it well: Why older adults are unreliable eyewitnesses. *Psychological Bulletin & Review*, *13*, 770–775. doi:10.3758/BF03193995

Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10-to 12-second rule for distinguishing accurate from inaccurate positive identification. *Journal of Applied Psychology, 87*, 951–962. doi:10.1037/0021-9010.87.5.951

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, *67,* 818-835. doi:10.1037/0022-3514.67.5.818

Erickson, W. B., Lampinen, J. M., & Moore, K. N. (2016). Eyewitness identifications by older and younger adults: A meta-analysis and discussion. *Journal of Police and Criminal Psychology, 31,* 108–121. doi:10.1007/s11896-015-9176-3

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13,* 8–20. doi:10.3758/BF03198438

Graham, J. E., Rockwood, K., Beattie, B. L., Eastwood, R., Gauthier, S., Tuokko, H., & McDowell, I. (1997). Prevalence and severity of cognitive impairment with and without dementia in an elderly population. *The Lancet, 349,* 1793–1796. doi:10.1016/S0140-6736(97)01007-6

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Academic Press, San Diego, CA.

Kocab, K., Martschuk, N., & Sporer, S. L. (2019). *Identification evidence by elderly witnesses: A meta-analysis*. Unpublished manuscript.

Kwong See, S. T., Hoffman, H. G., & Wood, T. (2001). Perceptions of an elderly eyewitness: Is the older eyewitness believable? *Psychology and Aging, 16,* 346–350. doi:10.1037/0882-7974.16.2.346

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in face memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7,* 1–35. doi:10.1037//1076-8971.7.1.3

Neal, T. M. S., Christiansen, A., Bornstein, B. H., & Robicheaux, T. R. (2012). The effects of mock jurors' beliefs about eyewitness performance on trial judgments. *Psychology, Crime, & Law, 18,* 49–64. doi:10.1080/1068316X.2011.587815

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, *28*, 151–160. doi:10.1016/j.concog.2014.06.007

Park, H. L., O'Connell, J. E., & Thompson, R. G. (2003). A systematic review of cognitive decline in the general elderly population. *International Journal of Geriatric Psychiatry, 18,* 1121–1134. doi:10.1002/gps.1023

Pearson, J. M., Law, J. R., Skene, J. A. G., Beskind, D. H., Vidmar, N., Ball, D. A., … Skene, J. H. P. (2018). Modelling the effects of crime type and evidence on judgments about guilt. *Nature Human Behaviour, 2,* 856–866. doi:10.1038/s41562-018-0451-z

Quattrone, G. A. & Jones, E. E. (1980). The perception of variability within ingroups and outgroups: Implications for the law of small numbers. *Journal of Personality and Social Psychology, 38,* 141–152.

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin, 138,* 146–174. doi:10.1037/a0025750

Ross, D. F., Dunning, D., Toglia, M. P., & Ceci, S. (1990). The child in the eyes of the jury: Assessing mock jurors' perceptions of the child witness. *Law and Human Behavior, 14,* 5–23. doi:10.1007/BF01055786

Saks, M., J., & Spellman, B. A. (2016). *The psychological foundations of evidence law.* New York and London: New York University Press.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103,* 403–428. doi:10.1037/0033-295X.103.3.403

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*, 46–62. doi:10.1037/a0014560

Schieber, F. (2006). Vision and aging. In J. E. Birren, & K. W. Schaie (Eds.), *The handbook of the psychology of aging* (pp. 129–161). Burlington, USA: Elvesier Academic Press.

Searcy, J. H., Bartlett, J. C., & Memon, A. (1999). Age differences in accuracy and choosing in eyewitness identification and race recognition. *Memory & Cognition*, *27*, 538–552. doi:10.3758/BF03211547

Searcy, J. H., Bartlett, J. C., Memon, A., & Swanson, K. (2001). Aging and lineup performance at long retention intervals: Effects of metamemory and context reinstatement. *Journal of Applied Psychology, 86*, 207–214. doi:10.1037/0021-9010.86.2.207

Sherif, M., & Sherif, C. W. (1969). *Social psychology.* New York: Harper & Row.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7,* 36–97. doi:10.1037//1076-8971.7.1.36

Sporer, S. L. (2008). Lessons from the origins of eyewitness testimony research in Europe. *Applied Cognitive Psychology, 22,* 737–757. doi:10.1002/acp.1479

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.

Viechtbauer, W. (2010). Metafor: Meta-analysis package for *R* (Version 1.9-2). [Software]. Retrieved from http://www.metafor-project.org/

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., ... Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*, 264–271. doi:10.1016/j.concog.2013.01.004

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*, 515–526. doi:10.1037/a0039510

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence–accuracy relationship. *Journal of Applied Research in Memory and Cognition*, *4*, 329–334. doi:10.1016/j.jarmac.2016.04.006

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

Wright, D. B., & Stroud, J. (2002). Age differences in lineup identification accuracy: People are better with their own age. *Law & Human Behavior*, *26*, 641–654. doi:10.1023/A:1020981501383

Appendix. Comparison of the Fixed Effects Model (FEM) and the Random Effect Model (REM) Results Without Outliers Based on Self-programmed Excel Macros, SPSS Macros by Lipsey and Wilson (2001), and the *R* package "metafor" by Viechtbauer (2010)

Table A1: *Fixed Effects Model (FEM) and Random Effect Model (REM) Results for Hits*

	Young Faces (k = 7)			Mixed-age Faces (k = 9)			Older Faces (k = 6)		
	Excel	SPSS	*R*	Excel	SPSS	*R*	Excel	SPSS	*R*
FEM									
Z	5.05	5.05	5.05	4.06	4.06	4.06	-2.30	-2.30	-2.30
p	< .001	< .001	< .001	< .001	< .001	< .001	.022	.022	.022
Q	12.54	12.54	12.54	9.78	9.78	9.78	7.25	7.25	7.25
$p(Q)$	.051	.051	.051	.281	.281	.281	.203	.203	.203
Hedges g_u	0.48	0.48	0.48	0.35	0.35	0.35	-0.24	-0.24	-0.24
[95% CI]	[0.30, 0.67]	[0.30, 0.67]	[0.30, 0.67]	[0.18, 0.52]	[0.18, 0.52]	[0.18, 0.52]	[-0.44, -0.04]	[-0.44, -0.03]	[-0.44, -0.03]
REM									
Z	3.52	3.52	3.50	3.55	3.55	3.60	-1.93	-1.93	-1.66
p	< .001	< .001	< .001	< .001	< .001	< .001	.053	.054	.098
VC or τ^2	.070	.070	.070	.015	.015	.015	.029	.029	.055
Hedges g_u	0.49	0.49	0.48	0.34	0.34	0.34	-0.24	-0.24	-0.24
[95% CI]	[0.22, 0.76]	[0.22, 0.76]	[0.21, 0.76]	[0.15, 0.53]	[0.15, 0.53]	[0.16, 0.54]	[-0.49, 0.00]	[-0.49, 0.00]	[-0.52, 0.04]

Note. VC = variance component (in Excel and SPSS); τ^2 = estimated heterogeneity using the Hedges estimator (in *R*).

Table A2: *Fixed Effects Model (FEM) and Random Effect Model (REM) Results for False Alarms*

	Young Faces (k = 7)			Mixed-age Faces (k = 9)			Older Faces (k = 7)		
	Excel	SPSS	*R*	Excel	SPSS	*R*	Excel	SPSS	*R*
FEM									
Z	9.17	9.17	9.17	11.57	11.57	11.57	7.49	7.49	7.49
p	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001
Q	7.93	7.93	7.93	10.09	10.09	10.09	6.07	6.07	6.07
$p(Q)$	.244	.244	.244	.259	.259	.259	.415	.415	.415
Hedges g_u	0.90	0.90	0.90	1.05	1.05	1.05	0.72	0.72	0.72
[95% CI]	[0.71, 1.09]	[0.71, 1.09]	[0.71, 1.09]	[0.87, 1.22]	[0.87, 1.22]	[0.87, 1.22]	[0.53, 0.91]	[0.53, 0.91]	[0.53, 0.91]
REM									
Z	7.88	7.88	7.60	10.25	10.25	9.92	7.45	7.44	7.49
p	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001
VC or τ^2	.022	.022	.029	.020	.020	.021	.001	.001	.000
Hedges g_u	0.89	0.89	0.90	1.07	1.07	1.05	0.72	0.72	0.72
[95% CI]	[0.67, 1.12]	[0.67, 1.12]	[0.67, 1.12]	[0.87, 1.27]	[0.86, 1.27]	[0.84, 1.25]	[0.53, 0.91]	[0.53, 0.91]	[0.53, 0.91]

Note. VC = variance component (in Excel and SPSS); τ^2 = estimated heterogeneity using the Hedges estimator (in *R*).

Table A3: *Fixed Effects Model (FEM) and Random Effect Model (REM) Results for Recognition Performance*

	Young Faces (k = 4)			Mixed-age Faces (k = 7)			Older Faces (k = 5)		
	Excel	SPSS	*R*	Excel	SPSS	*R*	Excel	SPSS	*R*
FEM									
Z	7.86	7.86	7.86	10.57	10.57	10.57	3.57	3.57	3.57
p	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001
Q	2.40	2.40	2.40	3.30	3.30	3.30	7.33	7.33	7.33
$p(Q)$	.493	.494	.494	.771	.771	.771	.119	.119	.119
Hedges g_u	1.06	1.06	1.06	0.98	0.98	0.98	0.40	0.40	0.40
[95% CI]	[0.80, 1.32]	[0.80, 1.32]	[0.80, 1.32]	[0.80, 1.17]	[0.80, 1.17]	[0.80, 1.16]	[0.18, 0.61]	[0.18, 0.61]	[0.18, 0.61]
REM									
Z	7.86	7.86	7.86	10.57	10.57	10.57	2.67	2.66	2.88
p	< .001	< .001	< .001	< .001	< .001	< .001	.008	.008	.004
VC or τ^2	.000	.000	.000	.000	.000	.000	.052	.052	.031
Hedges g_u	1.06	1.06	1.06	0.98	0.98	0.98	0.40	0.40	0.40
[95% CI]	[0.80, 1.32]	[0.80, 1.32]	[0.80, 1.32]	[0.80, 1.17]	[0.80, 1.17]	[0.80, 1.16]	[0.11, 0.70]	[0.11, 0.70]	[0.13, 0.66]

Note. VC = variance component (in Excel and SPSS); τ^2 = estimated heterogeneity using the Hedges estimator (in *R*).

Table A4: *Fixed Effects Model (FEM) and Random Effect Model (REM) Results for Response Criterion*

	Young and Mixed-age Faces (k = 10)		
	Excel	SPSS	*R*
FEM			
Z	6.46	6.46	6.46
p	< .001	< .001	< .001
Q	19.97	19.97	19.97
$p(Q)$	.018	.018	.018
Hedges g_u	0.47	0.47	0.47
[95% CI]	[0.33, 0.61]	[0.33, 0.61]	[0.33, 0.61]
REM			
Z	4.52	4.52	3.29
p	< .001	< .001	.001
VC or τ^2	.069	.069	.101
Hedges g_u	0.52	0.52	0.47
[95% CI]	[0.30, 0.75]	[0.30, 0.75]	[0.19, 0.75]

Note. VC = variance component (in Excel and SPSS); τ^2 = estimated heterogeneity using the Hedges estimator (in *R*).

DEUTSCHE ZUSAMMENFASSUNG

In der Aufklärung von Straftaten geben Augenzeugen/innen Informationen über Täter, die Straftat und mögliche Umstände. Insbesondere in Fällen, in denen andere forensische Beweise fehlen, stützen sich Ermittlungsbehörden auf Personenbeschreibungen und Personenidentifizierungen, um Straftäter zu finden bzw. zu überführen. Probleme mit dem Beweiswert von Personenidentifizierungen wurden wiederholt im Laufe der Geschichte des Strafrechts dokumentiert (Sporer, 2008). Mehrere tausend wissenschaftliche Artikel wurden seit Anfang der 1970er Jahre über das Wiedererkennen von Gesichtern bzw. Personen publiziert (z.B. Loftus, 1979; Shepherd, Ellis, & Davies, 1982; Sporer, 1992; Yarmey, 1979), jedoch blieben diese Studien von rechtlichen Entscheidungsträgern oftmals unbemerkt. Mitglieder des Rechtssystems (d.h. Polizeiermittler/innen, Staatsanwälte/innen und Richter/innen) begannen mit einer erhöhten Sensibilität auf das Problem der Falschidentifizierungen zu reagieren, als immer mehr fälschlich verurteilte Personen durch DNA-Analysen entlastet wurden und ein großer Teil dieser Fälle neben anderen Faktoren auch Falschidentifizierungen aufwies (Garrett, 2011). Zugleich ist jedoch zu bedenken, dass Justizirrtümer in diesen Fällen auf ein Zusammenspiel von multiplen Faktoren zurückzuführen waren, zu denen nicht nur Falschidentifizierungen sondern u.a. auch fälschliche Beurteilungen des Beweiswertes von Identifizierungsaussagen gehörten.

In Anlehnung an das integrative Modell von Augenzeugenaussagen und deren Evaluation (Sporer, 2008) können Probleme mit Personenidentifizierungen auf drei Ebenen auftreten: (1) der *Informationsverarbeitungsebene*, (2) der *Metagedächtnisebene* und (3) der *Beurteilungsebene*. Diese drei Ebenen waren Gegenstand der Untersuchung in der vorliegenden Dissertation.

Auf der *Informationsverarbeitungsebene* können Probleme während der Wahrnehmungsphase, des Behaltensintervalls und der Abrufphase auftreten. In diesem Zusammenhang werden zwei Forschungsparadigmen unterschieden, *Wiedererkennen von Gesichtern* und *Personenidentifizierungen*. In Studien zum *Wiedererkennen von Gesichtern* wird während der Wahrnehmungsphase eine große Anzahl von Gesichtern präsentiert. Diese

Gesichter werden in der Erinnerungsphase mit derselben Anzahl von Distraktoren präsentiert. Dabei sollen die Versuchspersonen bei jedem Gesicht entscheiden, ob sie es zuvor gesehen haben oder nicht. In Studien zu *Personenidentifizierungen* sehen die Versuchspersonen typischerweise ein inszeniertes Verbrechen mit ein oder zwei Zielpersonen/Tätern. Während des Abrufs sollen sie die Zielperson in einer Wahlgegenüberstellung bestehend aus einem Tatverdächtigen und fünf bis zehn Vergleichspersonen identifizieren und ihre subjektive Sicherheit, sich richtig entschieden zu haben, ausdrücken (Sporer, 1992). Verschiedene personenbezogene Faktoren können die Zuverlässigkeit der Identifizierungsaussagen beeinflussen, wie zum Beispiel die Aufmerksamkeit, die Gedächtnisleistung oder die Motivation, aber auch die Persönlichkeit und das Alter der/s Augenzeugen/innen. Von speziellem Interesse für diese Dissertation war das Alter. Studie 1 untersuchte anhand einer Metaanalyse den Alterseffekt beim Wiedererkennen von Gesichtern. Dabei wurde die Wiedererkennensleistung von jungen und älteren Erwachsenen miteinander verglichen. Derselbe Effekt wurde auch in Studie 2 dieser Dissertation untersucht, die sich mit der Identifizierungsleistung über die Lebensdauer beschäftigte.

Auf der *Metagedächtnisebene* evaluieren Augenzeugen/innen ihre eigenen Gedächtnisprozesse, indem sie metakognitive Strategien anwenden und ihre subjektive Sicherheit über ihre Erinnerung ausdrücken (Sporer, 2008). Auf dieser Ebene können Fehler auftreten, wenn Augenzeugen/innen eine hohe subjektive Sicherheit über ihre Entscheidung ausdrücken, die in Wirklichkeit falsch ist (*overconfidence*), oder eine niedrige subjektive Sicherheit, wenn ihre Antwort aber in Wirklichkeit richtig ist (*underconfidence*). Ein indirektes Indiz des Metagedächtnisses ist die Antwortlatenz, bzw. die Entscheidungszeit, die ein/e Augenzeuge/in für die Identifizierungsentscheidung benötigt. Diese Metagedächtnisprozesse waren Gegenstand der Untersuchung in Studie 2 der vorliegenden Dissertation, beginnend im Jugendalter bis zum hohen Alter.

Auf der *Beurteilungsebene* evaluieren Urteiler/innen Augenzeugenaussagen, die auf der Informationsverarbeitungs- und der Metagedächtnisebene beobachtet wurden, indem sie verbale und nonverbale Aspekte der Aussagen in Betracht ziehen. Urteiler/innen sind Kriminalbeamte/innen, Staatsanwälte/innen, Sachverständige, Geschworene und

Richter/innen, das heißt Personen, die die Augenzeugenaussagen angesichts vorhandener Beweismittel bewerten, die jedoch bis auf Sachverständige keine Gedächtnisexperten sind. Fehler können auf dieser Ebene entstehen, wenn Urteiler/innen Augenzeugenaussagen fehlerhaft interpretieren, beziehungsweise eine Alternativhypothese nicht betrachten (Sporer, 2008). Studie 3 untersuchte den Entscheidungsprozess von Geschworenen in einer Simulationsstudie, die Urteiler/innen, die über die Schuld des Angeklagten in einem Strafverfahren eines adversarischen Rechtssystems (wie z.B. in den USA, Großbritannien, Australien) entscheiden. Dabei war das Bewusstsein über die alterungsbedingten Gedächtnisveränderungen und dementsprechende Entscheidungen in einem Strafverfahren von besonderem Interesse.

Studie 1: Wiedererkennen von Gesichtern im Alter: Eine Metaanalyse

Qualitative (*narrative*) Reviews dokumentierten einen Alterseffekt beim Wiedererkennen von Gesichtern sowie bei Personenidentifizierungen (Bartlett, 2014; Bartlett & Memon, 2007). Zudem zeigten Metaanalysen einen starken Alterseffekt bei Personenidentifizierungen (Erickson, Lampinen, & Moore, 2015; Kocab, Martschuk, & Sporer, 2019), jedoch wurden bisher keine statistischen Analysen durchgeführt, die diesen Effekt beim Wiedererkennen von Gesichtern testeten. Das Ziel der Studie 1 war, diese Forschungslücke zu füllen und eine Metaanalyse zum Wiedererkennen von natürlichen (d.h. Fotos von natürlichen, nicht invertierten oder computergenerierten), unbekannten Gesichtern mit Menschen verschiedener Altersgruppen durchzuführen. Dabei wurde das Wiedererkennen von Gesichtern von psychisch und kognitiv gesunden jungen (< 30 Jahre) und älteren Erwachsenen (> 60 Jahre) statistisch über alle geeigneten Studien im Sinne der Signaldetektionstheorie (SDT; z.B. Macmillan & Creelman, 2005) verglichen.

Aus der gedächtnispsychologischen Literatur geht hervor, dass sich im natürlichen Alterungsprozess die Gedächtnisleistung verändert. Der Alterungsprozess wird mit einer Verringerung der kognitiven und Gedächtnisfähigkeit (Coleman & O'Hanlon, 2008), sowie mit systematischen Veränderungen der Sehfähigkeit und Verarbeitungsgeschwindigkeit (Salthouse, 1996; Schieber, 2006) in Zusammenhang gebracht. Diese Faktoren können wiederum das Wahrnehmen und das Wiedererkennen von Gesichtern beeinträchtigen (Lott, Haegerstrom-Portnoy, Schneck, & Brabyn, 2005). Zudem werden altersbedingte strukturelle Veränderungen spezifischer Gehirnareale mit der Tendenz assoziiert eine Person zu wählen, gleich ob sie die Zielperson war oder nicht (Huh, Kramer, Gazzaley, & Delis, 2006).

Die haupttheoretischen Ansätze beschreiben Gedächtnisprozesse und alterungsbedingte Veränderungen aus verschiedenen Perspektiven. Der begrenzte Verarbeitungsressourcenansatz (*limited processing resources approach;* z.B., Craik, 1986) postuliert, dass bei älteren Menschen die für das Enkodieren und den Abruf von Informationen vorhandenen Ressourcen begrenzt sind und auf andere Weise genutzt werden. Insbesondere scheint bei älteren Menschen das Verarbeiten von ressourcenintensiven Informationen beschränkt zu sein, vor allem wenn es sich um selbstinitiierte Verarbeitung handelt (Maylor, 2005). Eine Implikation dieses theoretischen Ansatzes auf das Wiedererkennen von Gesichtern wäre, dass ältere Menschen weniger in "tiefere" oder "elaboriertere" Gesichtsverarbeitungen engagieren können—Faktoren, die bei jüngeren Menschen das Wiedererkennen von unbekannten Gesichtern erleichtern (Bower & Karlin, 1974; Sporer, 2001). Die reduzierte Verarbeitungsgeschwindigkeitshypothese (*reduced processing speed hypothesis;* Salthouse, 1996) postuliert, dass die kognitive Leistung sich mit der alterungsbedingten Verringerung der mentalen Geschwindigkeit verschlechtert, weil (a) relevante Aufgaben nicht in einem begrenzten Zeitraum vollendet werden können und (b) relevante Informationen, die gleichzeitig dargeboten werden, nicht mehr vorhanden sind, wenn die Verarbeitung abgeschlossen ist. Folglich weisen ältere Menschen auch bei zeitlich unbegrenzten Aufgaben schlechtere Gedächtnisleistungen auf (Maylor, 2005), was sich darin wiederspiegelt, dass der Informationsabruf scheitert und falsche Erinnerungen produziert werden. Trotz der Unterschiede ähneln sich diese Ansätze insofern, als dass die Informationsverarbeitung mit ansteigendem Alter reduziert ist, was wiederum zu einer schwächeren kognitiven Leistung im Alter führt.

Zusätzlich zu den kognitiven Ansätzen ist es ebenso wichtig, das Wiedererkennen von Gesichtern aus der sozialpsychologischen Sicht zu betrachten. Nach dem *In-group/Out-group Modell* (Sporer, 2001) können Menschen Gesichter ihrer eigenen Gruppe (*in-group*) besser wiedererkennen, als einer Fremdgruppe (*out-group*). Gruppenzugehörigkeit bezieht sich auf Ethnizität, Geschlecht oder Alter. Vorteile zugunsten der jeweiligen Eigengruppe wurden bereits beim *Own-race, Own-sex,* und *Own-age Bias* demonstriert (Herlitz & Lovén, 2013; Meissner & Brigham, 2001; Rhodes & Anastasi, 2012; Shapiro & Penrod, 1986; Sporer, 2001; Wright & Sladden, 2003; Wright & Stroud, 2002). Nach Sporers Modell verarbeiten Beobachter/innen Gesichter der Eigengruppe eher automatisch, das heißt die Verarbeitung ist eher holistisch und auf die Konfiguration von Merkmalen gerichtet, was zu besserem Wiedererkennen führt. Gesichter der Außengruppe hingegen ordnen Beobachter/innen ohne persönlichem Interesse oder weiterer Verarbeitung rasch der Außengruppe (ohne weitere Verarbeitung) zu, was zu einer schlechteren Wiedererkennensleistung führt.

Die Wiedererkennensleistung wird in der Regel in Anlehnung an die Signaldetektionstheorie (SDT; z.B. Macmillan & Creelman, 2005) ausgewertet. Diese postuliert, dass das Vertrautheitsgefühl sich zwischen einem neuen und einem vorher gesehenen Stimulus (in diesem Fall Gesicht) unterscheidet, so dass die Vertrautheit für ein vorher gesehenes höher ist als für ein neues Gesicht (O'Toole, Bartlett, & Abdi, 2000). Versuchspersonen können dabei ein vorher gesehenes Gesicht (*Zielgesicht*) entweder richtig wiedererkennen (*Treffer*) oder fälschlich als ein neues Gesicht (*Distraktor*) zurückweisen (*falsche Zurückweisung*). Einen Distraktor können sie entweder richtig als ein neues Gesicht zurückweisen (*korrekte Zurückweisung*) oder fälschlicherweise als ein altes wiedererkennen (*falscher Alarm*). Da richtige und falsche Zurückweisungen komplementäre Werte zu Treffern und falschen Alarmen sind, werden diese nicht weiter betrachtet. Zwei weitere wichtige SDT Maße sind der *Leistungsindex d',* der Treffer und falsche Alarme kombiniert und den Grad der Diskrimination angibt, und die *Reaktionsneigung beta* (oder *C*), die angibt, ob Personen eher dazu neigen ein Gesicht auszuwählen (laxeres Entscheidungskriterium) oder Gesichter zurückzuweisen (konservatives Entscheidungskriterium). Nach Sporers (2002) In-group/Out-group Modell wird auch ein laxeres Entscheidungskriterium für Gesichter der Außengruppe erwartet. Diese Maße waren Gegenstand der Untersuchung in der vorliegenden Metaanalyse.

Es wurden 19 Studien eingeschlossen, die insgesamt 79 Vergleiche von jüngeren und älteren gesunden Erwachsenen (d.h., ohne kognitive, neurologische oder psychiatrische Einschränkungen) zum Wiedererkennen von unbekannten Gesichtern jungen, mittleren und älteren, sowie gemischten Alters berichteten. Getrennte Metaanalysen wurden zu jedem der einzelnen oben genannten SDT Maße durchgeführt. Gesichter verschiedener Altersgruppen wurden getrennt ausgewertet, um eine Unabhängigkeit zwischen den Einzeleffekten sicherzustellen. Hedges g_u war die Effektstärke, die angibt, wie viele Standardabweichungen sich zwei Gruppen von einander unterscheiden. Nach Cohens (1988) Empfehlungen werden die Effektstärken als klein ($g_u = 0.2$), mittelgroß ($g_u = 0.5$) und groß ($g_u \geq 0.8$) interpretiert.

Die Effekte wurden im Rahmen des *Modells fester Effekte* (*fixed effects model*) und des *Modells zufallsvariabler Effekte* (*random effects model*) gewichtet, bevor und nachdem mögliche Ausreißer nach Hedges und Olkin (1985) identifiziert und entfernt wurden. Hier werden nur die Ergebnisse des Modells zufallsvariabler Effekte nach der Entfernung der Ausreißer berichtet, die sich kaum von den Ergebnissen des Modells fester Effekte unterschieden. Zudem wurden hierarchische Metaregressionsanalysen durchgeführt, um mögliche Moderatoren zu identifizieren und dabei mögliche Interkorrelationen zwischen den einzelnen Prädiktorvariablen in Betracht zu ziehen und somit einen möglichen Konfundierungseffekt zu kontrollieren (Pigott, 2012). Die nicht-parametrische *Trim-and-fill* Methode schätzte einen möglichen Publikationsbias ab (Duval & Tweedie, 2000a, 2000b; Sutton, 2009).

Im Einklang mit alterungsbedingten sensorischen (Schieber, 2006), neurologischen (Coleman & O'Hanlon, 2008) und kognitiven Veränderungen (Salthouse, 1996) übertrafen junge Erwachsene ältere Erwachsene in fast allen abhängigen Variablen. Die Effektgrößen waren klein bis mittelgroß für Treffer und groß für falsche Alarme sowie die SDT-Maße Leistungsindex und Reaktionsneigung. In anderen Worten, junge Erwachsene erkannten nicht nur mit größerer Wahrscheinlichkeit junge und gemischte Zielgesichter ($g_u = 0.31$), sondern sie wiesen auch mit einer größeren Wahrscheinlichkeit Distraktoren zurück ($g_u = 0.95$). Dementsprechend war die allgemeine Wiedererkennensleistung bei jungen Erwachsenen viel höher als bei älteren Erwachsenen ($g_u = 1.01$).

Um einem Spiegeleffekt (*mirror effect*) nach der SDT (Glanzer & Adams, 1985) zu testen, wurden Treffer (*Sensitivität*) und falsche Alarme (1 – *Spezifität*) gegeneinander getrennt für junge und ältere Erwachsene geplottet. Das allgemeine Muster junger Erwachsenen folgte einem negativen Zusammenhang zwischen Treffern und falschen Alarmen—ein Muster, das auch eine Metaanalyse zum *Cross-race Effect* mit überwiegend jungen Erwachsenen ergab (Meissner & Brigham, 2001). Bei älteren Erwachsenen hingegen war der Zusammenhang zwischen Treffern und falschen Alarmen nicht vorhanden beziehungsweise positiv, was auf eine erhöhte Reaktionsneigung bei älteren Erwachsenen hinwies. Die metaanalytischen Ergebnisse zu der Reaktionsneigung bestätigten diese Annahme, da ältere Erwachsene in der Tat eine höhere Tendenz aufwiesen, jemanden auszuwählen als junge Erwachsene (junge und gemischte Gesichter: $g_u = 0.52$).

Die Metaanalyse untersuchte zudem einen möglichen Own-age Bias. In Anlehnung an das In-group/Out-group Modell (Sporer, 2001) war der Alterseffekt bei den abhängigen Variablen falsche Alarme ($g_u = 0.72$) und Leistungsindex *d'* ($g_u = 0.40$) für ältere Gesichter geringer als für jüngere Gesichter. Bei Treffern war der Alterseffekt negativ ($g_u = -0.24$), d.h. ältere Erwachsene erkannten Gesichter ihres Alters besser als junge Erwachsene. Diese Ergebnisse replizierten metaanalytische Befunde zum Own-age Bias von der Kindheit bis zum hohen Alter (Rhodes & Anastasi, 2012).

Metaregressionsanalysen ergaben, dass das Behaltensintervall signifikant mit Treffern, aber nicht mit falschen Alarmen zusammenhing, wenn das Regressionsmodell die verbleibenden Variablen (d.h. Alter der Gesichter, Alter der Versuchspersonen, Darbietungszeit, Anzahl der Zielgesichter, Veränderungen der Emotionen und/oder der Pose, und zusätzliche Aufgaben) kontrollierte. Je länger das Behaltensintervall war, umso größer war der Unterschied zwischen jungen und älteren Erwachsenen für Treffer. Betrachtet man den Unterschied zwischen Treffern und falschen Alarmen im Zusammenhang mit den unterschiedlichen Spiegeleffektmustern zwischen jungen und älteren Erwachsenen, deuten diese Ergebnisse darauf hin, dass die Prozesse des Wiedererkennens und vor allem des Vergessens im jungen und älteren Alter unterschiedlich funktionieren.

Die Alterseffekte in dieser Metaanalyse können als relativ robust interpretiert werden, wenn man die Ergebnisse mit der umfangreichen Metaanalyse von 128 Studien (Shapiro & Penrod, 1986) vergleicht, die deutlich kleinere Effekte für die meisten Variablen in der Literatur feststellten. Weiterhin waren die Effekte viel größer als die des Own-race Bias in der Metaanalyse von Meissner und Brigham (2001). Bedenkt man dabei, dass es sich bei den Versuchspersonen um gesunde Probanden/innen handelte, die keine kognitiven oder Gedächtnisdefizite aufwiesen, sind die Ergebnisse möglicherweise eine Unterschätzung des Alterseffekts der Allgemeinbevölkerung. Jedoch ist in diesem Zusammenhang zu bedenken, dass Trim-und-fill Analysen eine mögliche Publikationsverzerrung bei falschen Alarmen und bei der Wiedererkennensleistung feststellten, so dass die geschätzten Effekte für falsche Alarme (g_u = 0.85) und für die Wiedererkennensleistung (g_u = 0.93) etwas geringer als die errechneten Werte waren.

Studie 2: Subjektive Sicherheit Älterer Augenzeugen/innen: Ist Sie Diagnostisch für die Identifizierungsleistung?

Seit mehr als drei Jahrzehnten haben Studien den Beweiswert der subjektiven Sicherheit auf die Identifizierungsleistung untersucht. Während frühere Metaanalysen ergaben, dass die subjektive Sicherheit mit der Identifizierungsleistung nur moderat assoziiert war (Bothwell, Deffenbacher, & Brigham, 1987; Wells & Murray, 1983), zeigten neuere Metaanalysen, dass die subjektive Sicherheit stark mit Identifizierungsleistung assoziiert war, wenn nur Wähler/innen berücksichtigt werden, die eine positive Identifizierung machten und daher forensisch besonders relevant sind (Sporer, Penrod, Read, & Cutler, 1995). Vor Kurzem argumentierten Forscher nach einer Reihe von Re-analysen vorheriger Studien, dass die subjektive Sicherheit während der ursprünglichen Identifizierung ein zuverlässiger Indikator der Identifizierungsleistung ist (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted, Read, & Lindsay, 2016; Wixted & Wells, 2017). Diese Analysen wurden mit großen Datensätzen durchgeführt, jedoch ohne Berücksichtigung potentieller Moderatoren, wie zum Beispiel das Alter der Augenzeugen/innen oder die Ethnizität der Gesichter.

Demgegenüber weisen metakognitive Modelle darauf hin, dass sich im Alter metakognitive Entscheidungen von der Richtigkeit dieser Entscheidungen dissoziieren (Palmer, David, & Fleming, 2014), da ältere Erwachsene Schwierigkeiten aufweisen, Informationen effektiv zu nutzen und abzurufen, wenn sie Entscheidungen treffen (Hertzog & Hultsch, 2000; Shing, Werkle-Bergner, Li, & Lindenberger, 2009). Die Mehrheit der Studien wurde jedoch mit verbalem und nicht visuellem Stimulusmaterial durchgeführt, geschweige denn Gesichtern.

Jedoch zeigten auch Laborstudien, die diesen Zusammenhang bei Identifizierungen untersuchten, dass die subjektive Sicherheit bei älteren Erwachsenen einen geringeren Beweiswert hatte als bei jungen Erwachsenen (Searcy, Bartlett, & Memon, 1999; Searcy, Bartlett, Memon, & Swanson, 2001; Wright & Stroud, 2002). Diese Fragestellung wurde auch bei Feldstudien in einem Identifizierungsparadigma über die Lebensdauer untersucht (Palmer, Brewer, Weber, & Nagesh, 2013, Altersspanne 14–87; Sauer, Brewer, Zweck, & Weber, 2010, Altersspanne 15–85; Sauerland & Sporer, 2009, Altersspanne 15–84), jedoch wurde das Alter der Versuchspersonen nicht spezifisch analysiert und die damit assoziierten Veränderungen im Zusammenhang zwischen der subjektiven Sicherheit und der Identifizierungsleistung nicht adressiert. Das Ziel der Studie 2 war, diesen Zusammenhang in einer Re-analyse der Feldstudie von Sauerland und Sporer (2009) zu untersuchen, nämlich inwieweit sich der Beweiswert der subjektiven Sicherheit bei einer Wahlgegenüberstellung in Abhängigkeit des Alters verändert.

Zusätzlich zu der subjektiven Sicherheit wurde auch die Entscheidungszeit untersucht, die ein objektives metakognitives Maß darstellt. In Übereinstimmung mit dem Dualprozessmodell, welches besagt, dass richtige Entscheidungen schnell und automatisch sind, während falsche Entscheidungen zeitaufwendig und bewusst sind (Dunning & Perretta, 2002; Dunning & Stern, 1994), werden richtige Identifizierungen vermutlich schneller als falsche Identifizierungen vorgenommen (Sporer, 1992, 1993, 1994). Jedoch bestehen auch hier keine Studien, die diesen Zusammenhang in Abhängigkeit vom Alter untersuchten.

Die Re-analyse wurde mit 436 Wählern von insgesamt 720 Versuchspersonen durchgeführt, die zwischen 15 und 83 Jahre alt waren. Versuchspersonen wurden in sechs Altersgruppen unterteilt: 15–20 Jahre (n = 44); 21–30 (n = 149); 31–40 (n = 74); 41–50

(n = 58); 51–60 (n = 64); und 61–83 (n = 47). Zehn verschiedene weibliche und männliche Zielpersonen zwischen 20 und 37 Jahren wurden getestet, um eine Repräsentativität der Stimuluspersonen (*stimulus sampling*) zu erfüllen (Wells & Windshitl, 1999). Jede der Zielpersonen sprach Personen in einer Fußgängerzone an. Nachdem die Zielperson außer Sichtweite war, klärte eine zweite Person die zuvor angesprochene Person über die Studie auf und bat eine Identifizierung in einer fairen Wahlgegenüberstellung vorzunehmen.

Die Ergebnisse zeigten, dass das Alter der Versuchspersonen ein wichtiger Moderator für den Zusammenhang zwischen der subjektiven Sicherheit und der Identifizierungsleistung war. Während die subjektive Sicherheit über die Lebensdauer konstant blieb, stieg die Identifizierungsleistung vom Jugendalter bis zum jungen Erwachsenenalter leicht an und verringerte sich stetig bis zum höheren Alter. Dementsprechend war die postdiktive Aussagekraft der subjektiven Sicherheit bei Jugendlichen (15–20 Jahre) und jungen Erwachsenen (21–30 Jahre) stark ausgeprägt und nahm stetig bei jeder Altersgruppe ab 30 Jahren ab. Allgemein waren die Ergebnisse für junge Erwachsene im Einklang mit den Ergebnissen von Wixted et al. (2015), welche besagten, dass hohe subjektive Sicherheit einen hohen Beweiswert hatte (92.3% richtige Identifizierungen bei hoher subjektiver Sicherheit). Junge Erwachsene zeigten eine gute und überdurchschnittliche Identifizierungsleistung der gesamten Stichprobe in Bezug auf Maße der Kalibrierung, Über-/Unterschätzung der Richtigkeit, Diskriminationsindex ANRI (*adjusted normalized resolution index*) und SDT Leistungsindex d'. Demgegenüber verringerte sich die Aussagekraft der subjektiven Sicherheit mit ansteigendem Alter, während die eigene Überschätzung (*overconfidence*) anstieg. Die Diskrimination war fast Null ab dem Alter von 30 Jahren und die postdiktive Aussagekraft der subjektiven Sicherheit verringerte sich ab dem Alter von 40 Jahren. Darüber hinaus konnten Erwachsene ab dem Alter von 55 Jahren richtige von falschen Entscheidungen kaum unterscheiden. Der Altersunterschied bestand bei verschiedenen vorherigen Basisraten der Zielpersonpräsenz, wobei die Unterschiede zwischen den Altersgruppen am stärksten für kleine bis mittelgroße vorherige Basisraten der Zielpersonpräsenz waren.

Weiterhin waren in Anlehnung an das Dualprozessmodell (Dunning & Perretta, 2002;

Dunning & Stern, 1994) richtige Identifizierungen in der Tat schneller als falsche Identifizierungen. Jedoch variierte der Zusammenhang zwischen der selbstgeschätzten Entscheidungszeit und der Identifizierungsleistung in Abhängigkeit vom Alter, was wiederum auf metakognitive Defizite im ansteigenden Alter hinweist. Multiple Regressionsanalysen ergaben, dass das Alter der Versuchspersonen mit der Identifizierungsleistung zusammenhing, selbst wenn in den Analysen gleichzeitig subjektive Sicherheit und Entscheidungszeit (bzw. geschätzte Entscheidungszeit) kontrolliert wurden. Darüberhinaus war die Entscheidungszeit viel stärker mit der Identifizierungsleistung (reziproke *OR* [odds ratio] = 3.85) als die subjektive Sicherheit (*OR* = 1.93) assoziiert. Selbst die von den Versuchsteilnehmern selbst geschätzte Entscheidungszeit hatte eine signifikante Aussagekraft in der Identifizierungsleistung (reziproke *OR* = 1.63), die jedoch ab dem Alter von 50 Jahren stark eingeschränkt war. Diese Ergebnisse bestätigen theoretische Annahmen, dass metakognitive Prozesse je nach Alter variieren, und demonstrieren, dass es wichtig ist, Videoaufnahmen des gesamten Identifizierungsverfahrens anzufertigen, um die subjektive Sicherheit und die Entscheidungszeit für spätere Auswertungen festzuhalten (Sporer, 1992, 1993).

Studie 3: Bewusstsein der Geschworenen über Altersbedingte Veränderungen des Gedächtnisses und der Kognitiven Fähigkeit

Im Gerichtsverfahren mit Geschworenen liegt es an ihnen, über die Schuld des Täters zu entscheiden. Jedoch werden in Straf- und anderen Gerichtsverfahren in der Regel gegensätzliche Beweise und Aussagen über die Ereignisse dargeboten (Bornstein & Greene, 2011). In Fällen, in denen das Opfer der/die einzige/r Augenzeuge/in ist, werden Richter/innen und Geschworene mit der Frage konfrontiert, ob das Opfer eine aufrichtige, wahrheitsgemäße (d.h. Glaubwürdigkeit im Sinne des Mangels an Täuschung) *und* richtige, zuverlässige Aussage machte (d.h. Glaubwürdigkeit im Sinne des Mangels an Fehlern; Sporer, 2008). Die zwei Konstrukte schließen sich nicht gegenseitig aus und müssen daher gleichzeitig bei der Beurteilung der Glaubwürdigkeit von Augenzeugen/innen betrachtet werden. Diese Komplexität spiegelt sich auch in Anweisungen für Geschworene wider, da die

Jury die Möglichkeit hat, Beweise ganz oder teilweise zu akzeptieren oder zurückzuweisen (Judicial Commission of New South Wales, 2016; Queensland Courts, 2016).

Nach dem konzeptuellen Modell der Geschworenenentscheidungen (Devine & Caughlin, 2014) werden Charakteristika über den Fall (z.B. Charakteristika des Opfers, Art des Verbrechens), die Geschworenen (z.B. Geschlecht, Erfahrungen, rechtliches Vertrauen) und den Angeklagten (z.B. Ethnizität, kriminelle Vorgeschichte) mit den Fallrückschlüssen und letztlich mit dem Schuldurteil in Zusammenhang gebracht. Geschworene gelten allgemein als aktive Informationsverarbeiter (Bornstein & Greene, 2011), deren Entscheidungen nicht nur Beweise und gesetzliche Richtlinien (d.h. rechtliche Faktoren), sondern auch bestehende Stereotype und Einstellungen (d.h. außerrechtliche Faktoren) reflektieren (Sporer & Goodman-Delahunty, 2009). Wie Sherif und Sherif (1969) im Kontext der sozialen Wahrnehmung postulierten, ist der Beitrag interner Faktoren (z.B. vorherige Erfahrungen, soziale Einstellungen und vorübergehende Erwartungen) auf Urteile und Entscheidungen umso stärker, je mehrdeutiger eine Situation (oder ein zu beurteilender Stimulus) ist (siehe auch Sporer & Goodman-Delahunty, 2009). Studie 3 untersuchte, inwiefern Fallcharakteristika, nämlich das Alter des Opfers und die Schwere des Verbrechens sowie Einstellungen der Geschworenen mit ihren Entscheidungen in einer Simulationsstudie zusammenhängen.

Bisherige Studien zeigten widersprüchliche Ergebnisse in Bezug auf die wahrgenommene Glaubwürdigkeit von älteren Augenzeugen/innen. Während manche Studien auf einen positiven Zusammenhang hinweisen (z.B. Kwong See, Hoffmann, & Wood, 2001), gibt es auch Belege für einen negativen Zusammenhang (z.B. Neal, Christiansen, Bornstein, & Robicheaux, 2012), beziehungsweise dafür, dass ältere Erwachsene weniger zuverlässig und zugleich ehrlicher wahrgenommen werden als junge Erwachsene (z.B. Brimacombe, Quinton, Nance, & Garrioch, 1997).

Auch in Bezug auf die Schwere der Straftat bestehen gegensätzliche Ansätze. Die Schweregrad-Nachsicht-Hypothese (*severity-leniency hypothesis;* Kerr, 1978) geht davon aus, dass Geschworene in schwerwiegenden Fällen eher zu einem Freispruch neigen, da sie die Gefahr einer Verurteilung einer unschuldigen Person in schwereren Fällen (was wiederum zu höheren Strafen führen würde) vermeiden und daher mehr Beweise für die Schuld des Angeklagten verlangen. Ein alternativer Theorieansatz postuliert hingegen, dass Geschworene in schwerwiegenden Fällen, die emotional-geladene Informationen enthalten, eher dazu neigen zu verurteilen um die Möglichkeit, einen gewalttätigen Straftäter freizulassen, zu vermeiden (Bright & Goodman-Delahunty, 2004). Metaanalytische Ergebnisse unterstützen die letztere Hypothese (Grady, Reiser, Garcia, Koeu, & Scurich, 2018). Inwieweit das Alter diesen Zusammenhang moderiert, ist jedoch fraglich.

In Studie 3 nahmen insgesamt 204 jury-fähige Bürger/innen zwischen 18 und 85 Jahren aus Staaten, in denen Geschworenenverfahren üblich sind, an einer Onlinestudie teil. Nachdem sie eine fiktive Zusammenfassung eines Gerichtsverfahrens durchlasen, füllten sie einen kurzen Fragebogen mit Fragen über den Fall und das Opfer, sowie über allgemeine Einstellungen zum Rechtssystem und zu kognitiven Fähigkeit in höherem Alter aus. In diesem Fall wurde ein junger Mann eines Raubüberfalls angeklagt. Das Alter des Opfers (25 Jahre vs. 75 Jahre) und die Schwere der Straftat (Opfer verletzt vs. nicht verletzt) wurden systematisch variiert. Das weibliche Opfer war die einzige Augenzeugin, und sie identifizierte den Tatverdächtigen in einer Wahlgegenüberstellung mit einer hohen subjektiven Sicherheit innerhalb von wenigen Sekunden—Faktoren, die nach den Ergebnissen der Studie 2 einen hohen Beweiswert bezüglich der Identifizierungsrichtigkeit haben. Zudem sagte ein in dem Fall leitender Polizeibeamter gegen den Angeklagten aus. Eine vorherige Pilotstudie mit 40 Versuchspersonen hatte ergeben, dass die Manipulation des Alters erfolgreich war in Bezug auf das Urteil, die Schuldwahrscheinlichkeit und die Wahrscheinlichkeit, dass die Identifizierung richtig war.

Die Ergebnisse der Hauptstudie zeigten, dass Geschworene vor allem in schwerwiegenden Fällen dem jungen Opfer mehr glaubten als dem älteren Opfer. Dementsprechend ordneten sie den Fall der Staatsanwaltschaft stärker und den Fall der Verteidigung schwächer bei dem jungen als dem älteren Opfer ein. Diese Altersunterschiede wirkten sich jedoch nicht auf die wahrgenommene Wahrscheinlichkeit, dass die Identifizierung richtig war, die Schuldwahrscheinlichkeit oder das Urteil aus, obwohl Geschworene den Fragen zu kognitiven und Gedächtnisverschlechterungen im Alter moderat zustimmten (d.h. die Zustimmung zu den Fragen lag über dem Mittel der Skala). Eine mögliche Erklärung ist, dass die Zusammenfassung der Zeugenaussagen nicht den altersspezifischen Sprechstil reflektierte, welche mögliche Rückschlüsse auf die kognitive Fähigkeit der Augenzeugin erlaubt. Wie die Studie von Brimacombe, Quinton, Nance, und Garrrioch (1997) zeigte, waren Aussagen von älteren Zeugen/innen weniger richtig beziehungsweise weniger schlüssig. Entsprechend erachteten Geschworene Aussagen von älteren Zeugen/innen weniger glaubwürdig als von jungen Zeugen/innen, unabhängig davon, ob sie das Alter der/s Zeugen/in kannten oder nicht.

In Anlehnung an den gestalttheoretischen Ansatz von Sherif und Sherif (1969) zeigte Studie 3, dass die Beweiswürdigung (d.h. die Wahrnehmung, dass die Identifizierung richtig war) und die Schuldwahrscheinlichkeit ohne fallspezifische Informationen mit den allgemeinen Einstellungen zum Rechtssystem assoziiert waren. Demzufolge interpretierten Geschworene den Fall umso wahrscheinlicher im Sinne der Staatsanwaltschaft, je mehr Vertrauen sie in das Rechtssystem hatten. Ähnliche Ergebnisse zeigten Devine und Caughlin (2014) in einer Metaanalyse zu dem Zusammenhang zwischen Geschworenencharakteristika und ihren Entscheidungen. Zudem traten die Geschworenen in der vorliegenden Simulationsstudie dem Opfer und dem Fall allgemein skeptischer gegenüber, je mehr sie annahmen, dass die kognitive Kapazität und die Gedächtnisleistung im Alter abnehmen. Sobald jedoch fallspezifische Informationen in Betracht gezogen wurden, verloren die Geschworeneneinstellungen an Bedeutung.

Allgemein war die Verurteilungsrate so gering, dass keine inferenzstatistische Analyse durchgeführt werden konnten. Eine Erklärung dafür ist, dass in dem Fall Aussage gegen Aussage standen und weder unabhängige Augenzeugen/innen noch andere forensische Beweise gegen den Angeklagten erbracht wurden—Faktoren, die mit der Fallauswertung und den Schuldurteilen im Zusammenhang stehen (Pearson et al., 2018). Somit gingen die Geschworenen dieser Simulationsstudie offenbar nach dem Prinzip in dubio pro reo vor und sprachen den Angeklagten frei. Weitere Forschung ist in diesem Zusammenhang notwendig, wie zum Beispiel Darbietung eines inszenierten Strafverfahrens live oder auf Video und Gruppenentscheidungen, um die ökologische Validität dieses Simulationsparadigmas zu erhöhen und den Einfluss anderer Geschworener auf Gruppenentscheidungen abschätzen zu können.

Fazit

Zusammenfassend ging aus den Studien 1 und 2 hervor, dass die Wiedererkennens- und die Identifizierungsleistung auf der Informationsverarbeitungsebene im höheren Alter wesentlich schlechter war als im jungen Erwachsenenalter. Zudem zeigte Studie 2, dass die metakognitiven Prozesse im Alter eingeschränkt waren. Die tatsächliche Identifizierungsleistung und die subjektive Sicherheit waren zunehmend mit ansteigendem Alter dissoziiert, so dass ältere Erwachsene dazu neigten, ihre Identifizierungsleistung zu überschätzen. Diese Ergebnisse stehen im Einklang mit Gedächtnismodellen und metakognitiven Modellen des höheren Alters. Schließlich zeigte Studie 2, dass die tatsächliche Entscheidungszeit eine größere Bedeutung bezüglich der Richtigkeit der Identifizierungsleistung hatte als die subjektive Sicherheit. Daher sollten Urteiler/innen, das heißt Polizeiermittler/innen, Staatsanwälte/innen, Geschworene und Richter/innen, neben dem Alter und der subjektiven Sicherheit der Augenzeugen/innen die objektive Entscheidungszeit berücksichtigen, wenn sie Identifizierungen von Augenzeugen/innen bewerten. Eine Videoaufnahme der Identifizierungsentscheidung ist empfehlenswert, damit Urteiler/innen sich ein eigenes Bild über den Identifizierungsprozess machen können.

Auf der Beurteilungsebene waren die Ergebnisse von Studie 3 jedoch weniger eindeutig. Während Geschworene sich in der Simulationsstudie relativ bewusst darüber waren, dass sich die kognitive Kapazität und das Gedächtnis im Alter verschlechtern, wendeten sie dieses Wissen nicht unbedingt bei ihren Entscheidungen an. Das Alter des Opfers wirkte sich nicht auf die wahrgenommene Wahrscheinlichkeit, dass die Identifizierung richtig war, oder die Schuldwahrscheinlichkeit aus. Diese Ergebnisse deuten darauf hin, dass, um Justizirrtümer weitgehend zu vermeiden, Geschworene (und wahrscheinlich auch andere Urteiler/innen, die keine Gedächtnisexperten sind) von weiterführenden Informationen (z.B. durch Sachverständige) über altersspezifische Veränderungen des Gedächtnisses und des Metagedächtnisses profitieren würden.

Literaturverzeichnis

Bartlett, J. C. (2014). The older eyewitness. In T. Perfect & D. Lindsay (Eds.), *The SAGE handbook of applied memory* (pp. 654–674). London, UK: SAGE Publications Ltd.

Bartlett, J. C., & Memon, A. (2007). Eyewitness memory in young and older adults. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Vol. 2: Memory for people* (pp. 309–338). Mahwah, NJ: Lawrence Erlbaum.

Bornstein, B. H., & Greene, E. (2011). Jury decision making: Implications for and from psychology. *Current Directions in Psychological Science, 20,* 63–67. doi:10.1177/0963721410397282

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). An exploratory study of personality differences in eyewitness memory. *Journal of Applied Psychology*, *72*, 691–695. doi:0021-9010/87

Bower, G. H. & Karlin, M. B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology, 103,* 751–757. doi:10.1037/h0037190

Bright, D. A., & Goodman-Delahunty, J. (2004). The influence of gruesome verbal evidence on mock juror verdicts. *Psychiatry, Psychology and Law, 11,* 154–166, doi:10.1375/pplt.2004.11.1.154

Brimacombe, C. A., Quinton, N., Nance, N., & Garrioch, L. (1997). Is age irrelevant? Perceptions of young and old adult eyewitnesses. *Law and Human Behavior, 21,* 619–634. doi:10.1023/A:1024808730667

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Coleman, P. G., & O'Hanlon, A. (2008). Ageing and adaptation. In R. Woods, & L. Clare (Eds.), *Clinical psychology of ageing* (pp. 17–32). West Sussex, England: John Wiley & Sons.

Craik, F. I. M. (1986). A functional account of age differences in memory. In F. Klix, & H. Hagendorf (Eds.), *Human memory and cognitive capabilities: Mechanisms and* *performances* (pp. 409–422). Amsterdam: Elsevier Science Publishers, North-Holland.

Devine, D. J., & Caughlin, D. E. (2014). Do they matter? A meta-analytic investigation of individual characteristics and guilt judgments. *Psychology, Public Policy, and Law, 20,* 109–134. doi:10.1037/law0000006

Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10-to 12-second rule for distinguishing accurate from inaccurate positive identification. *Journal of Applied Psychology, 87*, 951–962. doi:10.1037/0021-9010.87.5.951

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, *67,* 818–835. doi:10.1037/0022-3514.67.5.818

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

Erickson, W. B., Lampinen, J. M., & Moore, K. N. (2016). Eyewitness identifications by older and younger adults: A meta-analysis and discussion. *Journal of Police and Criminal Psychology, 31,* 108–121. doi:10.1007/s11896-015-9176-3

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge: Harvard University Press.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition, 13,* 8–20. doi:10.3758/BF03198438

Grady, R. H., Reiser, L., Garcia, R. J., Koeu, C., & Scurich, N. (2018). Impact of gruesome photographic evidence on legal decisions: A meta-analysis. *Psychiatry, Psychology and Law, 25,* 503–521. doi:10.1080/13218719.2018.1440468

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition, 21,* 1306–1336. doi:10.1080/13506285.2013.823140

Hertzog, C., & Hultsch, D. F. (2000). Metacognition in adulthood and old age. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp. 417–466). Mahwah, NJ: Lawrence Erlbaum.

Huh, T. J., Kramer, J. H., Gazzaley, A., & Delis, D. C. (2006). Response bias and aging on a recognition memory task. *Journal of the International Neuropsychological Society, 12,* 1–7. doi:10.1017/S1355617706060024

Judicial Commission of New South Wales. (2016). *Criminal Trial Courts Benchbook: Trial procedure.* Retrieved from https://www.judcom.nsw.gov.au/publications/benchbks/criminal/the_jury.html

Kerr, N. L. (1978). Severity of prescribed penalty and mock jurors' verdicts. *Journal of Personality and Social Psychology, 36,* 1431–1442. doi:10.1037/0022-3514.36.12.1431

Kocab, K., Martschuk, N., & Sporer, S. L. (2019). *Identification evidence by elderly witnesses: A meta-analysis*. Unpublished manuscript.

Kwong See, S. T., Hoffman, H. G., & Wood, T. (2001). Perceptions of an elderly eyewitness: Is the older eyewitness believable? *Psychology and Aging, 16,* 346–350. doi:10.1037/0882-7974.16.2.346

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

Lott, L. A., Haegerstrom-Portnoy, G., Schneck, M. E., & Brabyn, J. A. (2005). Face recognition in the elderly. *Optometry and Vision Science, 82,* 874–881. doi:10.1097/01.opx.0000180764.68737.91

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Maylor, E. A. (2005). Age-related changes in memory. In M. L. Johnson (Ed.), *The Cambridge handbook of age and ageing* (pp. 200–208). Cambridge UK: Cambridge University Press.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in face memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7,* 1–35. doi:10.1037//1076-8971.7.1.3

Neal, T. M. S., Christiansen, A., Bornstein, B. H., & Robicheaux, T. R. (2012). The effects of mock jurors' beliefs about eyewitness performance on trial judgments. *Psychology, Crime, and Law, 18,* 49–64. doi:10.1080/1068316X.2011.587815

O'Toole, A. J., Bartlett, J. C., & Abdi, H. (2000). A signal detection model applied to the stimulus: Understanding covariances in face recognition experiments in the context of face sampling distributions. *Visual Cognition, 7,* 436–463. doi:10.1080/135062800394603

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

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness & Cognition*, *28*, 151–160. doi:10.1016/j.concog.2014.06.007

Pearson, J. M., Law, J. R., Skene, J. A. G., Beskind, D. H., Vidmar, N., Ball, D. A., … Skene, J. H. P. (2018). Modelling the effects of crime type and evidence on judgments about guilt. *Nature Human Behaviour, 2,* 856–866. doi:10.1038/s41562-018-0451-z

Pigott, T. E. (2012). *Advances in meta-analysis.* New York: Springer.

Queensland Courts. (2016). *Supreme and District Courts Benchbook: 24 general summing up directions.* Retrieved from http://www.courts.qld.gov.au/__data/assets/pdf_file/0018/86022/sd-bb-24-summing-up-general.pdf

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin, 138,* 146–174. doi:10.1037/a0025750

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition.

Psychological Review, 103, 403–428. doi:10.1037/0033-295X.103.3.403

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence–accuracy relationship for eyewitness identification. *Law & Human Behavior*, *34*, 337–347. doi:10.1007/s10979-009-9192-x

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*, 46–62. doi:10.1037/a0014560

Schieber, F. (2006). Vision and aging. In J. E. Birren, & K. W. Schaie (Eds.), *The handbook of the psychology of aging* (pp. 129–161). Burlington, USA: Elvesier Academic Press.

Searcy, J. H., Bartlett, J. C., & Memon, A. (1999). Age differences in accuracy and choosing in eyewitness identification and race recognition. *Memory and Cognition*, *27*, 538–552. doi:10.3758/BF03211547

Searcy, J. H., Bartlett, J. C., Memon, A., & Swanson, K. (2001). Aging and lineup performance at long retention intervals: Effects of metamemory and context reinstatement. *Journal of Applied Psychology, 86*, 207–214. doi:10.1037/0021-9010.86.2.207

Shapiro, P. N., & Penrod, S. (1986). A meta-analysis of facial identification studies. *Psychological Bulletin, 100,* 139–156. doi:10.1037/0033-2909.100.2.139

Shepherd, J. W., Ellis, H. D., & Davies, G. M. (1982). *Identification evidence: A psychological examination.* Aberdeen: Aberdeen University Press.

Sherif, M., & Sherif, C. W. (1969). *Social psychology.* New York: Harper & Row.

Shing, Y. L., Werkle-Bergner, M., Li, S.-C., & Lindenberger, U. (2009). Committing memory errors with high confidence: Older adults do but children don't. *Memory*, *17*, 168–179. doi:10.1080/09658210802190596

Sporer, S. L. (1992). *Das Wiedererkennen von Gesichtern.* Weinheim: Beltz/Psychologie Verlags Union.

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology*, *78,* 22–33. doi:10.1037/0021-9010.78.1.22

Sporer, S. L. (1994). Decision–times and eyewitness identification accuracy in simultaneous and sequential lineups. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 300–327). New York: Cambridge University Press.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7,* 36–97. doi:10.1037//1076-8971.7.1.36

Sporer, S. L. (2008). Lessons from the origins of eyewitness testimony research in Europe. *Applied Cognitive Psychology, 22,* 737–757. doi:10.1002/acp.1479

Sporer, S. L., & Goodman-Delahunty, J. (2009). Disparities in sentencing decisions. In M. E. Oswald, S. Bieneck, & J. Hupfeld-Heinemann (Eds.), *Social psychology of punishment of crime* (pp. 379–401). New York, NY: John Wiley & Sons Ltd.

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.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (Vol. 2, pp. 435–452). New York: Russell Sage Foundation.

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., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115–1125. doi:10.1177/01461672992512005

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*, 515–526. doi:10.1037/a0039510

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence–accuracy relationship. *Journal of Applied Research in Memory & Cognition*, *4*, 329–334. doi:10.1016/j.jarmac.2016.04.006

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

Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta Psychologica, 114,* 101–114. doi:10.1016/S0001-6918(03)00052-0

Wright, D. B., & Stroud, J. S. (2002). Age differences in lineup identification accuracy: People are better with their own age. *Law and Human Behavior, 26,* 641–654. doi:10.1023/A:1020981501383

Yarmey, A. D. (1979). *The psychology of eyewitness testimony.* London: MacMillan.

PUBLICATION STATUS

References marked with an asterisk indicate the basis for the dissertation or relevant publications. [Mit einem Stern gekennzeichnete Literaturangaben stellen die Grundlage für die Dissertation oder relevante Publikationen dar.]

* Martschuk, N., & Sporer, S. L. (2019). *Mock jurors' awareness of age-related changes in memory and cognitive capacity.* Manuscript submitted for publication.

* Martschuk, N., & Sporer, S. L. (2019). *Zuverlässigkeit und Glaubwürdigkeit älterer Augenzeugen.* Manuscript submitted for publication.

Goodman-Delahunty, J., Martschuk, N., Powell, M., & Westera, N. (2019). Special measures for children in court: Law in action in a multi-agency committee. Australian Social Work. doi:10.1080/0312407X.2019.1624796

Hale, S., Goodman-Delahunty, J., & Martschuk, N. (2019, in press). Interactional management in a simulated police interview: Interpreters' strategies. In M. Mason & F. Rock (Eds.), *The discourse of police interviews.* University of Chicago Press.

Goodman-Delahunty, J., Martschuk, N., Powell, M., & Westera, N. (2019). Prosecutorial discretion about special measure use in Australian cases of child sexual abuse. In P. Stenning & V. Colvin (Eds.), *The evolving role of the public prosecutor: Challenges and innovations* (pp. 169–187). Routledge.

* Martschuk, N., Sporer, S. L., & Sauerland, M. (2019). Confidence of older eyewitnesses: Is it diagnostic of identification accuracy? *Open Psychology, 1,* 132–151. doi:10.1515/psych-2018-0010

Goodman-Delahunty, J., & Martschuk, N. (2018). Securing reliable information in investigative interviews: Coercive and noncoercive strategies preceding turning points. *Police, Practice, & Research,* 1–20. doi:10.1080/15614263.2018.1531752

Hale, S., Goodman-Delahunty, J., & Martschuk, N. (2018). Interpreter performance in police interviews: Differences between trained and untrained interpreters. *The Interpreter and Translator Trainer,* 1–25. doi:10.1080/1750399X.2018.1541649

Martschuk, N., Goodman-Delahunty, J., Powell, M., & Westera, N. (2018). Similarities in modi operandi of institutional and non-institutional child sexual offending: Systematic case comparisons. *Child Abuse and Neglect, 84,* 229–240. doi:10.1016/j.chiabu.2018.08.002

* Martschuk, N., & Sporer, S. L. (2018). Memory of faces in old age: A meta-analysis. *Psychology and Aging, 33,* 904–923. doi:10.1037/pag0000282

Goodman-Delahunty, J., Hodgson, N., Martschuk, N., & Cossins, A. (2017). *Responses to submissions to the Royal Commission's Consultation Paper on Criminal Justice in relation to the Jury Reasoning Research.* Available at: http://www.childabuseroyalcommission.gov.au/getattachment/79755c1f-a91b-4e18-90ef-b86486af2a40/Jury-Reasoning-Research-%E2%80%93-response-from-researcher

Goodman-Delahunty, J., & Martschuk, N. (2017). Using pretest-posttest research designs to enhance jury decisions. *Newcastle Law Review, 12,* 86–103.

Goodman-Delahunty, J., Martschuk, N., & Cossins, A. (2017). Validation of Child Sexual Abuse Knowledge Questionnaire. *Psychology, Crime, and Law, 23,* 391–412. doi:10.1080/1068316X.2016.1258469

Goodman-Delahunty, J., Martschuk, N., & Cossins, A., (2017). What Australian jurors know and do not know about evidence of child sexual abuse. *Criminal Law Journal, 41,* 86–103.

Hale, S., Martschuk, N., Ozolins, U., & Stern, L. (2017). The effect of interpreting modes on witness credibility assessments. *Interpreting, 19,* 70–97. doi:10.1075/intp.19.1.04hal

Goodman-Delahunty, J., Cossins, A., & Martschuk, N. (2016). *Jury reasoning in separate and joint trials of institutional child sexual abuse: An empirical study.* Sydney: Royal Commission into Institutional Responses to Child Sexual Abuse. Available at: http://www.childabuseroyalcommission.gov.au/getattachment/b268080d-599a-4d44-a9c5-c3f8181bae96/Jury-reasoning-in-joint-trials-of-institutional-ch

Goodman-Delahunty, J., Cossins, A., & Martschuk, N. (2016). National jury research published. *Judicial Officers' Bulletin, 28,* 45–48. Available at https://search.informit.com.au/fullText;dn=158825891230494;res=IELHSS

Goodman-Delahunty, J. & Martschuk, N. (2016). Risks and benefits of interpreter-mediated police interviews. *Journal of Criminal Justice and Security, 18*, 451–471.

Goodman-Delahunty, J., Martschuk, N., & Cossins, A. (2016). Programmatic pretest-posttest research to reduce jury bias in child sexual abuse cases. *Oñati Socio-Legal Series, 6,* 283–314. Available at http://ssrn.com/abstract=2786072

Goodman-Delahunty, J., Martschuk, N., & Ockenden, E. (2016). Effects of terrorist charges and threatening criminal conduct on mock juror decisions. *Psychiatry, Psychology, and Law, 23,* 697–708. doi:10.1080/13218719.2015.1120247

Goodman-Delahunty, J., Schuller, R., & Martschuk, N. (2016). Workplace sexual harassment in policing: Perceived psychological injuries by source and severity. *Psychological Injury and Law, 9,* 241–252. doi:10.1007/s12207-016-9265-3

Goodman-Delahunty, J., Cossins, A., & Martschuk, N. (2014). *Persistent misconceptions about child sexual abuse: The impact of specialized educative information and deliberation in a simulated trial.* Published conference paper at the 3rd international conference on the Empirical Studies of Judicial Systems, Taipei, Taiwan.

Goodman-Delahunty, J., Martschuk, N., & Dhami, M. K. (2014). Interviewing techniques to secure cooperation by high value detainees. *Applied Cognitive Psychology, 28,* 883–897. doi:10.1002/acp.3087

* Sporer, S. L., & Martschuk, N. (2014). The reliability of eyewitness identifications by the elderly: An evidence-based review. In M. Toglia, D. Ross, J. Pozzulo, & E. Pica (Eds.), *The elderly witness in court* (pp. 1–37). New York: Psychology Press. ISBN:978-1-84872-613-0

* Sporer, S. L., & Martschuk, N. (2013). Alte Leute als Identifizierungszeugen: Eine evidenzbasierte Zusammenfassung neuerer Befunde. [Elderly as eyewitnesses: An evidence-based review of recent finding.] *Praxis der Rechtspsychologie [Praxis of Legal Psychology], 23,* 87–108. ISSN:0939-9062

ERKLÄRUNG

Ich erkläre: Ich habe die vorgelegte Dissertation selbständig und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

Natalie Martschuk