Using general and eyewitness-specific metamemory assessments to estimate performance in multiple identifications

Renan Benigno Saraiva^{1,2*}, Inger van Boeijen³, Lorraine Hope¹, Robert Horselenberg², Peter J. van Koppen^{2,4}

¹Department of Psychology, University of Portsmouth

²Department of Criminal Law and Criminology, Maastricht University

³Department of Clinical Psychological Science, Maastricht University, The Netherlands

⁴Department of Criminal Law and Criminology, VU University Amsterdam

*Corresponding author: Renan Benigno Saraiva. Department of Psychology, Faculty of Science, University of Portsmouth, United Kingdom (renan.saraiva@port.ac.uk).

Note: This article has not yet been peer-reviewed. Please contact the authors before citing.

Acknowledgements: This research was supported by a fellowship award from the Erasmus Mundus Joint Doctorate Program The House of Legal Psychology (EMJD-LP) with Framework Partnership Agreement (FPA) 2013-0036 and Specific Grant Agreement (SGA) 2013-0036

Abstract

Identifications made by eyewitnesses are compelling evidence for prosecuting a suspect, but inaccurate identifications can have severe consequences including the conviction of innocent persons. Based on metacognitive frameworks, we aimed to distinguish accurate from inaccurate identification decisions by using metamemory assessments. Participants ($n = 203$) first completed an assessment of general and eyewitness-specific metamemory domains, followed by eight successive lineup identifications. We found that self-rated ability in the eyewitness memory domain was predictive of correct identifications, false identifications, confidence and confidence-accuracy calibration. Curiously, higher self-rated ability in the more general memory domain was predictive of fewer correct identifications. We discuss the potential applied value of metamemory assessments as predictors of correct and false lineup identifications, as well as theoretical contributions to underlying mechanisms of confidence.

Keywords: Eyewitness testimony; Memory; Metamemory; Face recognition; Identification; Lineup

General Audience Summary

Eyewitnesses evidence is critical in the criminal justice system, especially in cases lacking other physical evidence. However, it is well documented that eyewitness's memory can be contaminated. In more severe cases, eyewitnesses may wrongly identify an innocent suspect, contributing to miscarriages of justice. To address this issue psychological research has focused on factors that can be used to distinguish accurate from inaccurate identifications. In the current study, we investigated the relation between metamemory and eyewitness identification performance. Metamemory can be defined as the knowledge about one's memory capabilities and about strategies that can aid memory. Participants ($n = 203$) first completed metamemory questionnaires, assessing their own memory capabilities. After this assessment, participants watched eight mock-crime videos and tried to identify the target in a

lineup parades for each of the eight videos. Results indicated that individuals who choose someone from a were more likely to be correct if they reported contentment with their ability to recognize faces and individuals. Also, we observed a relation between participants metamemory assessments and confidence expressed in the identification tasks. These initial results suggest that metamemory assessments is a promising tool for distinguishing accurate from inaccurate eyewitness identifications.

Using general and eyewitness-specific metamemory assessments to estimate performance in multiple identifications

Identifications are compelling evidence for prosecuting a suspect in criminal cases, but inaccurate identifications can have severe consequences including the conviction of innocent suspects (Wells & Olson, 2003). Distinguishing accurate from inaccurate eyewitnesses' identifications is thus an important but challenging issue in the criminal justice system. Much research has focused on factors that can postdict eyewitness identification accuracy, including mainly confidence statements (Brewer & Wells, 2006; Sporer, Penrod, Read, & Cutler, 1995), decision-time (Sauerland, Sagana, Sporer, & Wixted, 2018), and decision processes (Dunning & Stern, 1994; Sauerland & Sporer, 2007). More recently, tests of face recognition performance have also been proposed as informative estimators of identification accuracy and proclivity to choose (Baldassari, Kantner, & Lindsay, 2019; Russ, Sauerland, Lee, & Bindemann, 2018). However, very little work has focused on the relation between self-reported memory ability and eyewitness identification performance. In this study, we tested whether self-ratings of memory ability can be used to discriminate eyewitness's identification accuracy and confidence-accuracy realism, based on theoretical frameworks of metamemory.

Self-judgements about the accuracy of our memories are prevalent in many settings. A student taking a test with a guessing penalty must make that judgement, as must an interviewer relying on his or her notes, as must an eyewitness sworn to provide a truthful account. The study of *metamemory*, the knowledge about one's memory capabilities and about strategies that can aid memory, has expanded rapidly over the years (Dunlosky, Mueller, & Thiede, 2016; Koriat, Nussinson, Bless, & Shaked, 2008). Much of that work has focused on how beliefs about memory ability are formed and maintained (Hertzog & Dixon, 1994). Koriat et al. (2008) distinguish between two classes of cues that can influence

metamemory judgments: experienced-based and information-based cues. Experience-based cues are intrinsic to the subjective learning experience, such as how quickly or easily an item is processed or accessed (Frank & Kuhlmann, 2017). In contrast, information-based cues involve people's beliefs about their own memory capacities and limitations. When students are asked to judge their performance on a test, for example, their judgements may be based on preconceived notions about their competence in the domain tested (Dunning, Johnson, Ehrlinger, & Kruger, 2003). Likewise, eyewitnesses providing information about a crime (Perfect, 2004) or identifying suspects (Olsson & Juslin, 1999) produce confidence statements that are partially based on intrinsic cues of self-capacity (Leippe & Eisenstadt, 2014).

Considerable evidence supports the notion that eyewitness confidence has a positive, albeit not perfect, relationship with identification accuracy (Brewer & Wells, 2006; Wixted & Wells, 2017). That is because confidence judgements are malleable and may be contaminated by many different factors such as biased lineup instructions (Leippe, Eisenstadt, & Rauch, 2009), foil similarity (Charman, Wells, & Joy, 2011), and positive identification feedback (Wells & Bradfield, 1998). Intrinsic factors, such as memory self-efficacy, may also have an important role in eyewitness confidence statements (Leippe & Eisenstadt, 2014). Perfect (2004), for example, found that self-rated ability in eyewitness memory domain (i.e., face recognition and episodic details) was predictive of confidence judgements in a cued-recall task. Regarding identification tasks, Olsson and Juslin (1999), found that individuals who rated themselves as good face recognizers were more accurate and had a stronger confidence-accuracy relation in lineup identifications. This initial evidence suggests that self-ratings of memory ability may be useful to estimate confidence realism among eyewitnesses, however the literature on the subject is sparse.

Some evidence points to a positive relation between memory self-efficacy and memory performance in different tasks (Seeman, McAvay, Merrill, Albert, & Rodin, 1996; Valentijn et al., 2006). In the face matching literature, some studies show moderate to large correlations between self-reported face perception ability and performance in face matching tests (Gray, Bird, & Cook, 2017; Ventura, Livingston, & Shah, 2018). However, studies focusing specifically on face recognition show that individuals have limited insight into their ability to recognize unfamiliar faces (Bindemann, Attard, & Johnston, 2014; Bobak, Mileva, & Hancock, 2018). Bindemann et al. (2014) argue that individuals tend to overgeneralize their ability to recognize familiar faces to situations in which unfamiliar faces need to be identified. It is arguable, then, that the association between memory self-efficacy and memory functioning is the strongest when the self-rated ability is specific to the targeted memory task.

Current theoretical perspectives posit that different memory systems share some basic features (e.g., the means of acquiring new information), and they differ in some other features (e.g., their functions and underlying neural mechanisms; Schacter, Wagner, & Buckner, 2000; Tulving, 2007). Metamemory questionnaires often assess self-perceived performance and functioning of general memory (Troyer & Rich, 2002). Other metamemory questionnaires focus on specific memory domains, such as the Self-evaluation of Memory Systems Questionnaire (Tonković & Vranić, 2011), which measures diverse systems such as episodic memory and semantic memory, and the Eyewitness Metamemory Scale, which measures self-rated ability and endorsement of strategies for face and person recognition (Saraiva et al., 2019). In the current study, we tested both self-ratings of general memory ability and self-ratings of eyewitness memory ability to estimate identification performance.

Some initial evidence has been found linking self-ratings of memory ability and eyewitness identification performance (Olsson & Juslin, 1999; Perfect, 2004). However,

these studies measured memory self-capacity using items of untested reliability and validity. In the current study, we examined the relation between validated measures of metamemory and eyewitness identification performance. Additionally, in Olsson and Juslin (1999), the confidence-accuracy relation could only be analysed at a group level (i.e., individuals with high performance versus individuals with lower performance), because confidence-accuracy relations could not be computed for each participant. In the current study, we investigated how overconfident or underconfident each participant was in a series of lineup identifications. In this approach, it is possible to compute confidence-accuracy relation scores that are specific to each participant, which may be tested as an outcome of metamemory measures. Multiple lineup identification designs have minimal effects on eyewitness accuracy, choosing and confidence (Mansour, Beaudry, & Lindsay, 2017). Hence, we adopted a multiple lineup identification design to allow for the calculation of calibration and over/underconfidence statistics for each participant instead of calculating only group coefficients. We expected that: (H1) metamemory measures would be predictive of identification performance; (H2) individuals with higher metamemory scores would be better calibrated in their confidence-accuracy relationship; (H3) the relationship between metamemory measures and eyewitness identification performance would be stronger for metamemory measures specifically related to eyewitness identification domains.

Method

Participants

To estimate the required sample size a power analysis was conducted on a pilot dataset of 56 participants. Pilot data were used instead of simulated data because it may offer more precise estimates of the variance and random coefficients present in nested data (Snijders, 2005). Simulations were conducted on multilevel models including metamemory

factors as predictors of eyewitness performance scores (Green & MacLeod, 2016). The simulations showed that a sample of 200 participants would be required to detect an odds ratio of 1.30 with 80% power for each predictor. So, we set 200 as our target sample size. The chosen effect size was not established based on previous studies due to lack of investigations on metamemory scores as predictors of eyewitness performance. Instead, it was chosen as a minimum effect size to be of practical relevance (Gelman & Carlin, 2014).

Participants were undergraduate students at universities in the United Kingdom and The Netherlands, who participated in exchange for course credit. A total of 221 participants completed the study, but two participants were excluded for not passing at least four out of the five attention checks, and sixteen other participants were excluded because technical issues were reported (e.g., films not loading properly). The final sample ($n = 203$) comprised of 81% female participants and had a mean age of $M = 20.22$, ranging from 18 to 58 years ($SD = 4.08$).

Materials and Instruments

Stimuli event. Participants watched eight mock crime films, developed in previous studies by Colloff and colleagues (Colloff, Wade, & Strange, 2016; Colloff, in preparation). Each film was 30 seconds long and depicted a simulated crime (e.g., burglary, car theft, laptop theft, drink spiking). All the films showed different individuals acting as perpetrators and victims. Every participant watched all eight films, presented in a random order.

Identification tasks. For each film, participants were instructed to try to identify the perpetrator in the mock-crime, in either a target-present or target-absent simultaneous lineup. In the target present lineups the target (perpetrator) was presented with five fillers, and in the target-absent lineups, six fillers were shown without the target. Target-presence was randomized for each film, but every participant saw an equal number of target-present and target-absent lineups (i.e. four of each presence type). None of the target or fillers had

distinctive feature and all fillers were selected based on modal descriptions of the targets (see Colloff et al., 2016). Participants were instructed that the perpetrator may or may not be in the lineup, and they could evaluate the pictures for as long as needed to make a decision.

Metamemory assessment. Participants' self-ratings of memory ability were assessed using two different psychometric instruments: The Multifactorial Metamemory Questionnaire (MMQ, Troyer and Rich 2002) and the Eyewitness Metamemory Scale (EMS; Saraiva et al., 2019). The MMQ was developed to assess separate dimensions of subjective memory ratings, with 57 items divided into three sub-scales: Contentment, Ability, and Strategy. The contentment sub-scale comprises 18 items (e.g., "I am generally pleased with my memory ability"; $\alpha = 0.92$) rated from 1 (*strongly disagree*) to 5 (*strongly agree*), with higher scores indicating higher levels of memory contentment. The ability sub-scale comprises 20 items concerned with experiences with common memory errors over the past two weeks (e.g., "how often do you forget an appointment?"; $\alpha = 0.89$) from 1 (*all the time*) to 5 (*never*), with higher scores indicating higher levels of self-reported memory ability. The strategy sub-scale comprises 19 items focused on the use of memory strategies during the past two weeks (e.g., "how often do you use a timer or alarm to remind you when to do something?"; $\alpha = 0.82$). The items are assessed on a scale ranging from 1 (*never*) to 5 (*all the time*), with higher scores indicating more frequent use of memory strategies.

The EMS is an instrument comprising 23 items assessing three factors: Contentment, Discontentment and Strategies. All items are rated on a scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The EMS-Contentment factor comprises 10 items (e.g., "My ability to remember faces is much better than other people's ability to remember faces"; $\alpha = 0.82$) with higher scores indicating higher levels memory contentment with respect to ability to remember or recognise faces. The EMS-Discontentment factor has eight items (e.g., "Sometimes I have trouble recognizing a person that I know relatively well"; $\alpha = 0.82$) with

higher scores indicating higher memory discontentment with ability to remember or recognise faces. The EMS-Strategies factor comprises five items (e.g., “I often create a visual image in my mind of a face that I want to remember.”; $\alpha = 0.75$) with higher scores indicating higher endorsement of memory strategies to remember faces.

Procedure. The study was conducted online using the Qualtrics platform. Participants first completed the EMS and MMQ instruments, and each instrument was presented in a randomised order. After completing the metamemory assessment, participants were instructed that they would watch a series of films and answer some questions about them. A filler task lasting one minute was presented after each mock crime film. Participants were then presented with either a target-present or target-absent lineup and were asked to try to identify the perpetrator in the film or choose a ‘not-present’ option. For each lineup, participants were asked to provide a confidence statement about their choice using a scale ranging from 0% - not at all certain to 100% - totally certain. This procedure was repeated for all eight mock-crimes, with the film appearing first, followed by a two-minute filler task and then the lineup identification task. The order of the films was randomized for each participant. Demographic information was requested at the end of the experiment.

Analysis approach. In our design, each participant made multiple lineup decisions. Thus, the data were nested in two levels, with identification trials at Level 1 and participants at Level 2. In a similar approach to that of Mansour et al. (2017), we fitted separate multilevel mixed-effects models to examine the metamemory measures as predictors of lineup decisions. The first model tested for correct identifications and included only the target-present lineups, comparing the number of correct identifications of the target (coded as 1) versus all other decisions (coded as 0). The second model included only the target-absent data, comparing the number of correct rejections (coded as 1) versus foil identifications (coded as 0). For the whole data, we compared the number of false identifications (coded as

1) versus all other decisions (coded as 0). The last multilevel model was fitted to predict confidence statements in the lineup decisions. All effects were estimated using maximum likelihood, with logit link functions for models including categorical outcomes (e.g., accuracy) and link functions for models including continuous outcomes (i.e., confidence). The lme4 package available for R was used for all multilevel modelling (Bates, Mächler, Bolker, & Walker, 2014). Global models were fitted including all metamemory factors as predictors of each outcome variable (Burnham & Anderson, 2002). All predictors were centred around their grand mean, meaning that we subtracted the overall mean of that variable from each subject's score.

Calibration analysis was used to test our hypothesis that metamemory measures would be predictive of participants over/underconfidence. In the calibration approach, the aim is to verify if an eyewitness is more accurate about the information for which they show the greatest confidence. The confidence-accuracy relationship is computed across every level of confidence, typically in 10% increments (e.g., a scale that ranges from 0% - "not at all certain", to 100% - "totally certain"). A perfect calibration occurs when the groups confidence level equals the percentage of accurate answers for that group (e.g., witnesses who express 80% confidence are accurate in 80% of their identifications). Typically, a calibration index is computed, comprising the average squared discrepancy for each decision in a given confidence level and actual proportion of correct decisions in the same confidence group. A calibration index of 0 indicates perfect calibration, and, as such, values closer to 0 indicate a stronger confidence-accuracy relationship. Another related statistic is over/underconfidence (OU), which is computed by subtracting the mean accuracy from mean confidence for the entire witness sample, with scores that range from -1 (underconfidence) to +1 (overconfidence; see Brewer & Wells, 2006 for details on calculations of calibration scores). In this study, we computed calibration and OU scores for each participant using their

multiple identifications and associated confidence ratings. This way, each participant had calibration scores that could be included as outcomes in regression models. Multilevel models were not used for those analysis because the calibration scores are computed by aggregating values from all identification decisions made by the same participant. Thus, multiple regression models were used to analyse the metamemory measures as predictors of participants calibration and over/underconfidence. The Benjamini-Hochberg correction was applied to all p values from the regression models to account for multiple testing and false discovery rates (FDR; Benjamini & Hochberg, 1995). Emphasis is given to results with both raw p -values and FDR lower than .05, but results with significant raw p -values and FDR lower than .10 are also presented.

Results

Table 1 provides the means and standard deviations for all relevant lineup identification variables. Before testing our hypotheses, we calculated Intraclass Correlation Coefficients (ICC) to determine the extent to which our dependent variables varied among participants and lineup trials. Across the different models, we found that the ICC for participants ranged from .01 to .18 and ICC for lineups ranged from .06 to .12. We further conducted likelihood ratio tests comparing random-intercept models and fixed-intercept models for each outcome variable and found that random-intercept models fit the data significantly better than the equal-intercept models for all outcomes (see Table S1 in supplemental materials). Taken as a whole, these results support our use of random coefficient modelling (Burnham & Anderson, 2002). Table 2 provides descriptive statistics and correlation among all metamemory scales. Correlations ranged from $r = -.41$ to $r = .60$, and no indication of multicollinearity was observed (all VIF < 1.83).

[Tables 1 and 2 around here]

The raw p -values and adjusted p -values using Benjamini-Hochberg correction for all regression models are presented in Table 2. The odds ratio, 95% confidence intervals, and random parameter estimates for the models including metamemory scales as predictors of eyewitness categorical outcomes are presented in Table 3. As shown in Table 3, EMS-Contentment was the main predictor of eyewitness identification outcomes. For each unit increase in EMS-Contentment score, the odds of making a correct identification increased by a factor of 1.41, and the odds of making a false identification decreased by a factor of 0.79. Considering raw p -values lower than .05, EMS-Contentment was also an indicator of correct rejections (OR = 1.21, p = .02), but this effect was not significant after applying a correction for false detection rates (adjusted p = .08). Figure 1 presents the predictive probabilities for correct identifications, correct rejections, and false identifications as a function of EMS-Contentment, holding all other variables in the model at their mean value. MMQ-Contentment was also a significant predictor of correct identifications, but each unit increase in its score decreased the odds of correct identifications (OR = 0.74).

[Tables 3 and 4 around here]

Figure 1. Multilevel regression plots of predicted probabilities for model outcomes as a function of EMS-Contentment. Regression lines are predicted relationships and shaded polygons represent 95% confidence intervals.

Considering non-adjusted p -values $< .05$, we found that EMS-Strategies was predictive of correct rejections and false identifications, with each unit increase in its score decreasing the odds of correct rejections by a factor of 0.82, and increasing the odds of false

identifications by a factor of 1.19. To better examine this result, we explored a multilevel model including the metamemory scales as predictors of proclivity to choose in the lineup tasks. For this outcome, every target or filler identification was coded as ‘chooser’ and every lineup rejection was coded as ‘non-chooser’. The coefficients in the model revealed that EMS-Strategies was the only significant predictor of choosing (see Table S2). With each unit increase in EMS-Strategies, the odds of selecting someone from a lineup (i.e., target or filler) increased by a factor of 1.16 (95% CI = 1.04, 1.30, $p = .005$).

The multilevel model including metamemory scores as predictors of confidence revealed that EMS-Contentment, EMS-Strategy, and MMQ-Contentment were predictive of confidence statements (see Table 3). Higher scores in EMS-Contentment ($\beta = 0.34$, adjusted $p = .005$) were indicative of higher confidence, and higher scores in MMQ-Contentment were indicative of lower confidence ($\beta = -0.30$, adjusted $p = .03$). Results from the multiple regression models including metamemory scores as predictors of participants confidence-accuracy calibration and overconfidence are presented in Table 5. Higher scores in EMS-Contentment were indicative of better calibration ($\beta = -0.02$).

[Table 5 around here]

Discussion

We tested the effectiveness of general and domain-specific metamemory as predictors of eyewitness identification performance. The findings indicated that self-rated memory ability for face and person identification, as measured by the EMS, was predictive of eyewitness identification performance. Specifically, higher EMS-Contentment scores were associated with more correct identifications and fewer false identifications. These findings are consistent with previous studies showing a relation between self-perceived memory efficacy and memory performance (Kliegel & Jager, 2006; Valentijn et al., 2006). In contrast, Perfect (2004) observed that self-rated ability in the domains of eyewitness memory was not

a useful predictor of eyewitness performance in identification and cued-recall tests. However, the self-efficacy measure used in that study was one single-item, so inferences of memory self-efficacy from such a measure may be limited. The current data support the notion that self-perceived memory ability for face and person identification contribute to the estimation of objective identification performance. This initial finding signals that assessments of self-memory ability may allow for more accurate predictions of correct and false lineup identifications, although replication and further research is necessary before further application is considered.

Surprisingly, we observed that higher scores on self-rated general memory capacity, as measured by the MMQ, was indicative of fewer correct identifications in target-present lineups. That seems to be the reverse of what was observed with the EMS-Contentment factor, which showed a positive relationship with correct identifications. It may be the case that individuals with higher scores on MMQ-Contentment are overestimating their general memory ability, which may lead to more false identifications and fewer correct identifications. This finding is in line with Rickenbach, Agrigoroaei, and Lachman (2015), who observed that self-assessments of episodic memory ability are often inaccurate relative to actual performance. These results suggest that while higher self-ratings of face memory capacity are indicative of more correct identifications, higher self-ratings of general memory ability are indicative of fewer correct identifications. In other words, individuals who claim to have a strong general memory ability, but weak face memory ability, may be at a higher risk of committing false identifications.

Self-rated ability for face memory and person identification was also related to confidence statements in identification tasks, with higher scores in EMS-Contentment being associated to higher confidence. Koriat et al. (2008) argue that confidence judgements may be based partly on the learning experience, and partly on domain-specific beliefs (e.g., “My

memory is not so good”). In our data, confidence judgements were predicted by self-rated ability for face memory and person identification tasks, lending support to the notion that confidence statements stem in part from self-rated ability in a domain of knowledge (Brewer & Sampaio, 2012). Furthermore, we found that individuals with higher scores in EMS-Contentment may be generally more effective at regulating their confidence judgments to reflect the likely accuracy of their identification decisions. This finding is consistent with those of Olsson and Juslin (1999), who initially found that individuals reporting higher self-rated recognition skill show a more diagnostic confidence-accuracy relation. Interestingly, confidence in lineup identifications was predicted only by self-rated ability for face and person identification, but not by factors on general memory ability. This finding supports the notion of domain-specific memory self-efficacy, defined as an individual’s appraisal of his or her usual ability in a given memory domain (Hertzog & Dixon, 1994).

Endorsement of memory strategies for face and person identification, as measured by the EMS-Strategies factor, was related to more false identifications and fewer correct rejections. These specific findings had a higher risk of false discovery after correcting for multiple testing, which means they may represent spurious effects found by chance. Nevertheless, we decided to discuss these results given the existence of some theoretical support for the observed pattern of results. Exploring this result, we found that EMS-Strategies was the only significant predictor of eyewitness choice, with higher scores in EMS-Strategies indicating a higher probability of choosing someone from a lineup. This finding contributes to the literature on individual differences and eyewitness’s proclivity to choose. Baldassari and colleagues (2019), for example, found that individual differences in proclivity to choose in a face recognition task was predictive of false identifications in culprit-absent lineups. In fact, a growing literature suggests that individual differences related to performance on standardized tests may offer reliable indicators of eyewitness identification

performance (Bindemann, Avetisyan, & Rakow, 2012; Russ et al., 2018). We expand on those findings by suggesting that individuals who endorse more strategies to better remember faces may have a higher proclivity to choose in lineup identifications, contributing to more false identifications and fewer correct rejections. It may be the case that individuals who claim to endorse more strategies may choose someone from a lineup more frequently because they feel that such strategies help them encode stronger memory traces (Chua, Hannula, & Ranganath, 2012). It is important to note, however, that score on the EMS-Strategies factor alone cannot inform whether participants used any strategies that they claimed to use.

There are some limitations associated with current study. First, the metamemory assessment occurred prior to the lineup identification tasks. In determining our procedure, we reasoned that exposure to the identification tasks before the completion of the metamemory assessment would have affected witnesses' self-ratings of their memory ability to a greater degree than completing the measures would affect eyewitness performance (Olsson & Juslin, 1999). This may have been appropriate for the aims in the current study, but future investigation should examine the robustness of self-rated memory ability as predictors of eyewitness performance when measures are obtained after the identification tasks. We also implemented an eyewitness paradigm with a short delay between exposure to the crime and the identification tasks. In naturalistic contexts, the delay between witnessing a crime and completing an identification task can take much longer (Valentine, Pickering, & Darling, 2003). It may be the case that self-ratings of memory ability are stable (Lane & Zelinski, 2003) and could still be predictive of identification accuracy in situations involving longer delays between encoding and recognition. However, it remains to be determined whether the current findings would replicate when assessing witness performance for identifications made after longer delays.

Identifying factors that can distinguish accurate from inaccurate witnesses is an important challenge for the criminal justice system. In this paper we present initial evidence that self-rated memory ability for face and person identification may be a useful predictor of identification performance. Participants who reported higher contentment with their eyewitness memory ability had more correct identification, fewer false identifications, and were also better calibrated in their confidence-accuracy relation. If corroborated by further evidence, this finding could offer the basis for an assessment tool specifically tailored for eyewitness identification decisions. Such an assessment could be used in addition to other known predictors of identification accuracy, such as confidence statements, decision time, and decision processes, in order to better inform decisions regarding eyewitness identification evidence.

Data Availability Statement

All data and R data analysis script files can be found at:

https://osf.io/vu3ab/?view_only=26d1108258c8413e995320c7e0b0b084

Author Contributions

A.B. devised the project and the main conceptual ideas. C.D., D.E., and E.F., provided feedback on the ideas and procedure. A.B. and B.C. constructed the online survey. A.B. carried out the experiment. All authors contributed to the interpretation of the results. A.B. took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Acknowledgements

This research was supported by a fellowship award from the Erasmus Mundus Joint Doctorate Program The House of Legal Psychology (EMJD-LP) with Framework Partnership

Agreement (FPA) 2013-0036 and Specific Grant Agreement (SGA) 2013-0036. We thank Melissa Colloff for sharing materials for this study.

References

- Baldassari, M. J., Kantner, J., & Lindsay, D. S. (2019). The importance of decision bias for predicting eyewitness lineup choices: toward a Lineup Skills Test. *Cognitive Research: Principles and Implications*, 4(2), 1-13. doi: 10.1186/s41235-018-0150-3
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4*. *arXiv [stat.CO]*. Retrieved from <http://arxiv.org/abs/1406.5823>
- Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior*, 25, 475–491. doi: 10.1023/A:1012840831846
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57, 289-300. doi: 10.1111/j.2517-6161.1995.tb02031.x
- Bindemann, M., Attard, J., & Johnston, R. A. (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology*, 1, 986903. doi: 10.1080/23311908.2014.986903
- Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology. Applied*, 18, 277–291. doi: 10.1037/a0029635
- Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification predict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, 1, 96–103. doi: 10.1016/j.jarmac.2012.02.001

- Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2018). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology*, advanced online publication. doi: 10.1177/1747021818776145
- Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, 12, 11–30. doi: 10.1037/1076-898X.12.1.11
- Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language*, 67, 59–77. doi: 10.1016/j.jml.2012.04.002
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer. doi: 10.1007/b97636
- Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior*, 35, 479–500. doi: 10.1007/s10979-010-9261-1
- Chua, E. F., Hannula, D. E., & Ranganath, C. (2012). Distinguishing highly confident accurate and inaccurate memory: Insights about relevant and irrelevant influences on memory confidence. *Memory*, 20, 48–62. doi: 10.1080/09658211.2011.633919
- Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science*, 27, 1227–1239. doi: 10.1177/0956797616655789
- Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. New York, NY: Oxford University Press. doi:

10.1093/oxfordhb/9780199336746.013.14

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12, 83–87. doi:

10.1111/1467-8721.01235

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

Frank, D. J., & Kuhlmann, B. G. (2017). More than just beliefs: Experience and beliefs

jointly contribute to volume effects on metacognitive judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 43, 680–693. doi:

10.1037/xlm0000332

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9, 641-651. doi:

10.1177/1745691614551642

Gray, K. L. H., Bird, G., & Cook, R. (2017). Robust associations between the 20-item prosopagnosia index and the Cambridge Face Memory Test in the general population.

Royal Society Open Science, 4(3), 160923. doi: 10.1098/rsos.160923

Green, P., & MacLeod, C. J. (2016). SIMR : an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7, 493–498. doi:

10.1111/2041-210X.12504

Hertzog, C., & Dixon, R. A. (1994). Metacognitive development in adulthood and old age. In

J. Metcalfe & A. P. Shimamura (Eds.). (1994). *Metacognition: Knowing about knowing* (pp. 227-251). Massachusetts: MIT press.

Kliegel, M., & Jäger, T. (2006). Can the prospective and retrospective memory questionnaire (PRMQ) predict actual prospective memory performance? *Current Psychology*, 25,

182–191. doi: 10.1007/s12144-006-1002-8

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky & R.A. Bjork (Eds). *Handbook of metamemory and memory* (pp. 117-136). New York: Psychology Press.

Lane, C. J., & Zelinski, E. M. (2003). Longitudinal hierarchical linear models of the memory functioning questionnaire. *Psychology and Aging, 18*, 38–53. doi: 10.1037/0882-7974.18.1.38

Leippe, M. R., & Eisenstadt, D. (2014). Eyewitness confidence and the confidence-accuracy relationship in memory for people. In R.C. Lindsay, D. F. Ross, J. D., Read, & M. P. Toglia (Eds.). *The Handbook of Eyewitness Psychology: Memory for people* (pp. 377-425). New York: Psychology Press.

Leippe, M. R., Eisenstadt, D., & Rauch, S. M. (2009). Cueing confidence in eyewitness identifications: Influence of biased lineup instructions and pre-identification memory feedback under varying lineup conditions. *Law and Human Behavior, 33*, 194–212. doi: 10.1007/s10979-008-9135-y

Mansour, J. K., Beaudry, J. L., & Lindsay, R. C. L. (2017). Are multiple-trial experiments appropriate for eyewitness identification studies? Accuracy, choosing, and confidence across trials. *Behavior Research Methods, 49*, 2235–2254. doi: 10.3758/s13428-017-0855-0

Olsson, N., & Juslin, P. (1999). Can self-reported encoding strategy and recognition skill be diagnostic of performance in eyewitness identifications? *The Journal of Applied Psychology, 84*, 42–49. doi: 10.1037/0021-9010.84.1.42

Perfect, T. J. (2004). The role of self-rated ability in the accuracy of confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology, 18*, 157–

168. doi: 10.1002/acp.952

Rickenbach, E. H., Agrigoroaei, S., & Lachman, M. E. (2015). Awareness of memory ability and change: (in)accuracy of memory self-assessments in relation to performance.

Journal of Population Ageing, 8, 71–99. doi: 10.1007/s12062-014-9108-5

Russ, A. J., Sauerland, M., Lee, C. E., & Bindemann, M. (2018). Individual differences in eyewitness accuracy across multiple lineups of faces. *Cognitive Research: Principles and Implications*, 3(30), 1-17. doi: 10.1186/s41235-018-0121-8

Saraiva, R. B., van Boeijen, I. M., Hope, L., Horselenberg, R., Sauerland, M., & Van Koppen, P. (2019). Development and validation of the Eyewitness Metamemory Scale.

Retrieved from psyarxiv.com/mjfea

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(1), e0190416. 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

Schacter, D. L., Wagner, A. D., & Buckner, R. L. (2000). Memory systems of 1999. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford Handbook of Memory* (pp. 627-643). New York: Oxford University Press.

Seeman, T., McAvay, G., Merrill, S., Albert, M., & Rodin, J. (1996). Self-efficacy beliefs and change in cognitive performance: MacArthur studies on Successful Aging.

Psychology and Aging, 11, 538-551. doi: 10.1037/0882-7974.11.3.538

Snijders, T. A. B. (2005). Power and sample size in multilevel modeling. *Encyclopedia of Statistics in Behavioral Science*, 3, 1570–1573. doi: 10.1002/0470013192.bsa492

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in

- simultaneous and sequential lineups. *The Journal of Applied Psychology*, 78, 22-33. doi: 10.1037/0021-9010.78.1.22
- Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, 118, 315–327. doi: 10.1037/0033-2909.118.3.315
- Tonković, M., & Vranić, A. (2011). Self-evaluation of memory systems: Development of the questionnaire. *Aging & Mental Health*, 15, 830-837. doi: 10.1080/13607863.2011.569483
- Troyer, A. K., & Rich, J. B. (2002). Psychometric properties of a new metamemory questionnaire for older adults. *The Journals of Gerontology*, 57, 19–27. doi: 10.1093/geronb/57.1.P19
- Tulving, E. (2007). Are there 256 different kinds of memory. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 39-52). New York: Psychology Press.
- Valentijn, S. A. M., Hill, R. D., Van Hooren, S. A. H., Bosma, H., Van Boxtel, M. P. J., Jolles, J., & Ponds, R. W. H. M. (2006). Memory self-efficacy predicts memory performance: Results from a 6-year follow-up study. *Psychology and Aging*, 21, 165–172. doi: 10.1037/0882-7974.21.2.165
- Valentine, T., Pickering, A., & Darling, S. (2003). Characteristics of eyewitness identification that predict the outcome of real lineups. *Applied Cognitive Psychology*, 17, 969–993. doi: 10.1002/acp.939
- Ventura, P., Livingston, L. A., & Shah, P. (2018). Adults have moderate-to-good insight into their face recognition ability: Further validation of the 20-item Prosopagnosia Index in a Portuguese sample. *Quarterly Journal of Experimental Psychology*, 71, 2677-2679. doi: 10.1177/1747021818765652

- Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, 83, 360-376. doi: 10.1037/0021-9010.83.3.360
- Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology*, 54, 277-295. doi: 10.1146/annurev.psych.54.101601.145028
- Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, 18, 10–65. doi: 10.1177/1529100616686966

Table 1

Main descriptive statistics for identification performance scores

	All Lineups		Lineup 1		Lineup 2		Lineup 3		Lineup 4		Lineup 5		Lineup 6		Lineup 7		Lineup 8	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
Correct identifications	581	36	72	35	66	32	57	28	92	45	73	36	74	36	59	29	88	43
Correct rejections	521	32	49	24	70	34	69	34	43	21	46	23	77	38	74	36	93	46
False Identifications	386	24	74	36	43	21	37	18	57	28	81	40	31	15	47	23	16	08

<i>Means, standard deviations, and correlations of the metamemory measures</i>							
Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5
1. EMS-Contentment	3.96	0.86					
2. EMS-Discontentment	2.96	1.05	-.27** [-.39, -.14]				
3. EMS-Strategy	4.40	1.10	.23** [.10, .36]	.04 [-.10, .18]			
4. MMQ-Contentment	3.69	0.69	.30** [.17, .42]	-.22** [-.35, -.08]	.01 [-.13, .15]		
5. MMQ-Ability	3.50	0.56	.30** [.17, .42]	-.20** [-.33, -.07]	.10 [-.04, .24]	.60** [.50, .68]	
6. MMQ-Strategy	2.91	0.54	.01 [-.13, .15]	.15* [.01, .28]	.26** [.13, .39]	-.29** [-.41, -.16]	-.41** [-.52, -.29]

Note. * indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$. Values in brackets are 95% confidence intervals

Table 3.
Metamemory and Multiple Identifications

Raw p-values and adjusted p-values using Benjamini-Hochberg correction for all regression models

	Correct identifications		Correct rejections		False Identifications		Confidence		Calibration		Over/underconfidence	
	Raw <i>p</i>	BH correction	Raw <i>p</i>	BH correction	Raw <i>p</i>	BH correction	Raw <i>p</i>	BH correction	Raw <i>p</i>	BH correction	Raw <i>p</i>	BH correction
EMS-Contentment	<.001	.005	.02	.08	.002	.02	<.001	.005	.003	.03	.25	.50
EMS-Discontentment	.33	.57	.87	.94	.52	.72	.90	.94	.55	.73	.68	.86
EMS-Strategies	.42	.61	.03	.10	.02	.09	.02	.08	.76	.91	.03	.11
MMQ-Contentment	.005	.03	.38	.57	.11	.24	.004	.03	.08	.21	.69	.86
MMQ-Ability	.05	.16	.37	.57	.85	.94	.33	.57	.96	.96	.15	.32
MMQ-Strategy	.10	.25	.35	.57	.87	.94	.29	.56	.04	.11	.91	.94

Logistic multilevel models testing metamemory scores as predictors of eyewitness identification performance

	Correct identifications		Correct rejections		False Identifications		Confidence	
Fixed coefficients	OR	95% CI	OR	95% CI	OR	95% CI	β	95% CI
EMS-Contentment	1.41	1.17, 1.70**	1.21	1.02, 1.44	0.79	0.68, 0.91*	0.34	0.15, 0.51**
EMS-Discontentment	0.91	0.77, 1.09	0.99	0.84, 1.16	1.05	0.88, 1.22	-0.01	-0.18, 0.16
EMS-Strategies	1.04	0.87, 1.25	0.82	0.69, 0.97	1.19	1.03, 1.38	0.20	0.03, 0.38
MMQ-Contentment	0.74	0.59, 0.91*	0.91	0.75, 1.10	1.15	0.97, 1.36	-0.30	-0.50, -0.09*
MMQ-Ability	0.81	0.64, 1.01	1.12	0.91, 1.38	0.97	0.81, 1.16	0.13	-0.10 0.32
MMQ-Strategy	0.87	0.72, 1.06	1.10	0.92, 1.31	0.97	0.83, 1.14	-0.10	-0.29, 0.09
Random parameters	Variance	<i>SD</i>	Variance	<i>SD</i>	Variance	<i>SD</i>	Variance	<i>SD</i>
Level 2: Participants	0.05	0.23	0.02	0.13	0.19	0.44	0.93	0.96
Level 1: Lineup	0.29	0.54	0.45	0.67	0.37	0.61	0.30	0.55

Note. * indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$. All p -values in the regression models were adjusted using Benjamini-Hochberg false discovery rate

Multiple regressions testing metamemory scores as predictors of calibration and over/underconfidence

Fixed coefficients	Calibration		Over/underconfidence	
	β	95% CI	β	95% CI
EMS-Contentment	-0.022	-0.038, -0.007*	-0.020	-0.055, 0.014
EMS-Discontentment	0.003	-0.008, 0.015	0.005	-0.021, 0.032
EMS-Strategies	-0.001	-0.013, 0.009	0.031	0.004, 0.057
MMQ-Contentment	0.019	-0.002, 0.040	0.009	-0.039, 0.059
MMQ-Ability	0.001	-0.027, 0.028	0.047	-0.018, 0.112
MMQ-Strategy	0.025	0.001, 0.052	0.003	-0.055, 0.061

Note. * indicates $p < .05$. All p-values in the regression models were adjusted using Benjamini-Hochberg false discovery rate

Figure 1. Multilevel regression plots of predicted probabilities for model outcomes as a function of EMS-Contentment. Regression lines are predicted relationships and shaded polygons represent 95% confidence intervals.