Received: 3 February 2022 | Accepted: 29 September 2022

DOI: 10.1111/lcrp.12232

the british psychological society
promoting excellence in psychology

REGISTERED REPORT STAGE 2

Predicting and projecting memory: Error and bias in metacognitive judgements underlying testimony evaluation

Rebecca K. Helm | **Bethany Growns**

University of Exeter Law School, Exeter, UK

Correspondence
Rebecca K. Helm, University of Exeter Law School, St Luke's Campus, Heavitree Road, Exeter EX1 2LT, UK.
Email: r.k.helm@exeter.ac.uk

Funding information
UK Research and Innovation fellowship, Grant/Award Number: MR/T02027X/1

Abstract

Purpose: Metacognitive judgements of what another person would remember had they experienced a stimulus—that is social metamemory judgements, are likely to be important in evaluations of testimony in criminal and civil justice systems. This paper develops and tests predictions about two sources of error in social metamemory judgements that have the potential to be important in legal contexts—errors resulting from beliefs informed by own memory being inappropriately applied to the memory of others, and errors resulting from differential experience of an underlying stimulus.

Method: We examined social metamemory judgements in two experimental studies. In Experiment 1 (N = 323), participants were required to make either social metamemory judgements relating to faces or predictions relating to their own memory for faces. In Experiment 2 (N = 275), we manipulated participant experience of faces, holding the described experience of the person whose memory was being assessed constant and asked participants to make social metamemory judgements.

Results: As predicted, judgements relating to the memory of others were prone to inaccuracy. Whilst participants making predictions relating to their own memory performed above chance, participants making social metamemory

judgements performed no better than chance. Social metamemory judgements were also influenced by the way stimuli were experienced by an assessor, even where this experience did not correspond to the experience of the person whose memory they were assessing.

Conclusions: Having our own experiences of memory does not necessarily make us well-placed to assess the memory of others, and, in fact, our own experiences of memory can even be misleading in making judgements about the memory of others.

KEYWORDS

judgements of learning, juror decision making, memory, metacognition, perspective taking, social metamemory

BACKGROUND

Judgements about what a witness would be expected to remember if an event had occurred as alleged are important for the legal system. For example, imagine a case in which a witness cannot identify a defendant as the person they saw committing a crime. Finders of fact will have to judge whether they would expect the witness to remember the defendant had the defendant committed the crime. Or imagine a person alleging that they have been the victim of a crime but who cannot remember details about the setting they allege that the crime occurred in. An important question in determining whether the lack of memory for details undermines their credibility is what we would expect them to remember had the crime occurred as alleged. These judgements are a form of social metamemory judgement (see Tullis & Fraundorf, 2017). The importance of these judgements legally was demonstrated in the recent trial of Ghislaine Maxwell in which, at least allegedly, a topic of debate among jurors was whether victims of assault would be likely to remember certain details relevant to the offences (Gregory, 2022). Although existing research provides significant insight into the related question of what qualities people expect accurate testimony to contain (Bell & Loftus, 1989; Berman et al., 1995; Brewer & Burke, 2002; Semmler & Brewer, 2002; Wells & Leippe, 1981), no existing work in psychology and law has examined social metamemory judgements.

Knowing how social metamemory judgements are made and about likely inaccuracies in them has the potential to contribute to our understanding of evaluations of witness testimony, including defendant and complainant testimony. It has the potential to help explain existing findings in the literature (e.g. that people tend to believe accurate testimony contains good memory for peripheral details; Bell & Loftus, 1989; Wells & Leippe, 1981), and, importantly, to demonstrate ways that evaluations of testimony may be based on flawed assumptions about what others are likely to remember (which may help explain observed weaknesses in evaluations, e.g. Brigham & Bothwell, 1983; Kaminski & Sporer, 2018). This paper highlights the importance of social metamemory judgements for law and reports the results of two experiments that tested predictions relating to systematic errors in these judgements in the legal context that are likely to influence legal outcomes. Experiment 1 examined the potential for errors resulting from generalizations from one's own beliefs about memory and experience of a stimulus to others, even where the stimulus is presented in the same way to the assessor and the person whose memory is being assessed. Experiment 2 examined a way that these errors are likely to be exacerbated in legal contexts as a result of relevant stimuli being presented to finders of fact in ways that are different from how those stimuli were experienced by those whose memory they are assessing.

Metamemory judgements and 'social' metamemory

The majority of existing work on metamemory examines people's insight into what they will remember from a stimulus that they have experienced, known as judgements of learning. This work has examined judgements in the context of memory for words on word lists (Koriat et al., 2004), unfamiliar faces (Nguyen et al., 2018; Sommer et al., 1995) and recorded events (Dutton & Carroll, 2001). It provides insight into people's accuracy in assessing what they will remember (Rhodes, 2006), their sensitivity to factors likely to influence their own memory (Nguyen et al., 2018), and the cognitive processes that people draw on to make these judgements (Frank & Kuhlmann, 2017). Importantly, the work suggests that two separate classes of information inform these judgements—people's experience processing a stimulus (known as experience-based cues, Alter & Oppenheimer, 2009; Undorf & Erdfelder, 2011) and beliefs or theories about memory (known as belief-based cues, Frank & Kuhlmann, 2017; Koriat, 1997). Experience-based cues, such as processing fluency (the ease with which a stimulus is processed), incorporate factors into evaluations that are outside of conscious awareness. They improve the accuracy of judgements where the cue influencing the experience does in fact influence memory (Undorf & Erdfelder, 2015). However, they can also be misleading and worsen accuracy where the cue influencing experience does not in fact influence memory (Besken, 2016) or when they distract the person making a judgement from applying relevant beliefs and theories (Castel, 2008).

Judgements of learning are detached from social metamemory judgements relevant to the legal system since judgements of learning are made about one's own memory, and social metamemory judgements are made about the memory of others. However, the small amount of existing research examining social metamemory suggests that similar mechanisms underly social metamemory as underly traditional judgements of learning. Social metamemory, like judgements of learning, can be informed by both belief-based and experience-based cues (Koriat & Ackerman, 2010). Essentially, people draw on their experiences of a stimulus to make judgements about what others are likely to remember from it (see also Paulus et al., 2014). Provided that people making predictions have had an experience of an underlying stimulus, their personal experiences of the stimulus are likely to be important in assessing what other people will remember from it.

In one study, researchers found that people were only able to make accurate social metamemory judgements when they first experienced relevant stimuli and made judgements of learning relating to themselves (Koriat & Ackerman, 2010). However, social metamemory judgements are more susceptible to error than judgements of learning since social metamemory judgements are made based on the experiences and beliefs of one person but about the memory of another person. In this context, errors can occur as a result of misleading information, where people rely on cues relating to memory that may be related to their own memory performance, but which are unrelated to memory performance in the person whose memory they are judging (see Birch, 2005). Errors can also occur as a result of inadequate information, having insufficient information about another person, for example relating to their memory function and experience of a stimulus (Tullis & Fraundorf, 2017), which might influence the ability to make judgements about their memory on an individualized basis (see Lovelace, 1984). Experimental research in the context of the utility of mnemonic cues confirmed that, at least in that context, whilst social metamemory judgements are more accurate than chance, they are significantly less accurate than judgements of learning (Tullis & Fraundorf, 2017). However, to date, no research has examined social metamemory in a legally relevant context or investigated sources of error in social metamemory judgements in that context.

Social metamemory in the legal context

Social metamemory judgements, as well as related metacognitive judgements, are likely to be important when finders of fact are making determinations about the probative value that they can extract from witness testimony, specifically when they are deciding what the presence or absence of memory in a

witness can tell us about a crime or other alleged wrong. For example, as discussed above, judgements of whether a witness would be likely to remember a particular defendant had they seen them are likely to be important in determining what a lack of memory for that defendant in a witness tells us. If we expect that a witness who saw an offender commit a crime would remember that offender, a lack of memory for the defendant in that witness points towards the defendant not having been the offender (although of course, not conclusively). Importantly, finders of fact have often seen the stimuli that they are judging memory for, in court. For example, they will see the face of the defendant or a picture of a crime scene and will know objectively whether a witness remembered the defendant or accurately remembered details of the crime scene. They experience the stimuli (albeit in a limited way) themselves and then make a judgement about what a witness would be likely to remember had they experienced it.

In this context, there are two important risks with the potential to undermine the quality of metamemory judgements (over and above errors with traditional judgements of learning shown in other work, e.g., Nguyen et al., 2018). Both of these risks arise from people relying on their own experience of memory to judge the memory of others, something often recommended by legal systems (*R v JH and TG*, 2005, in the UK context). First, as discussed above, social metamemory judgements are susceptible to inaccuracy because they are made about other people, who assessors may lack information about and who may have different memory functions and experiences of stimuli than assessors do. People's beliefs about memory are likely to result from their own personal experiences of memory in the past (e.g. how their own memory has endured over time, or how often they can accurately remember the faces of people they have seen) and how memorable they themselves find specific stimuli, as well as other information presented to them (e.g. from the media or through education). As a result, whilst these beliefs may be accurate, and likely to facilitate accurate judgement, when applied to own memory, they may be misleading when applied to the memory of others, particularly where that memory differs from own memory in meaningful ways. The prediction that social metamemory judgements will be less accurate than judgements of learning, and that this inaccuracy will result at least partly from drawing on beliefs informed by own memory in making assessments, is examined in Experiment 1.

In the legal context, social metamemory judgements are also susceptible to another type of error because relevant stimuli are presented to finders of fact entirely differently from how they have been experienced by the person whose memory the finders of fact are assessing. For example, whilst a witness may have seen a person or crime scene for a short time with many distractions, a juror may see the person or crime scene for a longer period with no distractions. Expectations relating to the memorability of the stimulus may be subconsciously influenced by aspects of the experience of the finder of fact (experience-based cues), including those with no probative value when related to the original witness. For example, a juror who views a defendant for a long time may find that defendant more memorable than they would have if they had only seen them for a short time, and this increased memorability may increase their expectations that a witness who saw the defendant at a crime scene would remember them (even if that witness only saw the defendant for a short time). The prediction that an assessor's experience of a stimulus will influence social metamemory judgements even where this experience has no relevant probative value will be examined in Experiment 2.

Overview of study experiments and preregistration

The two experiments that we present in this paper investigated social metamemory judgements relevant to the legal system. Both experiments were submitted in a registered report and were preregistered prior to data collection. All data, materials and analysis scripts are available on OSF at https://osf.io/ms3f4. The pre-registration for Experiment 1 can be viewed at: https://osf.io/vk9we. The pre-registration for Experiment 2 can be viewed at: https://osf.io/cr8bj.

In Experiment 1, we examined both judgements of learning and social metamemory judgements about faces to test the predictions that social metamemory would be less accurate (and, relatedly, more biased), than judgements of learning in the context of identifications (Hypothesis 1), partly as a result

of assessors relying on their own experiences of memory, which may not be appropriate to apply to the memory of others (Hypothesis 2). We tested these predictions by comparing the accuracy of judgements of learning and social metamemory judgements and by examining whether the accuracy of social metamemory judgements improved when participants made judgements about others who shared a meaningful characteristic with them, making generalization more appropriate. We focused on memory abilities as one meaningful characteristic with the potential to make generalizations more appropriate. Note that matching on this feature only was expected to improve judgements, but not to bring accuracy to the level of judgements of learning, since participants were only matched based on one characteristic and not other characteristics or experiences of the stimuli.

In Experiment 2, we examined how errors in social metamemory might be exacerbated in a legal scenario, where the people assessing the memory of others are presented with the relevant stimulus differently from how that stimulus was experienced by the person whose memory they are assessing. We altered the experience of stimuli—faces—for assessors by manipulating the size of and length of exposure to (two experience-based cues shown to influence judgements of learning in the context of words, see Koriat & Ma'ayan, 2005; Price & Harrison, 2017) each face whilst keeping the described experience of the person whose memory for the person pictured was being assessed constant. In this design, we tested the prediction that changes in how the assessor experienced the face would impact their social metamemory judgements for the person they were assessing, despite having no probative value in those social metamemory judgements. Note that we are not implying in this study that size and length of exposure would not influence memory, but that the size and length of exposure of the assessor (the participant here) are not relevant to the memory of a witness where they are not related to the witness's own experience (e.g. if a witness saw a person for 15 s, the judgement of whether they would remember them should be based on this 15 s, and so should not be influenced by whether the person assessing the witness's memory saw the relevant stimulus for 5 s or 5 m).

EXPERIMENT 1

Method

Participants

Experiment 1 included 323 participants recruited from the Prolific survey platform (a platform shown to produce high-quality data and to filter out bots and inattentive participants see Peer et al., 2017). This sample size was chosen based on an a priori power analysis for detecting a medium effect ($d = 0.3$) in an independent samples t-test with 95% power using the *WebPower* package in *R* (suggested $N = 290$; Zhang & Mai, 2018), plus 10% to account for any attrition ($N = 319$). This sample size was increased very slightly to ensure sufficient power for Analysis 2 of our analysis plan (where we halve our sample and compare two scores for each participant within subjects, see below). Sample size for Analysis 2 is based on an a priori power analysis for detecting a medium effect ($d = 0.3$) in a paired samples t-test with 95% power, again using the *WebPower* package in *R* ($N = 146$; Zhang & Mai, 2018), plus 10% to account for any attrition ($N = 161$, note a total sample size of 322 is required to attain a half sample size of 161).[1] These sample sizes were expected to be sufficient in our mixed-effects design since our focus is on fixed effects in a regression model and individuals/trials are not expected to be clustered within groups.

To be eligible for the study, participants were required to live in the United Kingdom, to have normal or corrected-to-normal vision, to have a Prolific approval rating of at least 90% and to have completed at least 10 previous Prolific submissions. All participants were required to complete the experiment on

[1]Note that one additional participant was able to take our survey due to a technical error on the Prolific platform, and that participant was included in analyses.

a computer (not a cellular device or tablet). No participants were excluded based on our pre-registered exclusion criteria of failing two or more attention check questions (out of five total questions which were randomly distributed through the survey; 4.64% correctly answered four questions and 95.36% correctly answered all five questions). Participants in the final sample (N = 323) had an average age of 37.53 years (SD = 13.22, range = 18–75), and about half (49.85%) self-identified as female (48.92% as male and 1.24% as gender diverse). Each participant was compensated £5 for completing the approximately 30 minute experiment.

Design

The experiment had two experimental conditions (a 'self' condition and an 'other' condition). Participants in the self condition made predictions about their own memory (traditional judgements of learning) for faces, and participants in the other condition made predictions about the memory of another person (social metamemory judgements) for the same faces. Analyses examined metamemory accuracy, measured by proportion correct, sensitivity (d') and response bias (C) (i.e. Type 2 Signal Detection Theory; Fleming & Lau, 2014; Stanislaw & Todorov, 1999).

To calculate our signal detection measures (d' and C), we defined hits, misses, false alarms and correct rejections as follows (information relating to the other condition rather than the self condition is provided in brackets). We defined hits as trials where participants predicted they (another person) would remember a face and they (the other person) did remember the face, misses as trials where participants predicted they (another person) would not remember a face and they (the other person) did remember the face, false alarms as trials where participants predicted they (another person) would remember a face but they (the other person) did not remember the face, and correct rejections as trials where participants predicted they (another person) would not remember a face and they (the other person) did not remember the face. Higher d' values indicate higher sensitivity to whether an item would or would not be remembered independent of response bias (a tendency to respond 'yes' or 'no') with higher values indicating higher 'accuracy' in the task. Positive C values indicate an increased tendency to respond 'no', whilst negative C values indicate an increased tendency to respond 'yes'.

Materials and procedure

The experiment was hosted on Qualtrics, using face stimuli from a database of male faces used in previous work (Meissner et al., 2005; Nguyen et al., 2018). This database provides two headshots for each person, differing in facial expression and clothing. Each participant completed a study phase and then a subsequent test phase.

In the study phase, participants viewed 10 faces presented one at a time, for five seconds each (image sizes were set in Qualtrics to display consistently regardless of screen size). After viewing each face participants completed a short 30 seconds distractor task in which they were asked to generate as many items as possible for a variety of different semantic categories (as in Nguyen et al., 2018). This delay was designed to increase the accuracy of judgements by giving participants insight into their own long-term memory of the stimulus before making their judgement. Existing work has shown that typically delayed judgements of learning are significantly more accurate than immediate judgements of learning (known as the delayed judgement of learning effect, see Rhodes & Tauber, 2011).

After viewing each face and completing the distractor task, participants were then asked 'Think about the person whose face you just saw. If you saw a different photo of that person in 5 minutes time, do you think you would recognise them?' (self condition) or 'Other people in this study are viewing the same faces as you. Think about the person whose face you just saw. Imagine another person in the study saw that same face. If they then saw a different photo of that person in 5 minutes time, do you think they would recognise them?' (other condition). Participants were asked to respond either 'yes' or 'no'.

Participants then completed the test phase where they were shown a series of 20 faces one at a time. Half of those were faces of people they saw during the study phase (the same person but with a different facial expression and different clothing, as described above) and the other half were new faces. Participants were then asked to make a binary yes or no judgement of whether or not they saw each face before.

Following completion of the task, participants completed two scales measuring their objective and subjective memory ability for faces—the Cambridge Face Memory Test (CFMT, Duchaine & Nakayama, 2006) and the Eyewitness Metamemory Scale (EMS, Saraiva et al., 2019). These scales were completed at the end of the study to avoid possible contamination of responses in our metamemory task. The CFMT measures face recognition abilities through a forced choice paradigm in which participants study unfamiliar target faces and are then asked to identify target faces in 72 trials under conditions of increasing difficulty. The EMS scale measures self-report face recognition capacity. It contains 23 statements (e.g. 'My ability to remember faces is much better than other people's ability to remember faces' and 'If I witnessed a robbery, I would be able to recognise the perpetrator a month later') that participants must rate on a seven-point Likert scale from strongly disagree to strongly agree.

Results

Analysis 1: Comparing judgements of learning with social metamemory judgements

As per our pre-registered analysis plan, we tested Hypothesis 1 by comparing the judgements of participants in the self condition (i.e. judgements of learning, henceforth referred to as self judgements) with the judgements of participants in the other condition (i.e. social metamemory judgements, henceforth referred to as other judgements). In each condition, we scored each of the 10 judgements made in the study phase (relating to whether a face would be remembered) as accurate or inaccurate.

In the self condition, we scored each judgement made by the participant from the study phase against memory from their own responses in the test phase. Self judgements were classified as accurate if the participant predicted they would not remember the face at the study phase and did not remember it at the test phase or predicted they would remember the face at the study phase and did remember it at the test phase. Conversely, self judgements were scored as inaccurate if the participant predicted that they would not remember the face at the study phase and did remember it at the test phase or predicted they would remember the face at the study phase and did not remember it at the test phase.

In the other condition, each participant was yoked to a random person in the self condition, and their judgement in the study phase was scored against the test phase responses of the person that they were yoked to. Specifically, participants from the other and self conditions were randomly assigned an ID number from 1 to 160 within their condition, and then participants in each condition with corresponding IDs were yoked to one another. For example, if participant X from the self condition was randomly assigned ID '5' they were yoked to the participant in the other condition who was also randomly assigned ID '5'. Other judgements were classified as accurate if they predicted that another person would remember the face and the person they were yoked to did remember the face, or if they predicted another person would not remember the face and the person they were yoked to did not remember the face. On the other hand, the participant was classified as inaccurate in the judgement if they predicted that another person would remember the face and the person that they were yoked to did not remember the face, or if they predicted another person would not remember the face and the person they were yoked to did remember the face (scores created from this yoking are referred to as 'random' accuracy scores in Analysis 2 below).

Descriptive statistics

Descriptive statistics for self judgement and other judgement accuracy, sensitivity and response bias can be seen in Table 1. Self judgement accuracy and sensitivity were significantly above chance performance, whilst other judgement accuracy and sensitivity did not significantly differ from chance. Self judgement response bias was significantly below zero, but other judgement response bias did not significantly differ from zero.

TABLE 1 Descriptive statistics for other judgement and self judgement accuracy, results of one-sample *t*-tests comparing means against μ (i.e. the *t*-test one-sample test value for chance in each analysis)

	Mean (*SD*)	μ	*df*	*t*	*p*
Accuracy					
Other Judgement	0.50 (0.18)	.50	159	0.18	.858
Self Judgement	0.57 (0.16)	.50	159	5.34	<.001
Sensitivity					
Other Judgement	0.03 (0.73)	0	146	0.55	.583
Self Judgement	0.24 (0.70)	0	146	4.12	<.001
Response bias					
Other Judgement	0.04 (0.64)	0	146	0.75	.457
Self Judgement	−0.15 (0.71)	0	146	2.48	.014

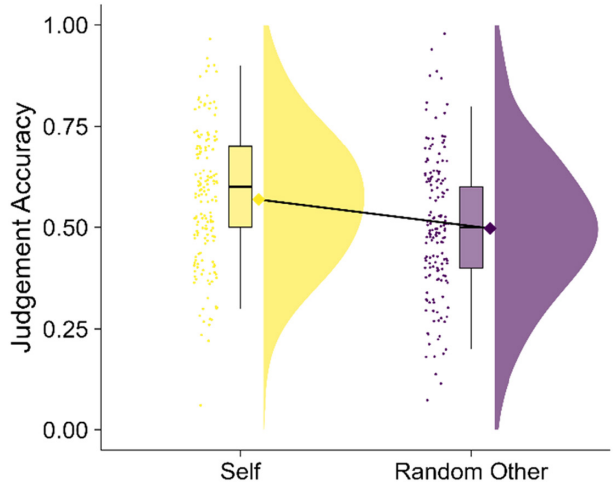

FIGURE 1 Average judgement accuracy between the self and other conditions. Raincloud plots depict (left-to-right) the jittered participants' averaged data points, box-and-whisker plots, means (represented by diamonds) and frequency distributions

Raw accuracy

A mixed-effects logistic regression model was run to test Hypothesis 1, using condition (self judgements vs. other judgements) to predict accuracy at the trial level with random effects included for participant and trial (allowing us to account for variability between stimuli and participants, Judd et al., 2012). We conducted this analysis using the *lme4* and *lmerTest* packages in *R* (Bates et al., 2015; Kuznetsova et al., 2017).

This analysis revealed the predicted effect of condition, such that selfjudgement accuracy was significantly higher than other judgement accuracy ($b = .29$, $z = 3.13$, $p = .002$, 95% CI [0.11, 0.47]; Figure 1). Means for each condition, shown in Table 1, show that self judgements were accurate 57% of the time and other judgements were accurate 50% of the time ($d = 0.42$).

Sensitivity and response bias

As an additional test of Hypothesis 1, we ran linear regression models to predict *d'* and *C* scores from condition (self judgements vs. other judgements) to assess whether sensitivity and response bias scores differed between the self and other conditions. We conducted these analyses using the *base stats* package in *R* (see Figure 2). Self judgement sensitivity was significantly higher than other judgement sensitivity

(b = .21, $t_{[1,292]}$ = 2.46, p = .015; Figure 2). Means for each condition, shown in Table 1, show that self judgement sensitivity was .24 and other judgement sensitivity was .03 (d = 0.29). Self judgement response bias was significantly lower than other judgement response bias (b = −.19, $t_{[1,292]}$ = 2.34, p = .020; although note this lower number actually represents more bias). Means for each condition, shown in Table 1, show that self judgement response bias was −.15 and other judgement response bias was .04 (d = 0.27). Whilst participants making self judgements had a tendency to respond yes (indicated by a negative C), participants making other judgements had a tendency to respond no (indicated by a positive C; Figure 2), although note that deviation from 0 (no bias) was only significant in self judgements and not in other judgements (see above).

Analysis 2: Improving social metamemory judgements through matching participants

In order to examine whether the accuracy of social metamemory judgements improves when participants make judgements about others who share a meaningful characteristic with them, we generated a second set of accuracy scores for each participant in the other condition by matching them to someone with similar memory performance as them as opposed to a random other. As per our pre-registered analyses, we first inspected the relationship between CFMT (M = 68.14, SD = 12.14) and EMS (M = 83.68, SD = 12.69) scores. As performance on both tasks was not strongly or significantly associated (r = .10, p = .082) and based on our pre-registered analytical decision, we yoked participants from the self condition with participants from the other condition within one of nine groups based on tertile splits of CFMT and EMS scores separately (as described below).

Specifically, we created nine separate groups based on performance on the two scales, such that participants were categorized as above average, average or below average on each scale (based on tertile splits). They were then assigned to one of nine groups based on those categorizations (above average CFMT and above average EMS, above average CFMT and average EMS, above average CFMT and below average EMS, average CFMT and above average EMS, average CFMT and average EMS, average CFMT and below average EMS, below average CFMT and above average EMS, below average CFMT and average EMS and below average CFMT and below average EMS) and yoked to a participant from the self condition in their equivalent memory group.

The numbers of participants in each of our nine groups differed in the self condition and other condition. Our sample size for this analysis was therefore limited by the number of matched pairs we could create in each group, which was equal to the number of participants in whichever of the self and other condition had the lower number of participants in that group. As a result, the sample size

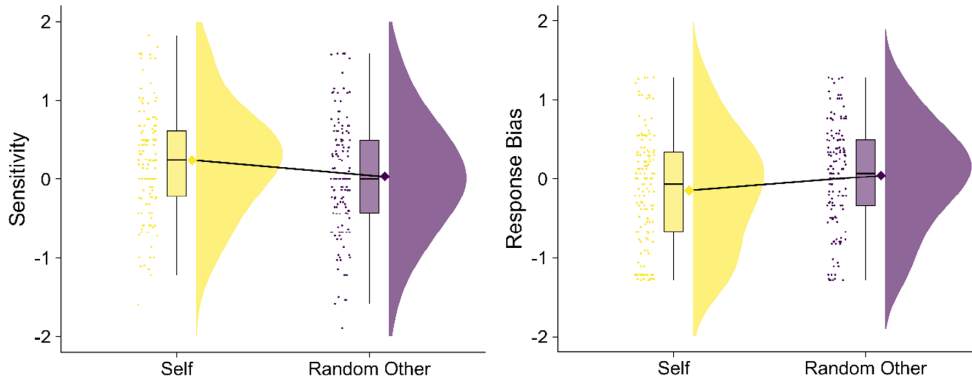

FIGURE 2 Sensitivity (left hand panel) and response bias (right hand panel) between the self and other conditions. Raincloud plots depict (left-to-right) the jittered participants' averaged data points, box-and-whisker plots, means (represented by diamonds) and frequency distributions

that we could generate a matched other judgement for was 136. Table 2 shows the sample sizes in each group in the self and other condition, and the overall number of matched pairs that could be created in each group.

In order to test whether this matching improved the accuracy of social metamemory, we compared these matched accuracy scores (i.e. social metamemory judgements about a matched-other, henceforth matched-other judgements) for each participant in the other condition with the original random accuracy scores for each participant (i.e. social metamemory judgements about a random other, henceforth random-other judgements), described above. Each person in the other condition therefore had two scores which were compared—a score when their predictions were assessed against a random other (their random-other judgement), and a score when their predictions were assessed against a matched other (their matched-other judgement).

Note that we also conducted exploratory analyses by yoking participants based on just CFMT scores (due to the CFMT being a performance measure of face recognition ability and therefore our most direct measure of face recognition ability, and to run an analysis including our full sample size), and the pattern of results was consistent with those reported below.

Descriptive statistics

Descriptive statistics for accuracy, sensitivity and response bias for matched-other and random-other judgements can be seen in Table 3. Judgement accuracy and sensitivity in both conditions did not significantly differ from chance, and response bias in both conditions did not differ significantly from 0.

Raw accuracy

To test Hypothesis 2, we ran a logistic mixed-effects regression model to predict raw accuracy at the trial level from condition (random-other or matched-other judgements), with random effects included for participants and trial. Random-other judgement accuracy did not significantly differ from matched-other judgement accuracy ($b = .03$, $t_{(1,18)} = 1.14$, $p = .269$, 95% CI [−0.07, 0.31]).

Sensitivity and response bias

As a further test of Hypothesis 2, we conducted a linear mixed-effects regression analysis to predict sensitivity and response bias from condition (random-other or matched-other judgements), with participant included as a random effect. Random-other judgement sensitivity did not significantly differ from matched-other judgement sensitivity ($b = .09$, $t_{[1256]} = 0.98$, $p = .327$, 95% CI [−0.09, 0.29]). Random-other judgement response bias also did not significantly differ from matched-other judgement response bias ($b = -.01$, $t_{[1,256]} = 0.08$, $p = .937$, 95% CI [−.16, .15]).

TABLE 2 Sample sizes for nine groups created during matching process

Cambridge face memory task tertile	Eyewitness metamemory scale tertile	*n* (self condition)	*n* (other condition)	*n* (matched pairs)
High	High	15	18	15
High	Moderate	24	17	17
High	Low	9	19	9
Moderate	High	13	14	13
Moderate	Moderate	16	21	16
Moderate	Low	20	15	15
Low	High	19	19	19
Low	Moderate	14	20	14
Low	Low	27	20	20

TABLE 3 Descriptive statistics for random-judgement and matched-judgement accuracy and results of one-sample *t*-tests comparing means against μ (i.e. the *t*-test one-sample test value for chance in each analysis)

	Mean (*SD*)	μ	*df*	*t*	*p*
Accuracy					
Random Judgement	0.51 (0.17)	.50	137	0.50	.620
Matched Judgement	0.48 (0.18)	.50	137	1.38	.169
Sensitivity					
Random Judgement	0.06 (0.73)	0	128	0.98	.332
Matched Judgement	-.03(0.78)	0	128	0.43	.667
Response bias					
Random Judgement	0.02 (0.63)	0	12	0.35	.730
Matched Judgement	0.03 (0.65)	0	12	0.45	.665

Discussion

The results of Experiment 1 supported our prediction that social metamemory judgements would be less accurate than judgements of learning in the context of identifications, both in terms of raw accuracy and sensitivity. Participants who made judgements of learning performed significantly better than chance when predicting what they thought they would remember—making correct predictions 57% of the time. Conversely, participants who made social metamemory judgements performed no better than chance when predicting what others would remember. These results are the first evidence that whilst eyewitnesses themselves may be able to predict what they will remember having seen a stimulus with some level of accuracy, others (e.g. jury members) may not be able to predict what a witness would remember had they seen a stimulus any better than someone randomly guessing would. However, it is also worth noting that the relatively low level of accuracy (57% accuracy, $d' = 0.24$) in even the self condition indicates that our task was difficult, even for participants making judgements relating to themselves.

Results did not provide support for our predictions in two respects. First, there was an unexpected response bias in self judgements that was not present in other judgements. Specifically, participants tended to overestimate their own memory ability but not the memory ability of others. This result, whilst unpredicted, is consistent with some previous work in other contexts (Kornell & Bjork, 2009, finding that participants underestimated forgetting on memory tasks involving word pairs; Sahar et al., 2020, finding that participants overestimated their performance and underestimated their errors in a memory task involving complex items held in visual working memory), and provides some evidence that people may be more likely to overestimate their own memory than to overestimate the memory of others.

Our prediction that accuracy in social metamemory judgements can be improved when people make judgements about others who share a meaningful characteristic with them—as opposed to judgements about random others—was not supported. This lack of support may have occurred because people are not driven by characteristics of their own memory when making judgements about other people's memory. Rather, inaccuracy might arise as a result of the lack of information relating to the other person leaving people to, essentially, guess. It is also possible that people do draw on beliefs informed by their own experience when making judgements about the memory of others, but that there is simply too much variability in the memorability of different faces for drawing on these beliefs to facilitate accurate judgement—even when a person is making judgements relating to someone with a similar memory capacity to them. Even people who are generally good at remembering faces may differ from others who are generally good at remembering faces in terms of which faces they find more or less memorable. The lack of an effect may also just have resulted from the difficulty of our task. Potentially an easier task could have created more variability in performance and, relatedly, could have revealed the predicted effects.

Future research should further examine inaccuracies in social metamemory judgements where the experience of the stimulus is the same for the assessor and the person being assessed and should examine potential methods to improve social metamemory judgements in that context. Such methods might include providing the assessor with information about the person they are assessing that has the potential to be probative in making predictions relating to their memory (e.g. information relating to their memory ability, see Andersen et al., 2014, or facial recognition ability, see Bindemann et al., 2012). Future research might also consider ways to make the task easier, for example by including faces with specific memorable features.

EXPERIMENT 2

Experiment 1 showed that inaccuracies in social metamemory judgements exist even when assessors experience a stimulus in the same way as the person whose memory they are making predictions about. Such predictions may be even more prone to inaccuracy in more realistic legal scenarios where the experience of the assessor differs from the experience of the person whose memory they are assessing. In these scenarios, judgements of the assessor may be biased by elements of the assessor's experience (e.g. how long they themselves saw the stimulus for) that are relevant to how well they may remember things, but which are clearly not relevant to the likely memory of another person who experienced the stimulus a different way (e.g. who viewed the stimulus for a different length of time). This possibility was examined in Experiment 2.

Method

Participants

Experiment 2 included 275 participants recruited online from the Prolific survey platform. This sample size was chosen based on an a priori power analysis for detecting a medium effect ($f = .25$) in a 2×2 within-subjects experimental design with 95% power using the *Webpower* package in *R* ($N = 250$; Zhang & Mai, 2018), plus 10% to account for attrition. To be eligible for the study, participants were required to meet the same inclusion requirements as in Experiment 1 and as in Experiment 1, participants were required to participate on a laptop or desktop computer. No participants were excluded based on our pre-registered exclusion criteria of failing two or more attention check questions (out of five total questions which were randomly distributed throughout the survey; 2.55% correctly answered four questions and 97.45% correctly answered all five questions).

Participants in the final sample ($N = 275$) had an average age of 40.23 years ($SD = 13.48$, *range* = 18–78), and the majority (70.18%) self-identified as female (28.00% as male and 1.82% as gender diverse). Each participant was compensated £5 for completing the approximately 30 minute experiment.

Design

All participants made social metamemory judgements relating to whether they thought a witness to a crime would remember the face of a specific person if they saw them at a crime scene. Each participant saw the faces of 20 people and made a judgement for each. The study utilized a 2 (size of faces) × 2 (exposure time) within-subjects design. Faces were either large (378 × 473 pixels) or small (76 × 95 pixels)[2] (image sizes were set in Qualtrics to display consistently regardless of screen size) and exposure time was either short (5 seconds) or long (25 seconds).

[2]Note that these sizes differ slightly from the sizes described in our registered report – 10 cm × 10 cm and 2 cm × 2 cm, due to the images used being rectangular.

Materials and procedure

As in Experiment 1, the experiment was hosted on Qualtrics, using face stimuli from a database of male faces used in previous work (Meissner et al., 2005; Nguyen et al., 2018). Each participant viewed twenty faces presented one at a time. These faces varied in size and exposure length per the conditions outlined above. After viewing each face participants completed a short 30 s distractor task to facilitate greater judgement accuracy (as in Experiment 1) and were then asked 'Imagine that a witness saw this person clearly (including their face) at a crime scene for approximately 15 seconds. If the witness saw this person's picture in a police lineup one day later, do you think the witness would recognise them?' Participants were asked to respond either 'yes' or 'no'. Fifteen seconds was chosen as the time period that the crime was said to have been viewed for since it differed equally from the time that participants saw the relevant faces in our short exposure time condition (5 s) and our long exposure time condition (15 s). We used four different stimulus sets so that each specific face used appeared in each cell of the design in order to remove the possibility of results being an artefact of specific stimuli. Participants were randomly assigned to a stimulus set, and within stimulus sets the order of faces seen was randomized.

Note that in this design we did not provide participants with an estimate of the distance that the witness was from the person that they saw (only noting that they saw the person clearly). In a short follow-up study, utilizing a cut-back version of the design focusing only on image size, we demonstrated that results replicated even where this specific estimate of distance was given. Details and results of this study are reported in the Supporting Information accompanying this manuscript, and materials and data are available on OSF (https://osf.io/ms3f4).

Results

Analyses examined whether the size of the face image the assessors saw and the length of time they viewed the image for influenced their judgements of whether another person would remember the person if they had seen them at a crime scene.

Descriptive statistics

Descriptive statistics for the proportion of faces participants predicted that a witness would remember, split by condition, are displayed in Table 4.

Regression analyses

First, we ran a linear regression model to examine the impact of our manipulations on predictions about what a witness would remember (averaged across trial) using the *base stats* package in *R*. We predicted social metamemory judgements from size (small or large) and exposure time (short or long) and the

TABLE 4 Descriptive statistics (means and standard deviations) for the proportion of faces participants predicted that a witness would remember in each condition

		Size	
		Small	**Large**
Exposure time	Short	0.37 (0.30)	0.65 (0.29)
	Long	0.40 (0.30)	0.69 (0.28)

interaction between size and exposure time. As shown in Figure 3, participants were significantly more likely to predict that a witness would remember a person from a crime scene when they themselves saw a large image of that person's face than when they saw a small image of that person's face ($b = -.29$, $t_{[3,1096]} = 11.51$, $p < .001$). Participants predicted that a witness would remember a person from a crime scene in 67% of cases when the participants saw a large picture of the person, compared to in 38.5% of cases when they saw a small picture of the person ($d = 1.10$). However, judgements did not significantly differ between short or long exposure time ($b = -.04$, $t_{[3,1096]} = 1.77$, $p = .077$), and the interaction between size and exposure time was also not significant ($b = .01$, $t_{[3,1096]} = 0.23$, $p = .822$).

Next, we ran a follow-up logistic mixed-effects regression model using the *lme4* and *lmerTest* packages in *R* to examine the impact of our manipulations on social metamemory judgements, controlling for participant and trial. Specifically, we predicted metamemory judgement at the trial level from size (small or large), exposure time (short or long) and the interaction between size and exposure time, with random effects included for participant and trial. As in our initial analysis, participants were significantly more likely to predict that a witness would remember a person from a crime scene when they (the participants) saw a large image of the person's face compared with a small image ($b = -1.54$, $z = 8.36$, $p < .001$, 95% CI [−1.91, −1.18]). However, again, judgements did not significantly differ between short and long exposure time ($b = -.24$, $z = 1.32$, $p = .188$, 95% CI [−0.61, 0.12]), and the interaction between size and exposure time was not significant ($b = -.46$, $z = .18$, $p = .860$, 95% CI [−0.46, 0.56]).

Discussion

The results of Experiment 2 (particularly when supported by our follow-up study reported in the Supporting Information accompanying this manuscript) support our prediction that changes in how an assessor (the person assessing the memory of another) experiences a stimulus (here a face) will impact their judgements of whether another person would remember that stimulus, despite that other

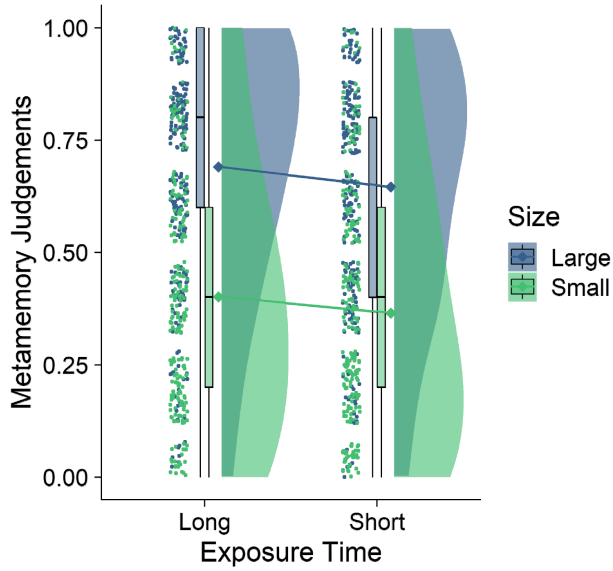

FIGURE 3 Social metamemory judgements of whether a witness would be likely to remember a face from a crime scene (proportion of stimuli of that type predicted to be remembered) split by the size of the image seen by the person making the metamemory judgement and the exposure time of the person making the metamemory judgement to the face. Raincloud plots depict (left-to-right) the jittered participants' averaged data points, box and whisker plots, means (represented by diamonds) and frequency distributions

person having experienced the stimulus differently. Specifically, our results show that the size of an image viewed by an assessor (something relevant to the likely memory of the stimulus for them; Price & Harrison, 2017) impacted whether they thought a witness who saw the person in the image at a crime scene would remember them. When participants saw a larger picture of the person, they were more likely to predict that a witness who saw that person at a crime scene would remember them, despite the size of picture that the assessor saw not being relevant to the size at which the witness whose memory they were assessing saw the face (made explicit in our follow-up study).

Contrary to our predictions, exposure time had no significant impact on social metamemory judgements. This lack of impact could have occurred for a number of reasons. It is possible that the specific exposure times that we used were insufficiently distinguishable to detect a difference, or that participants did not actually look at the stimulus for the entire time it was on the screen (thus minimizing the effect of the manipulation). It is also possible that the time that assessors are exposed to a face does not influence their judgements of whether a witness would remember the person whose face it is and, relatedly, that exposure time is not as salient as image size in making judgements about a face. Future research should explore these possibilities by examining how different exposure times (e.g. shorter and longer than 5 and 25 s) impact social metamemory judgements in experimental designs, taking additional steps to ensure participants pay attention during the entirety of each trial.

The finding that an assessor's experience of a face can bias their social metamemory judgements in at least one respect (the size of the image viewed by an assessor) is important. In litigation, finders of fact (e.g. jurors), experience stimuli in a very different way from how witnesses have experienced them. Finders of fact are often able to view the person alleged to have been involved in a crime clearly and close up, whilst witnesses often only see them at a distance. Although having the chance to engage with trial participants and evidence is important for finders of fact, this engagement may compromise memory assessments as it means that finders of fact experience relevant stimuli differently from how witnesses have (or have allegedly) experienced them.

GENERAL DISCUSSION

The experiments reported in this paper are the first to examine social metamemory in the legal context. Results support predictions in important ways, highlighting potential inaccuracy in social metamemory judgements. Firstly, results suggest that, as predicted, social metamemory judgements are less accurate than judgements of learning. In fact, whilst people performed better than chance in making judgements about their own memory, they performed no better than chance in making predictions about the memory of others. These findings depart slightly from previous work which found social metamemory judgements relating to word pairs to be more accurate than chance (Tullis & Fraundorf, 2017). This difference may be due to the enhanced complexity of faces when compared to words. Secondly, results show that, as predicted, changes in how an assessor experiences a picture of a person (specifically the size that they view the person's face at) can impact their social metamemory judgements of whether a witness would remember that person if they saw them at a crime scene. Thus, results show that social metamemory judgements are prone to inaccuracy and can also be biased by the experience of the assessor. Fundamentally, the results of these studies show that people are not necessarily good at making judgements about the memory of others. These findings have important applied implications.

Applied implications

Errors in social metamemory have the potential to influence evaluations of witnesses themselves (e.g. their honesty) and the probative value that decision makers draw from their testimony. Legal systems often presume that lay decision makers are well-placed to assess the accuracy of the memory of others

by virtue of having their own experiences with memory, and that memory can be effectively assessed based on common sense (see Akhtar et al., 2018). Implicit in these presumptions are that people's own experience of memory helps them effectively understand and assess the memory of others. In England and Wales, these sentiments have led to heavy reliance on lay decision makers in evaluations of testimony, and reluctance to introduce expert evidence on witness credibility (*R v Pendleton*, 2002; see also *State v Coley*, 2000, in the US context). However, this paper demonstrates that just because people can make somewhat effective judgements about their own memory, that does not mean they can make effective judgements about the memory of others (as demonstrated in Experiment 1). In fact, reliance on own memory in judging the memory of others can even be misleading (as demonstrated in Experiment 2). The work reported in this paper adds to a body of research highlighting difficulties that lay decision makers may have in assessing memory (Akhtar et al., 2018; Helm, 2021).

Limitations, future directions and conclusions

This work is intended as an initial exploration of some errors in social metamemory that have the potential to influence legal judgements. As such, it leaves important questions to be answered in future work. First, it will be important to examine empirically how social metamemory judgements (and errors in them) feed through into judgements in the context of legal disputes. For example, work should examine if social metamemory judgements do influence the weight given to a witness's testimony when that testimony is examined by jurors or mock jurors. It will also be important to examine social metamemory in contexts that more directly replicate how stimuli are experienced in real life. Although faces are important stimuli in legal cases, people have typically experienced them in the context of interactions or viewing events, rather than on computer screens. Assessors may be better at predicting what witnesses would remember from events the assessors have more experience with, such as personal interactions, than they are at predicting what others would remember when shown stimuli in an experimental study (although note that in some ways our design should have facilitated more accurate social metamemory, particularly in Experiment 1, where the assessor was experiencing the same thing as the person they were making judgements about). Because the way stimuli were experienced and assessed in this study did differ from how they are experienced and assessed in real life, we do not expect that effect sizes necessarily represent the size of effects in real-world scenarios. However, the work captures important cognitive tendencies that are likely to have an impact in real-world scenarios and that are important to explore and consider in those scenarios as well as in experimental work.

Ultimately, the fact that as humans we can be (and potentially often are) inaccurate when making predictions relating to the memory of others raises uncomfortable questions for legal systems, which rely on finders of fact being able to make accurate assessments of memory. If errors in social metamemory judgements feed through into errors in overall assessments of witness memory, the ability of finders of fact to evaluate witness testimony is undermined. And if finders of facts relying on their own experiences of memory does not facilitate accurate judgements relating to the memory of others, it is unclear what finders of fact should be basing their assessments of the memory of witnesses on, since they do not typically possess expertise on memory independently of their own experience. In this context, it is important to consider moving away from presumptions that finders of fact can evaluate memory appropriately, and towards designing and implementing infrastructure that can optimize their performance in making these evaluations. Future research on social metamemory should be used to inform tools or other interventions to improve evaluations of the memory of others (Pawlenko et al., 2012).

AUTHOR CONTRIBUTIONS

Rebecca Helm: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; validation; visualization; writing – original draft; writing – review and editing. **Bethany Growns:** Formal analysis; investigation; methodology; project administration; visualization; writing – review and editing.

FUNDING INFORMATION

This research is funded by a UK Research and Innovation fellowship to Rebecca K. Helm [MR/T02027X/1].

CONFLICTS OF INTEREST

The author has no real or potentially perceived conflicts of interest.

OPEN RESEARCH BADGES

This article has earned Open Data and Preregistered Research Designs badges. Data and the preregistered design and analysis plan are available at https://osf.io/ms3f4/?view_only=f6ddb94b127045b1b3b7a487bbe8d874, https://osf.io/vk9we/?view_only=8803087981134420bf72f707394f2f4b, https://osf.io/cr8bj/?view_only=bb6efd07b02747cda148048e2a632f8d.

DATA AVAILABILITY STATEMENT

Preregistration, data and materials underlying the proposed work are available on OSF at: https://osf.io/ms3f4.

ORCID

Rebecca K. Helm https://orcid.org/0000-0003-1429-3847
Bethany Growns https://orcid.org/0000-0002-6665-8134

REFERENCES

Akhtar, S., Justice, L. V., Knott, L., Kibowski, F., & Conway, M. A. (2018). The 'common sense' memory belief system and its implications. *The International Journal of Evidence and Proof*, *22*(3), 289–304. https://doi.org/10.1177/1365712718784045

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences*, *60*, 36–40. https://doi.org/10.1016/j.paid.2013.12.011

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bell, B. E., & Loftus, E. F. (1989). Trivial persuasion in the courtroom: The power of (a few) minor details. *Journal of Personality and Social Psychology*, *56*(5), 669–679. https://doi.org/10.1037/0022-3514.56.5.669

Berman, G. L., Narby, D. J., & Cutler, B. L. (1995). Effects of inconsistent eyewitness statements on mock-jurors' evaluations of the eyewitness, perceptions of defendant culpability and verdicts. *Law and Human Behavior*, *19*(1), 79–88. https://doi.org/10.1007/bf01499074

Besken, M. (2016). Picture-perfect is not perfect for metamemory: Testing the perceptual fluency hypothesis with degraded images. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1417–1433. https://doi.org/10.1037/xlm0000246

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*(2), 96–103. https://doi.org/10.1016/j.jarmac.2012.02.001

Birch, S. A. (2005). When knowledge is a curse: Children's and adults' reasoning about mental states. *Current Directions in Psychological Science*, *14*(1), 25–29. https://doi.org/10.1111/j.0963-7214.2005.00328.x

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior*, *26*(3), 353–364. https://doi.org/10.1023/a:1015380522722

Brigham, J. C., & Bothwell, R. K. (1983). The ability of prospective jurors to estimate the accuracy of eyewitness identifications. *Law and Human Behavior*, *7*(1), 19–30. https://doi.org/10.1007/BF01045284

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, *36*(2), 429–437. https://doi.org/10.3758/s13421-012-0249-6

Duchaine, B., & Nakayama, K. (2006). The Cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologica*, *44*(4), 576–585.

Dutton, A., & Carroll, M. (2001). Eyewitness testimony: Effects of source of arousal on memory, source-monitoring, and metamemory judgments. *Australian Journal of Psychology*, *53*(2), 83–91. https://doi.org/10.1080/00049530108255128

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 1–9. https://doi.org/10.3389/fnhum.2014.00443

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*(5), 680–693. https://doi.org/10.1037/xlm0000332

Gregory, A. (2022, January 8). 'Ghislaine Maxwell trial juror hires lawyer as second member of the jury reveals sexual abuse' *The Independent*. https://www.independent.co.uk/news/world/americas/ghislaine-maxwell-trial-jury-abuse-b1988402.html

Helm, R. K. (2021). Evaluating witness testimony: Juror knowledge, false memory, and the utility of evidence-based directions. *The International Journal of Evidence and Proof*, *25*(4), 264–285. https://doi.org/10.1177/13657127211031018

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kaminski, K. S., & Sporer, S. L. (2018). Observer judgments of identification accuracy are affected by non-valid cues: A Brunswikian lens model analysis. *European Journal of Social Psychology*, *48*(1), 47–61. https://doi.org/10.1002/ejsp.2293

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and Cognition*, *19*(1), 251–264. https://doi.org/10.1016/j.concog.2009.12.010

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting One's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*, 643–656. https://doi.org/10.1037/0096-3445.133.4.643

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*(4), 478–492. https://doi.org/10.1016/j.jml.2005.01.001

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*(4), 449–468. https://doi.org/10.1037/a0017350

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 756–766. https://doi.org/10.1037/0278-7393.10.4.756

Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other-race faces: A dual-process approach. *Applied Cognitive Psychology*, *19*, 545–567. https://doi.org/10.1002/acp.1097

Nguyen, T. B., Abed, E., & Pezdek, K. (2018). Postdictive confidence (but not predictive confidence) predicts eyewitness memory accuracy. *Cognitive Research: Principles and Implications*, *3*(1), 1–13. https://doi.org/10.1186/s41235-018-0125-4

Paulus, M., Tsalas, N., Proust, J., & Sodian, B. (2014). Metacognitive monitoring of oneself and others: Developmental changes during childhood and adolescence. *Journal of Experimental Child Psychology*, *122*, 153–165. https://doi.org/10.1016/j.jecp.2013.12.011

Pawlenko, N. B., Safer, M. A., Wise, R. A., & Holfeld, B. (2012). A teaching aid for improving jurors' assessments of eyewitness accuracy. *Applied Cognitive Psychology*, *27*(2), 190–197. https://doi.org/10.1002/acp.2895

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Price, J., & Harrison, A. (2017). Examining what prestudy and immediate judgments of learning reveal about the bases of metamemory judgments. *Journal of Memory and Language*, *94*, 177–194. https://doi.org/10.1016/j.jml.2016.12.003

R v JH and TG. (2005). EWCA Crim 1828.

R v Pendleton. (2002). 1 WLR 72.

Rhodes, M. G. (2006). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 65–80). Oxford University Press.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, *137*(1), 131–148. https://doi.org/10.1037/a0021705

Sahar, T., Sidi, Y., & Makovski, T. (2020). A metacognitive perspective of visual working memory with rich complex objects. *Frontiers in Psychology*, *11*, 1–14. https://doi.org/10.3389/fpsyg.2020.001

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

Semmler, C., & Brewer, N. (2002). Effects of mood and emotion on juror processing and judgments. *Behavioral Sciences & the Law*, *20*(4), 423–436. https://doi.org/10.1002/bsl.502

Sommer, W., Heinz, A., Leuthold, H., Matt, J., & Schweinberger, S. R. (1995). Metamemory, distinctiveness, and event-related potentials in recognition memory for faces. *Memory & Cognition*, *23*(1), 1–11. https://doi.org/10.3758/BF03210552

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/bf03207704

State v Coley. (2000). 32 S.W. 3d 831, Tenn.

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language*, *95*, 124–137. https://doi.org/10.1016/j.jml.2017.03.003

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264–1269. https://doi.org/10.1037/a0023719

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition*, *43*, 647–658. https://doi.org/10.3758/s13421-014-0479-x

Wells, G. L., & Leippe, M. R. (1981). How do triers of fact infer the accuracy of eyewitness identifications? Using memory for peripheral detail can be misleading. *Journal of Applied Psychology*, *66*(6), 682–687. https://doi.org/10.1037/0021-9010.66.6.682

Zhang, M., & Mai, Y. (2018). WebPower: *Basic and Advanced Statistical Power Analysis*. R Package version 0.6. https://CRAN.R-project.org/package=WebPower

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Helm, R. K. & Growns, B. (2022). Predicting and projecting memory: Error and bias in metacognitive judgements underlying testimony evaluation. *Legal and Criminological Psychology*, *00*, 1–19. https://doi.org/10.1111/lcrp.12232