the british psychological society
promoting excellence in psychology

Legal and Criminological Psychology (2020), 25, 111–132

www.wileyonlinelibrary.com

Eyewitness metamemory predicts identification performance in biased and unbiased line-ups

Renan Benigno Saraiva[1,2]*, IngervanBoeijen[3], Lorraine Hope[1], Robert Horselenberg[2], Melanie Sauerland[3] and Peter J.vanKoppen[2,4]

[1]Department of Psychology, University of Portsmouth, UK
[2]Department of Criminal Law and Criminology, Maastricht University, The Netherlands
[3]Department of Clinical Psychological Science, Maastricht University, The Netherlands
[4]Department of Criminal Law and Criminology, VU University Amsterdam, The Netherlands

Purpose. Distinguishing accurate from inaccurate identifications is a challenging issue in the criminal justice system, especially for biased police line-ups. That is because biased line-ups undermine the diagnostic value of accuracy post-dictors such as confidence and decision time. Here, we aimed to test general and eyewitness-specific self-ratings of memory capacity as potential estimators of identification performance that are unaffected by line-up bias.

Methods. Participants (N = 744) completed a metamemory assessment consisting of the Multifactorial Metamemory Questionnaire and the Eyewitness Metamemory Scale and took part in a standard eyewitness paradigm. Following the presentation of a mock-crime video, they viewed either biased or unbiased line-ups.

Results. Self-ratings of discontentment with eyewitness memory ability were indicative of identification accuracy for both biased and unbiased line-ups. Participants who scored low on eyewitness metamemory factors also displayed a stronger confidence–accuracy calibration than those who scored high.

Conclusions. These results suggest a promising role for self-ratings of memory capacity in the evaluation of eyewitness identifications, while also advancing theory on self-assessments for different memory systems.

Eyewitnesses play a major role in the criminal justice system, especially in cases lacking other physical evidence. In many jurisdictions, suspects are more likely to be prosecuted if an eyewitness identifies them as the perpetrator of a crime. However, as with other types of evidence, eyewitness identifications can be in error or contaminated (Wixted, Mickes,

**Correspondence should be addressed to Renan Benigno Saraiva, Department of Psychology, Faculty of Science, University of Portsmouth, King Henry Building, Portsmouth PO1 2DY, UK (email: renan.saraiva@port.ac.uk).*

DOI:10.1111/lcrp.12166

& Fisher, 2018). Researchers have identified some factors that can be used to distinguish accurate from inaccurate witnesses, including early statements of confidence (Brewer & Wells, 2006; Wixted & Wells, 2017), decision time during the identification (Sauer, Brewer, & Wells, 2008; Sauerland & Sporer, 2009; Sporer, 1993), and self-reported decision process (Dunning & Stern, 1994; Smith, Lindsay, & Pryke, 2001). However, when eyewitnesses are exposed to biased line-ups, the value of post-dictors such as confidence and decision time is undermined (Charman, Wells, & Joy, 2011; Key *et al.*, 2017).

We tested whether general and eyewitness-specific self-ratings of memory efficacy can be used to discriminate identification performance, based on theoretical frameworks of metamemory. In particular, we aimed to investigate the efficacy of metamemory factors as post-dictors of eyewitness identification for biased and unbiased line-ups. Metamemory refers to the knowledge and awareness that an individual has about their own memory capabilities (Dunlosky & Bjork, 2008). This introspective knowledge is often used to monitor and control one's own memory performance. Research on metacognitive judgements has expanded rapidly, focusing on how well people think they have learned new information (i.e., judgements of learning; Double, Birney, & Walker, 2018) and how well people feel they recognize a particular piece of information (i.e., feeling of knowing; Koriat, 2000).

One predominant view is that metacognitive judgements are inferential in nature, involving a variety of heuristics and cues that have some degree of validity in predicting objective memory performance (Koriat, Ma'ayan, & Nussinson, 2006). Such cues can be divided into experience-based (the subjective learning experience) or information-based (people's beliefs about their own memory capacities and limitations; Koriat, Nussinson, Bless, & Shaked, 2008). For example, metamemory judgements can be influenced by how quickly or easily an item is processed or accessed (Frank & Kuhlmann, 2017) and by preconceived notions about one's own competence in the domain tested (Dunning, Johnson, Ehrlinger, & Kruger, 2003). Understanding metamemory judgements in forensic settings is important because eyewitnesses may produce confidence statements or identification decisions that are partially based on intrinsic cues of self-efficacy (Leippe & Eisenstadt, 2014). That is, confidence judgements produced by eyewitnesses in forensic relevant tasks (e.g., line-up identifications) may not depend only on memory trace strength, but also on other intrinsic cues related to self-perceived memory efficacy (Brewer & Sampaio, 2012).

Brewer and Sampaio (2012) argue that confidence judgements result from the integration of two key components: information related to products and processes of the memory, and the individual's metamemory beliefs. In this prediction, confidence judgements are based partly on the learning experience, and partly on domain-specific beliefs (e.g., 'My memory is not so good'; Hertzog & Dixon, 1994). Studies investigating the role of domain-specific beliefs in eyewitness confidence reports are sparse. Olsson and Juslin (1999), for example, found that individuals who considered themselves to be good face recognizers were more accurate and had a stronger confidence–accuracy relation in line-up identifications. Similarly, Perfect (2004) found that self-rated efficacy in the domain of eyewitness memory (i.e., face recognition and episodic details) was predictive of confidence judgements in a cued-recall task. These initial findings suggest that expressions of confidence in eyewitness settings may be influenced by witnesses' beliefs about their own memory efficacy.

A complementary branch of metamemory research has focused on elucidating the relation between self-rated memory efficacy and objective memory performance. Some longitudinal studies have shown a positive relation between MSE and memory

performance in different tasks (Seeman, McAvay, Merrill, Albert, & Rodin, 1996; Valentijn *et al.*, 2006). Regarding face recognition ability, different tests of subjective and objective performance have been proposed as post-dictors of identification accuracy and proclivity to choose (Bindemann, Brown, Koyas, & Russ, 2012/2016; Grabman, Dobolyi, Berelovich, & Dodson, 2019; Russ *et al.*, 2018). For example, Grabman *et al.* (2019) found that individuals with stronger objective face recognition ability have a stronger eyewitness confidence–accuracy relationship. In the face matching literature, moderate-to-large correlations between self-reported face perception ability and performance in objective face matching tests have been documented (Gray, Bird, & Cook, 2017; Shah, Sowden, Gaule, Catmur, & Bird, 2015; Ventura, Livingston, & Shah, 2018). However, studies focusing specifically on face recognition tasks have shown that individuals have limited insight into their ability to recognize unfamiliar faces (Bobak, Mileva, & Hancock, 2018). It has been argued that individuals tend to overgeneralize their ability to recognize familiar faces to situations in which unfamiliar faces need to be identified (Bindemann *et al.*, 2014). Therefore, it might be expected that the association between self-ratings of memory efficacy and objective memory functioning should be strongest when the self-rated efficacy is specific to the targeted memory task.

Contemporary memory models propose that memory consists of relatively independent systems (Baddeley, 2000; Tulving, 2007). Different memory systems can share some basic features (e.g., the means of acquiring new information), but they differ in some other features (e.g., functions, operating principles, and underlying neural mechanisms; Schacter, Wagner, & Buckner, 2000; Tulving, 2007). Thus, it can be expected that perceived lack of ability in one domain (e.g., semantic memory) may not be predictive of perceived failure in an eyewitness-relevant domain (e.g., memory of faces or episodic memory). In a meta-analysis of 107 studies, Beaudoin and Desrichard (2011) found that the association between memory self-efficacy and performance was stronger for perceived efficacy for a specific memory task compared to perceived global memory efficacy. That is, individuals' assessments of their likely performance on a specific memory task were more closely related to memory accuracy than their self-ratings about their own general memory efficacy. Beaudoin and Desrichard (2011) argue that concurrent memory self-efficacy is more likely to be related to memory performance than global memory self-efficacy because individual's global memory results from the aggregation of individual's appraisals of their performance across distinct domains. Importantly, the same meta-analysis did not find evidence that domain memory self-efficacy is more strongly related to memory performance than global memory self-efficacy. However, this result was obtained in analyses with low statistical power due to the small number of effect sizes associated with specific assessments of domain memory self-efficacy (Beaudoin & Desrichard, 2011). Therefore, it is still unclear whether the relation between memory performance and memory self-efficacy is stronger for domain memory self-efficacy or global memory self-efficacy.

One challenge in investigating domain memory self-efficacy is that psychometric tools assessing global memory self-efficacy are more prevalent than tools assessing domain memory self-efficacy. Global memory self-efficacy is typically assessed using scales that include a variety of items related to many different memory domains and tasks (e.g., remembering important dates, remembering names, remembering facts). The Multifactorial Metamemory Questionnaire (MMQ; Troyer & Rich, 2002) is one example of instrument used to assess self-perceived performance and functioning of global memory. An instrument assessing memory self-efficacy for a specific domain that is more relevant for the current research is the Eyewitness Metamemory Scale, which measures self-rated efficacy and endorsement of strategies for face and person recognition (Saraiva, van

Boeijen, Hope, Horselenberg, & van Koppen, 2019; Saraiva *et al.*, 2019). In the current study, we tested self-ratings of general memory efficacy (MMQ) and self-ratings of eyewitness memory efficacy (EMS) as predictors of line-up identification performance. We also aimed to test identification performance for both biased and unbiased line-ups, given that other predictors of identification accuracy (e.g., confidence and decision time) are dependent on line-up bias (Charman *et al.*, 2011; Key *et al.*, 2017).

A line-up can be considered biased when the suspect differs noticeably from other line-up members so that the suspect 'stands out' among the line-up options (Wells *et al.*, 1998). In such instance, the line-up fillers are implausible and do not serve as functional alternatives to the suspect (Tredoux, 1999). One important issue with biased line-ups is that they undermine the effectiveness of post-dictors of accuracy such as identification confidence and decision time (Charman *et al.*, 2011; Key *et al.*, 2017). That is because subjective likelihood judgments are often based on comparisons between the chosen option and each of the individual alternatives. If implausible alternatives are present, there is increased perceived support for the chosen option, consequently inflating confidence judgements (Windschitl & Chambers, 2004). Charman *et al.* (2011) found that the presence of highly dissimilar fillers inflates witnesses' confidence in mistaken identifications. Similarly, Key *et al.* (2017) suggested that when the suspect stands out, witnesses tend to be overconfident and faster (regardless of accuracy) compared to witnesses exposed to unbiased line-ups. Taken together, these findings suggest that confidence and decision time, normally effective post-dictors of identification accuracy, have little diagnostic value if the identification decision was made from a biased line-up.

Our predictions about the relationship between line-up fairness and self-assessments of memory efficacy draw from the literature on metamemory and task difficulty. In unbiased line-ups, eyewitnesses need to rely more on their memory trace of the perpetrator to recognize one of the line-up members as a match of the remembered suspect's appearance. In contrast, biased line-ups may be perceived as easier because fillers are less similar to the suspect and therefore are implausible options. This perceived lower difficulty in biased line-ups creates a potentially misleading heuristic for metamemory judgements based on perceptual fluency. It has long been known that manipulations of perceptual fluency during retrieval can produce memory illusions (Jacoby & Whitehouse, 1989). One example is the belief that a more easily perceived test item is likely to be an old item. As such, memory misattribution may occur if perceptual ease is mistakenly assumed to indicate the stimulus's prior presentation (Higham & Vokey, 2000). In fact, under conditions of perceptual ease, metacognitive calibration tends to be weak, erring on the side of overconfidence (Chandler, 1994). Therefore, it might be expected that – if self-ratings of memory efficacy are related to identification performance – this relation will be weaker for biased compared to unbiased line-ups. However, it is important to acknowledge that this specific prediction is limited because perceptual fluency occurs at time of test and may be only loosely related to self-ratings of metamemory ability. Nevertheless, in order to examine this research question, we hypothesize that the relationship between metamemory factors and eyewitness identification performance to be weaker for biased than unbiased line-ups.

The current study

The purpose of this study was to test general and eyewitness-specific self-ratings of memory efficacy as predictors of eyewitness identification performance, for both biased and unbiased line-ups. We hypothesized self-ratings of memory efficacy to be related to

eyewitness identification accuracy (H1) and that this relationship would be stronger for self-efficacy in eyewitness-specific memory domains compared to self-efficacy in general memory domains (H2). Furthermore, we predicted individuals with higher self-ratings in the metamemory factors would display a stronger confidence–accuracy calibration than individuals with lower ratings (Olsson & Juslin, 1999; H3). Finally, we expected the relation between metamemory factors and eyewitness identification performance to be weaker for biased than unbiased line-ups (H4).

Method

The data, analysis code, and preregistration of this study can be found on the following Open Science Framework repository: https://osf.io/ymkz9/?view_only=49c11c762050470fbe45880af51512ee

Participants

A total of 1,103 participants completed the study[1]. We applied several exclusion criteria to ensure data quality: (1) 34 cases were removed for taking more than 90 min to complete the experiment, and (2) 97 cases for completing the experiment in under 15 min; (3) 95 cases were removed for not passing at least four out of five attention checks; and (4) 44 cases were removed due to suspicious bot activity (Prims & Motyl, 2018). The final sample (N = 744) had a mean age of M = 29.98, ranging from 18 to 72 years (SD = 12.63), and was comprised of 63% female participants (four participants chose not to disclose gender). Most participants were workers from Amazon Mechanical Turk (54%), followed by university students (34%) and participants found through social media (12%). Participants recruited via Amazon Mechanical Turk received US$1.00, students received course credits, and participants from social media were entered into a prize drawing for two £50 Amazon vouchers.

Our sample size was determined based on the confidence–accuracy calibration analysis, given that it is the most demanding analysis in our design. There are no clear guidelines on sample size requirements for calibration analysis, so we evaluated previous studies and reasoned that 400 choosers would provide stable estimates for calibration curves with five confidence levels (Sauer, Brewer, Zweck, & Weber, 2010; Sauerland & Sporer, 2009). Each participant completed two identifications, so our total amount of observations was 1,488, of which 815 were choosers. The final number of choosers surpassed our initial target of 400 choosers because we permitted data collection to continue during a period of optimal recruitment. There were no concerns of overpowering the experiment as our target sample size was based on achieving stability for the calibration analysis, in which case more data provide better stability.

Materials and instruments

Eyewitness Metamemory Scale (EMS)

The EMS contains 23 items divided into three factors: Contentment, Discontentment, and Strategies (Saraiva, van Boeijen, *et al.*, 2019; Saraiva *et al.*, 2019). All items are rated on a

[1] *Part of the data used in this study was also used in a separate study focused on the development and validation of the Eyewitness Metamemory Scale.*

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'; $\alpha = .85$) with higher scores indicating higher levels of memory contentment with respect to the ability to remember or recognize faces. The EMS-Discontentment factor has eight items (e.g., 'Sometimes I have trouble recognizing a person that I know relatively well'; $\alpha = .89$) with higher scores indicating higher memory discontentment with ability to remember or recognize 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'; $\alpha = .81$) with higher scores indicating higher endorsement of memory strategies to remember faces. Importantly, the items comprising the EMS-Strategies factor do not necessarily depict efficient or inefficient memory strategies, so higher scorers on this factor indicate endorsement of a greater number of different strategies, rather than the use of more efficient strategies.

General metamemory instruments

In addition to the EMS, participants also completed the Multifactorial Memory Questionnaire (Troyer & Rich, 2002), and the Squire Subjective Memory Questionnaire (Bergen, Brands, & Jelicic, 2010; Squire, Wetzel, & Slater, 1979). The MMQ consists of three factors: Contentment ($\alpha = .92$), Ability ($\alpha = .92$), and Strategy ($\alpha = .88$). The contentment factor has 18 items (e.g., 'I am generally pleased with my memory ability') rated from 1 (*strongly agree*) to 5 (*strongly disagree*), with higher scores indicating higher memory contentment. The ability factor has 20 items related to experiences with common memory errors over the past 2 weeks (e.g., 'How often do you forget an appointment?') from 1 (*all the time*) to 5 (*never*), with higher scores indicating better self-reported ability. The strategy factor has 19 items concerning the use of memory strategies during the past 2 weeks (e.g., 'How often do you use a timer or alarm to remind you when to do something?'). The items are assessed on a scale ranging from 1 (*never*) to 5 (*all the time*), with higher scores indicating greater use of memory strategies. The SSMQ consists of 18 items related to the development of memory functioning (e.g., 'My ability to recall things when I really try is'), rated on a 9-point scale ranging from −4 (*worse than ever*) to 4 (*better than ever before*).

Stimulus event

Participants viewed a 75-s film depicting a thief stealing a phone from a victim (adapted from Sauerland *et al.,* 2009, Experiment 4). There were two versions of the video counterbalancing the role of two actresses (victim and perpetrator) to better generalize the results to different suspects. Thus, in one version actress A was the perpetrator, while in the other version actress B was the perpetrator.

Line-ups

Every participant received two line-ups, one for the perpetrator and one for the victim in the stimulus event. All line-ups were presented in a simultaneous format and could be either target-present or target-absent. Target-present line-ups consisted of five fillers and the target (i.e., victim or perpetrator) and target-absent line-ups consisted of six fillers. Target presence and the position of each member in the line-up were randomized for every line-up presentation. Pilot tests were conducted to construct fair and unfair line-ups. In those tests, participants read a description of the target and were asked to select the

person who best matched this description from a line-up of six members. Tredoux's *E* was used as a measure of line-up fairness (Tredoux, 1998). Tredoux's *E* takes a minimum value of 1 and a maximum value that equals the nominal line-up size (six in this case). If some line-up members are selected less often than expected by chance, *E* values decrease towards 1 depending on the number of line-up members falling below chance levels of choosing. Four pilot tests were conducted with a total of 123 participants, adapting the line-ups to create sufficiently fair and unfair line-ups. The final four fair line-ups (i.e., target present and target absent for each of the two targets) had Tredoux' *E* values ranging from 3.81 and 4.57, while the four unfair line-ups had Tredoux' *E* values ranging from 1.54 to 2.56.

Procedure

Participants were recruited to an online experiment presented via Qualtrics. First, participants completed the EMS, followed by the MMQ and SSMQ. The EMS was always shown first, while the MMQ and SSMQ were presented in random order. Participants then watched the mock-crime film, followed by a 5-min filler task. Next, the first line-up was presented and participants were asked to identify the target or choose a 'not-present' option, while also providing a confidence judgement on a scale that ranged from 0% (*not confident at all*) to 100% (*totally confident*). After a 5-min filler task, participants received the second line-up. The order of line-up presentation was randomized for every participant (i.e., either perpetrator first or victim first). Finally, some demographic information including gender, age, and educational level was requested.

Results

In our analyses, we tested choosers and non-choosers separately for two reasons. First, it has been documented that post-dictors of identification performance have different associations for choosers versus non-choosers (e.g., Sauerland & Sporer, 2007, 2019; Sporer, Penrod, Read, & Cutler, 1995). Second, triers of fact are more specifically concerned with eyewitnesses who choose someone from a line-up, rather than eyewitnesses who reject a line-up (Mickes, 2015). Before conducting our main analyses, we examined whether line-up performance was significantly affected by the role of line-up targets (i.e., perpetrator vs. victim line-ups). Logistic regression models showed that identification accuracy did not vary as a function of line-up target role for choosers ($p = .13$) or non-choosers ($p = .86$; see Table 1), so the data from both perpetrator and victim line-ups were aggregated for the following analyses. Tables 2 and 3 present the main descriptive statistics of the line-up identification data.

Metamemory as predictors of eyewitness identification accuracy

First, we focused on the relation between metamemory and eyewitness identification accuracy by fitting regression models with metamemory factors as predictors of eyewitness identification accuracy for choosers and non-choosers. The metamemory factors included as predictors in the regression models were as follows: EMS-Contentment, EMS-Discontentment, EMS-Strategies, MMQ-Contentment, MMQ-Ability, MMQ-Strategy, and SSMQ-Memory Development. For choosers, correct identifications were coded as 1 and incorrect identifications (filler or innocent-suspect identifications) were

Table 1. Descriptive statistics of line-up performance per line-up target role (i.e., perpetrator vs. victim)

	All line-up identifications (*N* = 1488)		Perpetrator line-up (*N* = 744)		Victim line-up (*N* = 744)	
	n	%	*n*	%	*n*	%
Choosers						
Correct identifications	461	31.0	251	33.7	210	28.22
Filler identifications	354	23.7	174	23.3	180	24.2
Non-choosers						
Correct rejections	481	32.3	227	30.5	254	34.1
Incorrect rejections	192	12.9	92	12.3	100	13.4
Target-present Line-ups						
Guilty-suspect identifications	461	31.0	251	33.7	210	28.2
Filler identifications	98	6.58	44	5.91	54	7.25
Line-up rejections	192	12.9	92	12.3	100	13.4
Target-absent line-ups						
Innocent-suspect identifications	86	5.77	45	6.04	41	5.51
Filler identifications	170	11.4	85	11.4	85	11.4
Line-up rejections	481	32.3	227	30.5	254	34.1

Table 2. Descriptive statistics of line-up performance per line-up bias condition

	Biased (*N* = 750)		Unbiased (*N* = 738)	
	n	%	*n*	%
Choosers (*N* = 815)				
Correct identifications	266	35.5	195	26.4
Filler identifications	148	19.7	206	27.9
Non-choosers (*N* = 673)				
Correct rejections	237	31.6	244	33.1
Incorrect rejections	99	13.2	93	12.6
Target-present line-ups (*N* = 751)				
Guilty-suspect identifications	266	66.3	195	55.7
Filler identifications	36	9.0	62	17.7
Line-up rejections	99	24.7	93	26.6
Target-absent line-ups (*N* = 737)				
Innocent-suspect identifications	56	16.0	30	7.7
Filler identifications	56	16.0	114	29.4
Line-up rejections	237	67.9	244	62.9
	d'		*d'*	
	1.41		1.60	

coded as 0. For non-choosers, correct rejections were coded as 1 and incorrect rejections were coded as 0.

Each participant made two line-up decisions. Thus, the data were nested in two levels, with identification trials at Level 1 and participants at Level 2. Accordingly, we first tested for the necessity of using mixed-effects models in order to account for the nested components of the data (Hox, Moerbeek, & Van de Schoot, 2017). Mixed-effect models allow for the simultaneous examination of the effects of individual-level and group-level

Table 3. Descriptive statistics of line-up performance per level of identification confidence

	Line-up confidence level										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Correct identifications	1	5	6	26	23	44	51	83	74	73	75
Filler identifications	3	10	14	29	39	59	77	54	37	24	8
Correct rejections	5	2	8	11	25	40	52	74	86	83	95
Incorrect rejections	6	3	3	10	12	18	25	33	36	21	25
Total choices	15	20	31	76	99	161	205	244	233	201	203

variables on outcomes. That is, the variables composing a mixed-effect model are conceptualized in a hierarchical manner, so that observations can be nested within individuals, individuals can be nested within groups, groups can be nested within communities, and so on. Mixed-effect models are recommended when examining nested data because such models account for the fact that individual observations are, in general, not completely independent, while standard regression models assume observations are independent. In our data, it might be expected that the outcomes observed in the line-up identifications are not independent because each participant completed two identification trials. Intra-class correlation coefficients (i.e., the average correlation measured between observations in the same level) were examined to assess the necessity of using mixed-effect models in our data. We also compared models accounting for nesting in the data (i.e., random-intercept models) with models that did not account for nesting (i.e., fixed-intercept models) in order to determine whether mixed-effect models had a better fit to the data.

The R package lme4 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, subtracting the overall mean of that variable from each subject's score. Across the different models, we found that the intra-class correlation coefficients for participants ranged from .00 to .11 and ICC for line-ups were all .00. We further conducted likelihood ratio tests comparing random-intercept models and fixed-intercept models for each outcome variable and found no evidence that random-intercept models fit the data significantly better than the fixed-intercept models for all outcomes (see Table S1). Taken together, these results do not support the use of random coefficient modelling (Burnham & Anderson, 2002). Therefore, we proceeded with estimating logistic regression models with no random effects. Table 4 provides descriptive statistics and correlation among all metamemory scales. Correlations ranged from $r = -.55$ to $r = .61$. Two out of six diagnostic tests pointed to the presence of multicollinearity in the model, but inspection of variance inflation factor, tolerance, Farrar-Glauber F-tests, and partial correlations revealed negligible multicollinearity, so we proceeded without adopting remedial measures. The Benjamini-Hochberg correction was applied to all *p*-values from the regression models to account for multiple testing and false discovery rates (Benjamini & Hochberg, 1995). Odds ratio (OR) was examined as a measure of effect size in the logistic regression models.

We hypothesized that self-ratings of memory efficacy would be related to eyewitness identification accuracy (H1) and that this relationship would be stronger for self-efficacy in eyewitness-specific memory domains compared to self-efficacy in general memory domains (H2). Our first set of model testing focused on choosers, fitting one model for

Table 4. Means, standard deviations, and correlations of the metamemory factors

Variable	*M*	*SD*	1	2	3	4	5	6
1. EMS-contentment	4.20	1.02						
2. EMS-discontentment	3.53	1.09	-.32**					
3. EMS-strategy	4.50	1.09	.41**	.05				
4. MMQ-contentment	3.62	0.70	.30**	-.55**	.07**			
5. MMQ-ability	3.57	0.63	.31**	-.38**	.17**	.55**		
6. MMQ-strategy	2.86	0.64	.08**	.24**	.19**	-.31**	-.44**	
7. SSMQ	5.89	1.21	.61**	-.22**	.35**	.47**	.41**	.05*

Note. EMS = Eyewitness Metamemory Scale, *M* = Mean, MMQ = Multifactorial Metamemory Questionnaire, *SD* = Standard deviation, SSMQ = Squire Subjective Memory Questionnaire.
*$p < .05$; **$p < .01$; ***$p < .001$; *** indicates $p < .001$

biased line-ups and another model for unbiased line-ups (see Table 5). Among choosers, higher scores in EMS-Discontentment (i.e., memory discontentment with ability to remember or recognize faces) were indicative of lower accuracy for both biased (OR = 0.57, $p < .001$) and unbiased line-ups (OR = 0.56, $p < .001$; see Figure 1). None of the other metamemory factors were significant predictors of choosers accuracy for biased and unbiased line-ups. We then repeated the same steps for the non-choosers subset, fitting logistic regression models using metamemory factors as predictors of accuracy for biased and unbiased line-ups (see Table 6). The results showed that none of the metamemory factors were significant predictors of non-choosers identification accuracy for both biased and unbiased line-ups.

In addition to the initial regression models, we conducted exploratory analyses with regression models including identification confidence as an additional estimator of identification accuracy. In those models, the metamemory measures and confidence were included as predictors of identification accuracy for choosers in biased and unbiased line-ups. The aim of this analysis was to further examine whether the predictive value of the metamemory measures (i.e., information-based judgements) when accounting for the variance explained by identification confidence (i.e., experience-based judgements). The results showed that confidence was a significant predictor of accuracy for both biased and unbiased line-ups (see Table 7). Similar to what was observed in the previous model, EMS-Discontentment was a significant predictor of identification accuracy for both biased and unbiased line-ups, with higher scores of EMS-Discontentment being indicative of less accuracy identifications. Surprisingly, EMS-Contentment was also a significant predictor of identification accuracy for unbiased line-ups, with higher scores of EMS-Contentment being indicative of less accurate identifications.

Metamemory and confidence–accuracy calibration

Calibration analyses were carried out to examine the relation between metamemory measures and the confidence–accuracy relationship in identification tasks. Following Brewer and Wells (2006), calibration curves were created by plotting the proportion of correct responses against five categories of confidence (0–20%, 30–40%, 50–60%, 70–80%, and 90–100%). We first produced calibration curves for choosers versus non-choosers, and biased line-ups versus unbiased line-ups (see Figure 2). The diagonal line represents perfect calibration, such that each level of confidence is equivalent to the level of accuracy for decisions made with that level of confidence. Observations above this line indicate

Table 5. Logistic regression models of metamemory factors as predictors of identification accuracy among choosers

	Biased line-ups			Unbiased line-ups		
Predictor	*B* (SE)	*p*	OR [95% CI]	*B* (SE)	*p*	OR [95% CI]
EMS-contentment	0.05 (.13)	.75	1.05 [0.80, 1.37]	−0.19 (.15)	.51	0.83 [0.61, 1.11]
EMS-discontentment	−0.55 (.14)	<.001***	0.57 [0.43, 0.75]	−0.57 (.14)	<.001***	0.56 [0.42, 0.74]
EMS-strategies	0.08 (.12)	.73	1.08 [0.84, 1.39]	−0.17 (.12)	.51	0.84 [0.65, 1.08]
MMQ-contentment	0.23 (.16)	.46	1.26 [0.92, 1.72]	−0.16 (.15)	.55	0.85 [0.63, 1.13]
MMQ-ability	0.08 (.15)	.73	1.09 [0.81, 1.45]	−0.03 (.15)	.83	0.96 [0.72, 1.30]
MMQ-strategy	0.21 (.13)	.43	1.23 [0.95, 1.60]	0.19 (.13)	.46	1.21 [0.93, 1.57]
SSMQ	−0.43 (.15)	.05	0.65 [0.47, 0.87]	0.03 (.16)	.83	1.03 [0.75, 1.42]

Note. All *p*-values in the regression models were adjusted using Benjamini-Hochberg false discovery rate.
EMS = Eyewitness Metamemory Scale, MMQ = Multifactorial Metamemory Questionnaire, OR = Odds ratio, SSMQ = Squire Subjective Memory Questionnaire.
***$p < .001$.

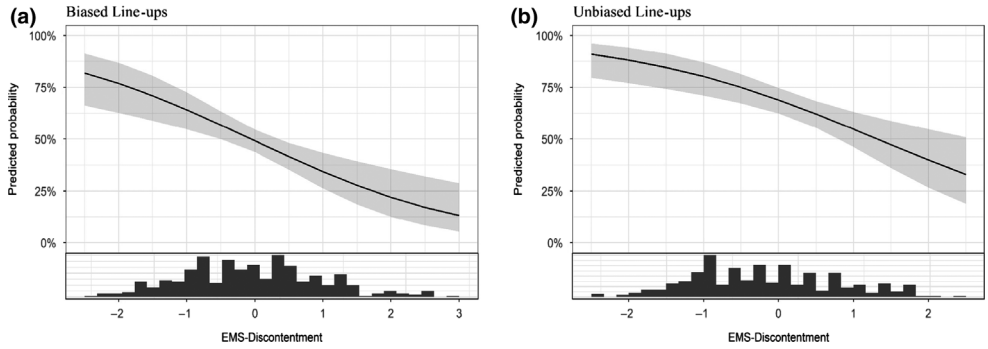

Figure 1. Predicted probabilities of identification accuracy among choosers as a function of EMS-Discontentment for biased and unbiased line-ups. The shaded polygon represents 95% confidence intervals.

underconfidence, and observations below this line indicate overconfidence. We computed three calibration statistics: calibration index, over/underconfidence, and resolution (see Brewer & Wells, 2006). Calibration (*C*) represents how far a given calibration curve is from a perfect calibration. It ranges from 0 (perfect calibration) to 1, and lower values represent better calibration. Over/underconfidence (*O/U*) indicate if a curve strays more above or below the perfect calibration line, with values ranging from −1 (very underconfident) to 1 (very overconfident). The Normalized Resolution Index (*NRI*) represents how well confidence discriminates accurate from inaccurate identifications, ranging from 0 = no discrimination to 1 = perfect discrimination. *NRI* is equivalent to η^2 in a one-way analysis of variance and can be interpreted as the percentage of variance of the outcome variable accounted for by confidence judgements, so *NRI* values of 0.01, 0.06, and 0.14 can be interpreted as small, medium, and large effects, respectively (Cohen, 1988; Yaniv, Yates, & Smith, 1991). Following Palmer, Brewer, Weber, and Nagesh (2013), we used a jackknife procedure to compute standard errors for each calibration statistic, which were then converted to 95% inferential confidence intervals (Tryon, 2001). If the confidence intervals do not overlap, that represents a significant difference. For choosers, the resolution statistic showed a high capability to discriminate between accurate and

Table 6. Logistic regression models of metamemory factors as predictors of identification accuracy among non-choosers

	Biased line-ups			Unbiased line-ups		
Predictor	*B* (SE)	*p*	OR [95% CI]	*B* (SE)	*p*	OR [95% CI]
EMS-contentment	−0.23 (.18)	.51	0.79 [0.55, 1.13]	−0.21 (.17)	.52	0.81 [0.57, 1.14]
EMS-discontentment	−0.26 (.16)	.43	0.77 [0.56, 1.05]	−0.31 (.16)	.36	0.73 [0.53, 1.01]
EMS-strategies	0.06 (.14)	.75	1.06 [0.80, 1.41]	0.09 (.13)	.73	1.10 [0.84, 1.43]
MMQ-contentment	0.36 (.16)	.20	1.43 [1.03, 1.99]	−0.20 (.19)	.56	0.81 [0.55, 1.19]
MMQ-ability	−0.12 (.15)	.70	0.88 [0.65, 1.20]	0.09 (.17)	.73	1.09 [0.77, 1.54]
MMQ-strategy	0.05 (.14)	.75	1.06 [0.79, 1.41]	0.13 (.16)	.67	1.14 [0.84, 1.56]
SSMQ	−0.09 (.17)	.73	0.91 [0.64, 1.28]	0.11 (.18)	.73	1.11 [0.78, 1.59]

Note. All *p*-values in the regression models were adjusted using Benjamini-Hochberg false discovery rate. OR = Odds ratio. EMS = Eyewitness Metamemory Scale. MMQ = Multifactorial Metamemory Questionnaire. SSMQ = Squire Subjective Memory Questionnaire.

Table 7. Logistic regression models of metamemory factors and confidence as predictors of identification accuracy among choosers

	Biased line-ups			Unbiased line-ups		
Predictor	*B* (SE)	*p*	OR [95% CI]	*B* (SE)	*p*	OR [95% CI]
Confidence	0.82 (.01)	<.001***	2.27 [1.80, 2.99]	0.76 (.12)	<.001***	2.21 [1.69, 2.77]
EMS-contentment	−0.16 (.14)	.25	0.84 [0.63, 1.12]	−0.46 (.17)	.005**	0.62 [0.44, 0.87]
EMS-discontentment	−0.49 (.14)	<.001***	0.60 [0.45, 0.80]	−0.62 (.15)	<.001***	0.53 [0.39, 0.71]
EMS-strategies	−0.01 (.13)	.93	0.98 [0.75, 1.28]	−0.07 (.13)	.55	0.92 [0.71, 1.20]
MMQ-contentment	0.26 (.16)	.11	1.30 [0.94, 1.80]	−0.16 (.15)	.30	0.84 [0.62, 1.15]
MMQ-ability	0.06 (.15)	.69	1.06 [0.78, 1.44]	−0.03 (.15)	.83	0.96 [0.71, 1.32]
MMQ-strategy	0.27 (.13)	.04	1.31 [1.01, 1.72]	0.23 (.14)	.10	1.26 [0.95, 1.67]
SSMQ	−0.44 (.16)	<.001***	0.63 [0.45, 0.87]	0.01 (.17)	.80	1.04 [0.74, 1.46]

Note. OR = Odds ratio. EMS = Eyewitness Metamemory Scale. MMQ = Multifactorial Metamemory Questionnaire. SSMQ = Squire Subjective Memory Questionnaire. All *p*-values in the regression models were adjusted using Benjamini-Hochberg false discovery rate.
**$p < .001$; ***$p < .001$.

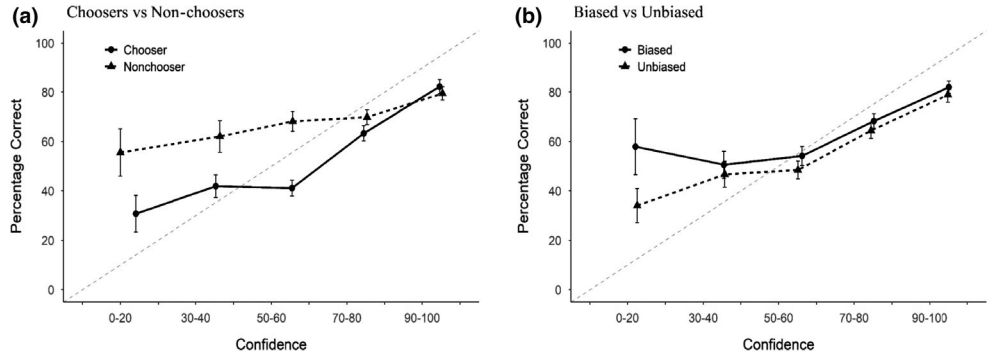

Figure 2. Confidence–accuracy calibration curves comparing choosers and non-choosers (a) and biased versus unbiased line-ups (b). The dotted diagonal grey line represents perfect calibration. The points on the curves are positioned in the mean confidence of the respective confidence group.

inaccurate identification decisions, for both biased (*NRI* = 0.12) and unbiased line-ups (*NRI* = 0.11; see Table 8). However, choosers tended to be more overconfident in unbiased line-ups (*O/U* = 0.14) compared to biased line-ups (*O/U* = 0.03).

Next, we compared calibration statistics between high and low scorers on each of the metamemory measures. Following Olsson and Juslin (1999), individuals above the 66th percentile were selected as high scorers and individuals below the 33th percentile as low scorers. For this analysis, we focus on choosers, because triers of fact are more specifically concerned with eyewitnesses that choose someone from a line-up, rather than eyewitnesses that reject a line-up (Mickes, 2015; Wixted & Wells, 2017). Each metamemory group (low scorers and high scorers) had a mean sample size of n = 271. Inspection of the confidence intervals suggested that low scorers in the EMS-Contentment, EMS-Discontentment, EMS-Strategies, and SSMQ were significantly less overconfident than higher scorers in those components (see Figure 3). The calibration curves for those measures reveal that lower scorers were generally better calibrated than high scorers, especially for higher levels of confidence (see Figure 4). A similar pattern of results was observed for both biased and unbiased line-ups (see supplemental materials).

Discussion

We investigated the diagnostic value of self-ratings of memory efficacy on eyewitness identification accuracy and confidence, examining the relationship between memory self-

Table 8. Calibration statistics for choosers and non-choosers across biased and unbiased conditions

	C [95% CI]	O/U [95% CI]	NRI [95% CI]
Biased line-ups			
Choosers	0.01 [0.01, 0.02]	0.03 [−0.01 0.07]	0.12 [0.06, 0.18]
Non-choosers	0.03 [0.01, 0.04]	0.04 [−0.01, 0.09]	0.02 [−0.01, 0.04]
Unbiased line-ups			
Choosers	0.03 [0.01, 0.04]	0.14 [0.09, 0.18]	0.11 [0.05, 0.17]
Non-choosers	0.03 [0.01, 0.04]	−0.03 [−0.08, 0.01]	0.04 [−0.01, 0.09]

Note. C = Calibration index, NRI = Normalized resolution index, O/U = Over/underconfidence index.

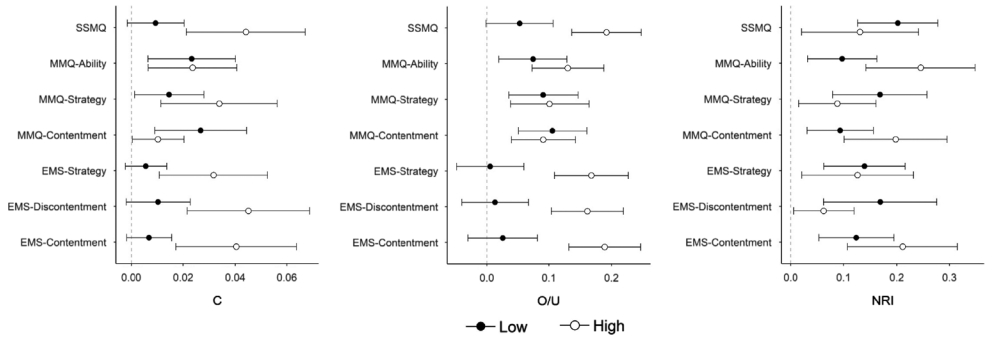

Figure 3. Inferential confidence intervals of calibration statistics for high and low scorers in each metamemory measure for all line-ups.

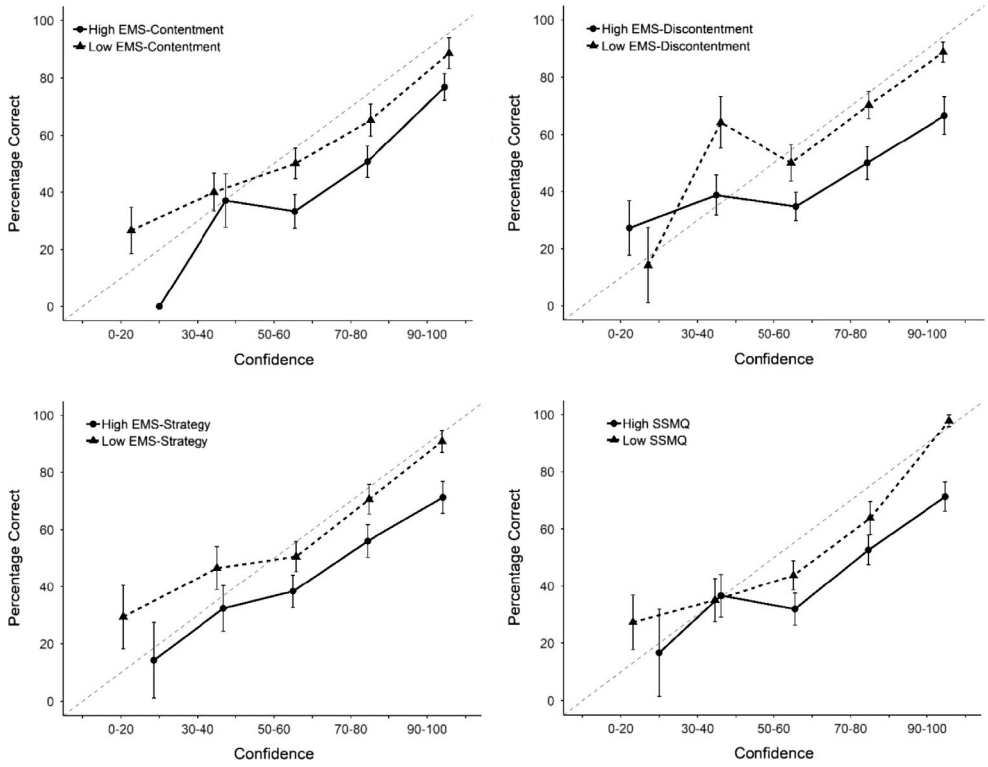

Figure 4. Calibration curves (all line-ups) comparing low and high scorers in the EMS-Contentment, EMS-Discontentment, EMS-Strategy and SSMQ metamemory factors. The points on the curves are positioned in the mean confidence of the respective confidence group.

efficacy and identification performance for both biased and unbiased line-ups. Our results revealed two key findings. First, higher discontentment with face recognition and person identification efficacy (EMS-Discontentment) was indicative of more inaccurate identifications for choosers in both biased and unbiased line-ups. Second, low scorers in EMS-Contentment, EMS-Discontentment, EMS-Strategies, and SSMQ were less overconfident and were generally better calibrated than high scorers, especially for higher levels of

confidence. These findings contribute to an ongoing debate concerning the relationship between behavioural and self-reported face recognition efficacy. While some research suggests that individuals have only limited insight into their own face recognition efficacy (Bindemann *et al.*, 2014; Bobak *et al.*, 2018), other studies report that self-ratings of face recognition efficacy are moderately to strongly related to objective performance (Livingston & Shah, 2018; Ventura *et al.*, 2018).

Focusing specifically on eyewitness-identification paradigms, the current research provides initial evidence for a relation between self-reported memory discontentment and accuracy in line-up identification settings. Most notably, we expected the relation between self-ratings of memory efficacy and identification performance to be weaker in biased line-ups compared to unbiased line-ups, but this relation was similar for both conditions. This finding has important implications given that other post-dictors of eyewitness identification performance are undermined in identifications made on biased line-ups (Charman *et al.*, 2011; Key *et al.*, 2017). In other words, although confidence and decision time have reduced diagnostic value of accuracy in biased line-ups, the same is not true for self-ratings of eyewitness memory efficacy. Charman *et al.* (2011) demonstrate that biased line-ups reduce the diagnostic value of confidence because confidence is inflated when the line-up target is compared with implausible fillers. The authors also propose a scaling effect explanation for this finding, based on the fact that witnesses must generate anchor points when providing a similarity score between two faces on a subjective scale (such as a 1 to 7 scale). During an identification, these anchor points may be affected by external factors, such as the dissimilarity between fillers and the target. In the case of self-ratings of memory efficacy, it is less likely that broader ratings (i.e., 'how good is your memory for faces?') will be affected by situational factors such as filler dissimilarity. An individual who often distrusts their ability to recognize unfamiliar faces is unlikely to change this self-assessment when exposed to a biased or unbiased line-up. Therefore, specific self-ratings of eyewitness memory efficacy may be useful estimators of accuracy independently of line-up fairness. If replicated, this finding may have important applied implications given the practical difficulties in producing unbiased line-ups without computerized systems (Memon *et al.*, 2011).

Another goal of the current study was to further investigate the relation between self-ratings of memory efficacy and eyewitness confidence–accuracy relationship. Lower scores in all eyewitness metamemory factors (i.e., EMS-Contentment, EMS-Discontentment, and EMS-Strategies) were indicative of a stronger confidence–accuracy relation among choosers, while higher scores in these factors were related to more overconfidence. This finding indicates that individuals who do not hold overly strong positive or negative opinions about their face recognition ability (low EMS-Contentment and low EMS-Discontentment) are better calibrated when reporting their confidence, while individuals with a stronger opinion (i.e., either for low or high memory ability) tend to exaggerate their confidence assessments. Both individuals with high EMS-Contentment and high EMS-Discontentment tended to be overconfident in their identifications. In contrast, Olsson and Juslin (1999) observed that individuals who rated themselves as good face recognizers had a more diagnostic confidence–accuracy relationship. However, in that study the authors acknowledge as a limitation having used single items of unknown validity and reliability, so inferences of memory self-efficacy from such a measure may be limited. The current data support the notion that individuals highly content with their own memories tend to exaggerate their confidence (Rickenbach, Agrigoroaei, & Lachman, 2015). The relation between higher discontentment and higher overconfidence seems less straightforward. One possible explanation for this result is that choosers who

are generally discontent with their own memories may overestimate their confidence precisely because they have selected someone from a line-up. In other words, if an individual is discontent with their memory efficacy, but nevertheless select someone from a line-up, the selection may be followed by inflated confidence. Finally, individuals who claimed to endorse more memory strategies to encode faces were also more likely to be overconfident, possibly because those individuals 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.

Our prediction that eyewitness-specific metamemory factors would have a stronger relation to identification performance compared to general metamemory factors was somewhat supported. We come to this conclusion because EMS-Discontentment was the strongest predictor in the models testing metamemory factors as predictors of identification accuracy among choosers. This pattern of results supports the utility of domain-specific memory self-efficacy, defined as an individual's appraisal of their usual efficacy in a given memory domain (Hertzog & Dixon, 1994). Our findings also suggest that assessments of self-efficacy focused on eyewitness-specific domains (e.g., face and person identification) are more valuable than assessments of *general* memory efficacy in distinguishing accurate from inaccurate identifications among choosers.

The current study has a number of limitations. First, although we tested two different targets in our eyewitness paradigm, we only used one mock-crime video. We reasoned that the inclusion of multiple target events could generate noise and affect the power of our analyses, so it remains to be determined whether the current findings would replicate when assessing witness performance for different types of target events. Second, the metamemory assessment occurred prior to the line-up identification tasks. In planning our procedure, we reasoned that exposure to the identification tasks before the completion of the metamemory assessment would have affected self-ratings of memory efficacy to a greater extent than completing the assessments would affect eyewitness performance (Olsson & Juslin, 1999). This may have been appropriate for the aims of the current study, but future investigation should examine the robustness of self-rated memory efficacy as predictors of eyewitness performance when measures are obtained after the identification tasks. Additionally, the fact that the EMS was always completed before the other general metamemory questionnaires may have influenced responses in the general metamemory questionnaires. We decided to always present the EMS questionnaire first because the data used in this study were also used for testing the development and validity of this scale, so including other measures before the EMS could have impaired its development. However, a recent study adopting counterbalancing procedures shows no evidence of order effects regarding the presentation of eyewitness-specific and general metamemory measures (Saraiva, van Boeijen, *et al.*, 2019; Saraiva *et al.*, 2019).

Taken together, our findings contribute to the ongoing challenge of distinguishing accurate from inaccurate eyewitness identifications in the criminal justice system. We present initial evidence that choosers who report higher discontentment with their face and person identification efficacy are more likely to commit a false identification. Further work is necessary to determine the generalizability of these results to different target events and for metamemory assessments obtained after the identification tasks. Furthermore, metamemory assessments can increase the diagnostic value of confidence, given the observation that individuals with stronger opinions about their face recognition efficacy tend to be overconfident. This is of importance because confidence statements are often used to discriminate accurate from inaccurate witnesses, but little is known

about whether confidence statements are affected by individual differences related to self-efficacy.

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.

Conflicts of interest

All authors declare no conflict of interest.

Author contribution

Renan Saraiva, M.D. (Conceptualization, Data curation, Formal analysis, Funding acquisition, Project administration, Resources, Writing – original draft, Writing – review & editing); Inger van Boeijen (Data curation, Formal analysis, Resources, Writing – review & editing); Lorraine Hope, Robert Horselenberg, Melanie Sauerland and Peter J. van Koppen (Supervision, Writing – review & editing).

Data availability statement

The data, analysis code, and pre-registration of this study can be found on the following Open Science Framework repository: https://osf.io/ymkz9/?view_only=49c11c762050470fbe45880af51512ee

References

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. Retrieved from http://arxiv.org/abs/1406.5823.

Beaudoin, M., & Desrichard, O. (2011). Are memory self-efficacy and memory performance related? A meta-analysis. *Psychological Bulletin*, *137*, 211–241. https://doi.org/10.1037/a0022106

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*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bergen, S., Brands, I., & Jelicic, M. (2010). Assessing trait memory distrust: Psychometric properties of the Squire Subjective Memory Questionnaire. *Legal and Criminological Psychology*, *15*, 373–384. https://doi.org/10.1348/135532509X471960

Bindemann, M., Attard, J., & Johnston, R. A. (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology*, *1*(1), 986903. https://doi.org/10.1080/23311908.2014.986903

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012/2016). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*, 96–103. https://doi.org/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, *72*(4), 872–881. https://doi.org/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. https://doi.org/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. https://doi.org/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, NY: Springer.

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, *22*, 273–280. https://doi.org/10.3758/BF03200854

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. https://doi.org/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. https://doi.org/10.1080/09658211.2011.633919

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

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory*, *26*, 741–750. https://doi.org/10.1080/09658211.2017.1404111

Dunlosky, J., & Bjork, R. A. (2008). The integrated nature of metamemory and memory. In J. Dynlosky, & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 11–28). New York, NY: Psychology Press.

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/xlm0000332

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition*, *8*, 233–243. https://doi.org/10.1016/j.jarmac.2019.02.002

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*, 160923. https://doi.org/10.1098/rsos.160923

Hertzog, C., & Dixon, R. A. (1994). Metacognitive development in adulthood and old age. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 227–251). Cambridge, UK: MIT Press.

Higham, P. A., & Vokey, J. R. (2000). Judgment heuristics and recognition memory: Prime identification and target-processing fluency. *Memory & Cognition*, *28*, 574–584. https://doi.org/10.3758/BF03201248

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology. General*, *118*, 126–135. https://doi.org/10.1037/0096-3445.118.2.126

Key, K. N., Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Cash, D. K., & Lane, S. (2017). Line-up fairness affects postdictor validity and "don't know" responses. *Applied Cognitive Psychology*, *31*, 59–68. https://doi.org/10.1002/acp.3302

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, *9*, 149–171. https://doi.org/10.1006/ccog.2000.0433

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*, 36–69. https://doi.org/10.1037/0096-3445.135.1.36

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.), *A handbook of memory and metamemory* (pp. 117–136). New York, NY: Psychology Press.

Leippe, M. R., & Eisenstadt, D. (2014). Eyewitness confidence and the confidence-accuracy relationship in memory for people. InR. C. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology: Volume II: Memory for people* (pp. 377–425). New York, NY: Routledge.

Livingston, L. A., & Shah, P. (2018). People with and without prosopagnosia have insight into their face recognition ability. *Quarterly Journal of Experimental Psychology*, *71*, 1260–1262. https://doi.org/10.1080/17470218.2017.1310911

Memon, A., Havard, C., Clifford, B., Gabbert, F., & Watt, M. (2011). A field evaluation of the VIPER system: A new technique for eliciting eyewitness identification evidence. *Psychology, Crime & Law*, *17*, 711–729. https://doi.org/10.1080/10683160903524333

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. https://doi.org/10.1016/j.jarmac.2015.01.003

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. https://doi.org/10.1037/0021-9010.84.1.42

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. https://doi.org/10.1037/a0031602

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. https://doi.org/10.1002/acp.952

Prims, J., & Motyl, M. (2018). A tool for detecting low quality data in internet research. GitHub: https://github.com/SICLab/detecting-bots.

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. https://doi.org/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. https://doi.org/10.1186/s41235-018-0121-8

Saraiva, R. B., Van Boeijen, I. M., Hope, L., Horselenberg, R., Sauerland, M., & Van Koppen, P. J. (2019). Development and validation of the Eyewitness Metamemory Scale. *Applied Cognitive Psychology*, *33*(5), 964–973. https://doi.org/10.1002/acp.3588.

Saraiva, R. B., van Boeijen, I. M., Hope, L., Horselenberg, R., & van Koppen, P. (2019). *Using general and eyewitness-specific metamemory assessments to estimate performance in multiple identifications*. Doi: https://doi.org/10.31219/osf.io/tay75.

Sauer, J. D., Brewer, N., & Wells, G. L. (2008). Is there a magical time boundary for diagnosing eyewitness identification accuracy in sequential line-ups? *Legal and Criminological Psychology*, *13*, 123–135. https://doi.org/10.1348/135532506X159203

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

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. https://doi.org/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. https://doi.org/10.1037/a0014560

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, NY: 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. https://doi.org/10.1037/0882-7974.11.3.538

Shah, P., Sowden, S., Gaule, A., Catmur, C., & Bird, G. (2015). The 20-item prosopagnosia index (PI20): Relationship with the Glasgow face-matching test. *Royal Society Open Science*, *2*, 150305. https://doi.org/10.1098/rsos.150305

Smith, S. M., Lindsay, R., & Pryke, S. (2001). Postdictors of eyewitness errors: Can false identifications be diagnosed in the cross-race situation? *Psychology, Public Policy, and Law*, *7*, 153–159. https://doi.org/10.1037/1076-8971.7.1.153

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

Squire, L. R., Wetzel, C. D., & Slater, P. C. (1979). Memory complaint after electroconvulsive therapy: Assessment with a new self-rating instrument. *Biological Psychiatry*, *14*, 791–801.

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior*, *22*, 217–237. https://doi.org/10.1023/A:1025746220886

Tredoux, C. G. (1999). Statistical considerations when determining measures of lineup size and lineup bias. *Applied Cognitive Psychology*, *13*, 9–26. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+<S9:AID-ACP634>3.0.CO;2-1

Troyer, A. K., & Rich, J. B. (2002). Psychometric properties of a new metamemory questionnaire for older adults. *The Journals of Gerontology*, *57*, 19–27. https://doi.org/10.1093/geronb/57.1.P19

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. https://doi.org/10.1037//1082-989X.6.4.371

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, IL. Cambridge, UK: 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. https://doi.org/10.1037/0882-7974.21.2.165

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*(12), 2677–2679. https://doi.org/10.1177/1747021818765652

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*, 603–647. https://doi.org/10.1023/A:1025750605807

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 198–215. https://doi.org/10.1037/0278-7393.30.1.198

Wixted, J. T., Mickes, L., & Fisher, R. P. (2018). Rethinking the reliability of eyewitness memory. *Perspectives on Psychological Science*, *13*, 324–335. https://doi.org/10.1177/1745691617734878

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. https://doi.org/10.1177/1529100616686966

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

Received 5 July 2019; revised version received 8 January 2020

Supporting Information

The following supporting information may be found in the online edition of the article:

Table S1. Model Fit Indices for Fixed Intercept Models (Model 1) and Random Intercept Models (Model 2) for Each Condition

Figure S1. Inferential confidence intervals of calibration statistics for high and low scorers in each metamemory measure for biased lineups.

Figure S2. Calibration curves comparing low and high scorers in the EMS-Contentment, EMS-Discontentment, EMS-Strategy and SSMQ metamemory factors for biased lineups.

Figure S3. Inferential confidence intervals of calibration statistics for high and low scorers in each metamemory measure for unbiased lineups.

Figure S4. Calibration curves comparing low and high scorers in the EMS-Contentment, EMS-Discontentment, EMS-Strategy and SSMQ metamemory factors for unbiased lineups.