

The Role of Self-rated Ability in the Accuracy of Confidence Judgements in Eyewitness Memory and General Knowledge

TIMOTHY J. PERFECT*

University of Plymouth, UK

SUMMARY

It is argued that confidence stems in part from self-rated ability in a domain of knowledge and that in eyewitness memory such perceptions are erroneous. Two experiments tested these hypotheses. In both experiments participants rated their relative ability in the domains of eyewitness memory and general knowledge and subsequently took tests of each, giving confidence ratings for each item attempted. In both studies, self-rated ability predicted performance for general knowledge, but not eyewitness memory. Across participants confidence ratings were significant predictors of accuracy for general knowledge, but not for eyewitness memory. In Experiment 1 self-rated ability was predictive of confidence ratings for both domains, although this effect was weaker in Experiment 2. The argument that the accuracy of confidence judgements in eyewitness memory is undermined by a lack of insight into relative expertise is therefore supported. Copyright © 2004 John Wiley & Sons, Ltd.

The confidence with which a witness is able to identify a perpetrator in a lineup, or recall a particular detail from a crime is intuitively believed to be a good predictor of likely accuracy. However, the results of numerous psychological investigations indicate that such intuitions maybe misguided. Over the past 15 years or so, there have been several meta-analyses of work on the confidence-accuracy (C-A) relation in eyewitness memory indicating that the relation is either entirely absent (Wells & Murray, 1984), weak (Bothwell, Deffenbacher, & Brigham, 1987) or at best moderate (Read, Lindsay, & Nicholls, 1998). Each of these meta-analyses has advanced explanations as to why the C-A relation is weaker than is intuitively expected, and it is not the intention to review that work here. Instead I take a different point of departure.

In the eyewitness confidence literature there have been relatively few attempts to compare the ability of individuals to judge their likely accuracy in eyewitness memory with another domain of knowledge. One exception is the study by Perfect, Watson, and Wagstaff (1993) who compared performance on tests of general knowledge and eyewitness memory. They were careful to ensure that both mean performance, and variability in performance were matched across domains. There were two important findings with regard to the C-A relation. First, when calculated across items using a Goodman-Kruskal gamma correlation (Nelson, 1984), the C-A relation was highly reliable for both general knowledge and eyewitness memory, and in fact did not differ across domains. Second, and

*Correspondence to: Professor Tim Perfect, Department of Psychology, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK. E-mail: tperfect@plymouth.ac.uk

more pertinent here, when calculated across individuals, the C-A relation was reliable for general knowledge, but not for eyewitness memory.

This pattern has since been replicated on several occasions, utilizing both recall and recognition paradigms (Hollins & Perfect, 1997; Perfect & Hollins, 1996, 1999; Robinson & Johnson, 1996 (immediate test conditions)). Given that performance across domains has been matched, variability controlled for, and personality factors made redundant by the use of the same individuals for both domains, it is clear that an alternative explanation for the weak C-A relation in eyewitness memory is required.

The theoretical argument advanced in Perfect and Hollins' research stems from an observation by Wells, Lindsay, and Ferguson (1979). They pointed out that people are unable to calibrate their knowledge about face identification ability because they do not receive adequate feedback about the veracity of their memories. Indeed, social etiquette may lead to misleading feedback. If you say 'hello' to someone you believe you recognize in the street, convention dictates that they will respond likewise, even if they do not recognize you. Thus you may receive positive reinforcement for a false positive identification. Similarly, you may fail to say hello to someone you ought to, and social convention will again generally ensure that you do not receive negative reinforcement for misses.

Perfect and Hollins (1996, 1999; Hollins & Perfect, 1997) have advanced a two-stage argument that develops Wells et al.'s (1979) observation. They point out that metacognitive judgements such as confidence ratings use various heuristics, only some of which will be useful in eyewitness memory. Heuristics such as speed of retrieval, vividness of mental imagery and so forth enable individuals to distinguish between items. Such strategies apply equally to general knowledge and eyewitness memory, and so C-A relations based on such heuristics are robust for both domains. Hence in those studies where a within-subjects C-A relation is calculated, there is no difference across domains (e.g. Perfect et al., 1993). However, such heuristics only allow for discrimination across items, and cannot apply across individuals. For there to be a correlation across individuals, it is necessary that different individuals use the same ratings of confidence for the same level of accuracy. This does not occur for eyewitness memory because people lack knowledge about how good they are and, by implication, how confident they ought to be.

This theoretical account can explain the pattern of C-A relations across items and across individuals for general knowledge and eyewitness memory, but as yet there is no direct evidence that people do lack insight into their relative ability at eyewitness memory. Two previous studies have looked at this issue indirectly. Hollins and Perfect (1997) examined the effect of interleaving questions about an eyewitnessed event with general knowledge questions. When compared with a standard blocked design (each domain tested separately), mixing the questions in this manner improved the C-A relation for eyewitness memory. Hollins and Perfect (1997) argued that this occurred because mixing the questions lead participants to anchor their eyewitness confidence to their general knowledge confidence, thus leading the entire group to use the confidence scale in a more similar fashion. More recently, Perfect, Hollins, and Hunt (2000) showed that giving feedback to witnesses about their performance increased the C-A relation for eyewitness memory for the entire group on a subsequent test trial. However, the feedback had to be of a particular kind to be effective: participants had to be told how they had performed compared to others. Simply being told that they were accurate or not had no impact on the C-A relation on subsequent trials.

However, what is lacking is a direct demonstration that individuals do know their relative standing in general knowledge but do not know how good they are at eyewitness memory, as this account requires. The experiments reported here were designed to provide such evidence. In Experiment 1 participants were asked to rate their relative ability (compared to their peers) at face identification, and sports knowledge as exemplars of eyewitness memory and general knowledge respectively. They were then given recognition tests at both, involving confidence judgements in the normal manner. Because of the prior work by Hollins and Perfect (1997) demonstrating that eyewitness memory judgements are improved by mixing them with general knowledge questions, a fixed order was used in which eyewitness memory performance was tested first. This provides an uncontaminated measure of the eyewitness memory judgement, but of course does mean that there are potential carry over effects for the general knowledge test. However, given the previous lack of an impact on general knowledge, this was judged to be of less concern.

The expectation is that absolute judgements of confidence stem in part from a person's belief in how good they are in that domain as has been shown previously by Trafimow and Sniezek (1994) and Zackay (1998). In line with this previous research it is predicted that self-rated ability in a domain will correlate with post-test confidence for both eyewitness memory and general knowledge. However, it is also predicted that these beliefs will lack validity for eyewitness memory, but be predictive for a domain of general knowledge. Finally, since post-test confidence stems in part from beliefs that are invalid in eyewitness memory, we expect to replicate the standard finding of weaker C-A relations for eyewitness memory than for general knowledge.

EXPERIMENT 1

Method

Participants

Sixty-three undergraduate students from the Department of Psychology at the University of Bristol participated in the study. There were 19 male and 44 female students, and their mean age was 21.4 ($SD = 4.45$) years.

Procedure

Participants were tested as a single large group during an experimental demonstration class. They were told that participation was optional and that responses would be collected anonymously. Each participant was given a response sheet which requested that they rate how good they think they are at recognizing faces, and answering questions about sport, compared to their peers. For each knowledge domain, a 10-point scale was used, with percentage ranges indicating where in the population the participant thought they would fall. This ranged from 1 = 0–9% of people would be worse than me (*i.e.* I am amongst the worst), to 10 = 90–99% of people would be worse than me (*i.e.* I am amongst the best).

Participants were then shown three black and white photographs of female faces on the overhead projector at the front of the lecture room. Each photograph was shown individually for 5 s, with a 5 s gap between faces, and participants were instructed to try to remember each person. The photographs (and those used as distractors in the lineups) were unfamiliar to the participants prior to the test.

After the final face had been shown, participants saw a simultaneous photographic lineup of six faces, which included one of the targets. This was shown in an array with three rows of two faces, each labelled with a letter from A to F. Distractors were chosen to be roughly matched on age, hair length and hair colour. A different photograph of the target was used, to ensure that face recognition rather than picture recognition was achieved. Participants were told that each lineup contained one of the faces they had seen previously, and they were instructed to choose a person from each lineup even if they had to guess. The targets appeared in positions B (top right), E (bottom left) and C (middle left) for the three trials. They then rated their confidence in their choice on a 6-point scale which ranged from 1 (*It could be any of the 6: i.e. I am guessing*) to 6 (*It could only be the person I have chosen: i.e. I am certain*). This procedure was repeated for three lineups.

Participants then attempted three questions about British sport. These items maintained the format of the previous lineups. That is, there were six alternative answers, one of which was the target, participants were instructed to guess if necessary and participants rated their confidence in their choice using the same 6-point scale.

Results

As a group, participants believed that they would be better at recognizing faces than answering questions about sport. The mean rating for sport questions was 4.09 ($SD = 2.77$), whilst for face recognition it was 6.13 ($SD = 1.79$). A repeated-measures *t*-test revealed that this difference was significant, $t(62) = 5.37, p < 0.001$. Performance on the recognition tests was not significantly different, $t(62) = 1.09$, although it was in the opposite direction to the ratings. Participants answered 1.05 ($SD = 1.00$) sports questions correctly, but only recognized 0.89 ($SD = 0.76$) faces (each test out of 3). The fact that, overall, self-ratings and performance do not relate should not be over-interpreted; at the time of rating, participants did not know the difficulty of the tests they would have to take. However, this cannot account for the fact that participants post-choice confidence ratings were more positive for eyewitness memory than for sports items, $t(62) = 5.38, p < 0.001$. The mean confidence rating for face identification was 3.75 ($SD = 0.90$) whilst for sports questions it was 2.57 ($SD = 1.57$).

More pertinent are the interrelations among self-rated ability, actual performance and post-choice confidence. These data are shown in Table 1, with correlations above the diagonal for the lineup tasks, and below the diagonal for the sports questions.

Table 1. Experiment 1: Pearson correlations between self-rated ability (SRA), accuracy (a1–a3) and mean confidence (c1–c3) for face recognition (above the diagonal) and sports knowledge (below the diagonal)

	SRA	a1	a2	a3	c1	c2	c3
SRA		0.11	−0.06	0.09	0.26	0.15	0.31
a1	0.42		−0.00	0.04	0.07	0.27	0.27
a2	0.23	0.20		0.12	−0.05	0.11	0.06
a3	0.37	0.39	0.26		0.04	0.06	−0.10
c1	0.72	0.68	0.37	0.44		0.37	0.40
c2	0.63	0.48	0.46	0.46	0.66		0.49
c3	0.47	0.39	0.29	0.59	0.60	0.53	

Note: Correlations above diagonal are values for face identification test, and below the diagonal for sports questions. Shaded cells represent significant correlations (one tailed, with Bonferroni correction). SRA = self-rated ability in the domain of knowledge; a1 = accuracy on item 1; a2 = accuracy on item 2; a3 = accuracy on item 3; c1 = confidence for item 1; c2 = confidence for item 2; c3 = confidence for item 3.

The pattern of correlations is quite different for sports questions and face recognition. For sports questions, self-rated ability is positively associated with subsequent performance on all items, although with Bonferroni adjustment, only the correlation for item 1 remains significant. Overall collapsed across items, for sports questions self-rated ability is predictive of accuracy, $r = 0.48$, $p < 0.001$. For eyewitness memory, the self-rated ability correlates very weakly with performance on individual items. Collapsed across items the relation between self-rated ability and accuracy remains weak, $r = 0.06$. Performance on the individual sports questions relates strongly to confidence for those items, with all correlations being significant. However, for face recognition, this is not the case. This is also true if one collapses across items $r = 0.76$, $p < 0.001$ for sport, and $r = 0.18$, for lineups respectively. Where there is some similarity across domains is that self-rated ability is predictive of post-choice confidence. Collapsed across items, the correlation for sport is $r = 0.70$, $p < 0.001$, whilst for lineups it is $r = 0.31$, $p < 0.05$. For eyewitness memory this relationship is relatively weak because it is independent of performance, whereas for sports knowledge there is congruency between self-rated ability, actual ability and confidence.

Discussion

The initial predictions were all fully supported. Participants' self-rated ability was predictive of their standing within the group at sports knowledge, even though they did not know the questions they would be asked. However, this was not true for face recognition. Self-rated ability at recognizing faces was unrelated to relative ability on the lineups.

The relations between performance and post-choice confidence were as expected, and replicated previous work (e.g. Perfect et al., 1993; Perfect & Hollins, 1996, 1999; Robinson & Johnson, 1996—immediate test conditions). Across individuals there was no relation between confidence and accuracy for the eyewitness materials, but in the same individuals there was an extremely robust relation between confidence and performance on the sports questions. The overall average C-A relation for eyewitness memory ($r = 0.18$) is quite close to that reported in meta-analyses of face identification studies in eyewitness memory (e.g. $r = 0.25$, for 35 studies involving recognition testing by Bothwell et al., 1987). Given that the performance measures in this study are restricted in range (being binary for each question, and only ranging from 0–3 for the overall measure), it is perhaps inappropriate to draw too strong conclusions regarding the absolute values of the correlations. Instead our interest is in the comparison between the two forms of memory test. Here, explanations based on personality factors clearly have no role, because the same individuals performed each test.

Overall difficulty has been advanced as a potential mediator of the C-A relations (Kebbell, Wagstaff, & Covey, 1996), and is not matched across items here. It is hard to see how this can explain the discrepancy between the domains, however. Overall there was no significant difference between the two domains in mean performance. At the level of items, whilst there were considerable differences between the lineups and sports questions, these differences did not predict which items would show a robust C-A relation. For the first lineup, 24% were correct, for lineup 2 it was 54% and for lineup 3 it was only 11%. However, as Table 1 indicates, there was no relation between level of performance and the correlation observed with confidence. The corresponding values for the sports items were 30%, 29% and 46%, but once again there was relation between mean performance and the

strength of the C-A relation. If item difficulty fully explained the C-A relation then one would expect the highest C-A relation for lineup 2, but this was not the case. In fact the correlation for lineup 2 was $r = 0.11$, whilst for the sports items the correlations ranged from $r = 0.46$ to $r = 0.68$ despite the questions being objectively more difficult than lineup 2.

The evidence presented here has been used to argue that eyewitnesses do not know how good, or bad, they are compared to others. Clearly self-rated ability does not predict subsequent performance in the present experiment. However, the danger in building a theoretical argument on the null hypothesis must be acknowledged. One alternative interpretation of the present data are that individuals usually can predict their relative ability at face identification, but that for some reason the present task is unrepresentative of that domain of ability. However, there are a number of counter-arguments to this view.

The first is that there was no a priori way of knowing whether the face recognition task was any less (or more) representative of face recognition ability than the sports questions were of general knowledge. Even if this were true, it would be irrelevant for the present purposes. The lineup task was selected on ecological grounds. Lineups are a common form of identification task. Whether or not they represent a valid test of face recognition ability is irrelevant with regards to the performance of real eyewitnesses; this is the task that eyewitnesses have to do, and clearly people are not aware how good they are at it. Moreover, one must ask whether the face-identification performance of a real-life witness is likely to be more typical of their face-identification ability than the present experimental paradigm? It is hard to see how. A real-life witness might only get a brief glimpse of the perpetrator, perhaps under distressing circumstances, and almost certainly in incidental learning conditions. Our 'witnesses' saw the faces in clear view, under no duress, and knew they would be tested. It seems intuitively more likely that the latter situation would produce a face-identification task that is more representative of general face-identification ability than the former, and yet it was this situation in which people were unable to predict their performance. Thus, it is argued that the present paradigm is likely to reflect the ability of real witnesses to predict their likely success. It appears that one of the reasons that C-A relations are weak in eyewitness memory is the fact that people literally do not know how good (or bad) they are.

The final reason for believing that the null effect for eyewitness memory may be genuine is that there is prior research showing the same pattern in a different domain. For example, Zackay (1998) recently studied the ability of bank managers to predict the outcome of banking trials given differing amounts of diagnostic information. As in the present study he measured pre-test confidence as well as accuracy and post-item confidence. His findings mirrored the present ones for eyewitness memory. Pre-test confidence was unrelated to accuracy, yet was predictive of post-item confidence. Thus, rather than questioning why people are apparently unable to predict their performance in tests of face recognition, perhaps we should ask how it is they are so good at predicting their general knowledge ability.

Nevertheless it remains the case that so far the poor predictive ability of eyewitnesses has only been demonstrated using a single task (lineups for faces) in a single sample. Performance in this experiment was generally low, and notwithstanding the discussion about item difficulty, there remains a concern about assessing ratings when performance is around chance. In fact, across the three lineups, only performance on lineup 2 was significantly greater than chance, $t(62) = 5.89$, $p < 0.001$. For lineups 1 and 3, $t(62) = 1.31$ and -0.95 respectively. For this reason it was decided that a replication was necessary, using a different paradigm and sample. For Experiment 2 we used memory for the details

of an event presented on video, and tested memory using recall rather than recognition. For comparative purposes we also tested a range of aspects of general knowledge in the same manner. If these data replicate the pattern seen in Experiment 1, this would show that the effect is more generalized than merely recognition of faces and sports facts. The predictions regarding self-rated ability, post-test confidence and accuracy remained the same as Experiment 1.

EXPERIMENT 2

Method

Participants

Forty volunteers from the University of Bristol were recruited to take part in this experiment. Their average age was 20.0 ($SD = 1.95$) years.

Procedure

Participants were tested in small groups of up to eight people at a time. Their first task was to rate their relative ability in general knowledge, and eyewitness memory on a scale ranging from 1 (*better than others*) to 10 (*worse than others*). They were then instructed to watch a 10 min video clip, and to try to remember as many details as possible. They were not informed about the nature of the test that would follow.

After a 5 min filled delay, participants then took a written test for their memory of the eyewitness event. There were 48 questions in total, with 12 each in the domains of people (e.g. 'What colour shirt was Sam, the radio operator, wearing?'), objects (e.g. 'What fell out of the killer's pocket?'), actions (e.g. 'Where did the killer take the knife from when they were in the hut?') and verbal (e.g. 'What instruction did Sarah shout at the last man to die?'). Participants were asked to recall as many answers as they could, and to rate their confidence in their answer using a 5-point scale ($1 = I$ am absolutely certain that the answer I have given is correct to $5 = I$ am not at all certain that the answer I have given is correct). For those items that participants were unable to recall, participants were asked to make a feeling of knowing judgement. These data will not be discussed here.

Once all participants had completed this test, they were given a 48-item test of general knowledge in the same format. There were equal numbers of questions from the domains of geography, history, science and sport.

Results and discussion

Because of an administrative error, the rating scales used in Experiment 2 were in reverse order compared to Experiment 1. That is, high confidence was indicated by lower numerical ratings on the scale. Thus, a positive association between confidence and performance should be indicated by negative correlation. However, for ease of comparison with Experiment 1, we have reversed the signs of the correlations here, so that a positive association is indicated by a positive correlation.

On average, the sample thought that they were better at eyewitness memory than in the domains of general knowledge specified, $t(39) = 4.44$, $p < 0.001$. The mean rating for ability in eyewitness memory was 5.41 ($SD = 1.51$), compared to 6.28 ($SD = 1.36$) for general knowledge, where lower ratings indicate greater confidence. This difference was reflected in performance, with more questions being answered correctly in eyewitness

memory (mean 22.9 out of 48, $SD = 5.24$) than in general knowledge (mean = 19.4, $SD = 7.58$), $t(39) = 3.07$, $p < 0.01$. Because participants could select which questions to attempt in the recall test, we also looked at performance in terms of the accuracy of attempted questions. Here performance was marginally superior for eyewitness memory, with 70.7% ($SD = 11.4\%$) of attempted eyewitness questions answered correctly, and 64.8% ($SD = 11.7\%$) of attempted general knowledge questions, $t(39) = 1.96$, $p < 0.06$. Participants were also more confident about the answers they provided for eyewitness memory than for general knowledge, $t(39) = 2.86$, $p < 0.01$. The mean confidence rating for eyewitness memory was 2.26 ($SD = 0.46$) whilst for general knowledge it was 2.46 ($SD = 0.65$), where higher confidence is indicated by lower ratings, as described above.

Having examined the mean performance levels, the next analyses examined the relationships between the self-ratings, confidence ratings and performance. Unexpectedly, there was a significant difference between the within-subject gamma correlations, $t(39) = 3.07$, $p < 0.01$. Within-subject C-A correlations, assessed by the Goodman-Kruskal gamma correlation (Nelson, 1984) were higher for general knowledge $G = 0.74$ ($SD = 0.22$), than for eyewitness memory, $G = 0.58$ ($SD = 0.24$).

For general knowledge, participants' self-ratings were predictive of their success on the final test, as measured by the total number of correct answers given on the test, $r = 0.51$, $p < 0.001$. However, for eyewitness memory, the self-rating bore no relation to test performance, $r = 0.15$. The difference between these two correlations was significant, $Z_{r_1-r_2} = 1.77$, $p < 0.05$. Thus, as in Experiment 1, participants know their relative standing in general knowledge, but do not in eyewitness memory.

We also examined the relationship between confidence ratings and performance. Since we gathered confidence ratings only for those items attempted, the analysis of the relation between confidence and accuracy looked at the correlation between the mean confidence in items attempted and proportion correct. That is, this analysis examines whether those who are more confident on average are more likely to be accurate on average. For general knowledge, this relationship was significant, $r = 0.41$, $p < 0.01$, whilst for eyewitness memory it was not, $r = 0.28$, $p < 0.09$. This value for the eyewitness memory correlation is close to that reported in Bothwell et al.'s (1987) meta-analysis. The correlations for general knowledge and eyewitness memory did not significantly differ from one another, $Z_{r_1-r_2} = 0.64$, $p < 0.26$.

As before, we also examined the relationship between the self-ratings made before the test, and the mean confidence in expressed answers. For both tests both associations were in the predicted direction, but neither correlation was significant, $r = 0.25$, $p < 0.11$ for general knowledge, and $r = 0.29$, $p < 0.07$ for eyewitness memory.

The evidence from Experiment 2 was therefore quite clear in replicating the lack of predictive ability for peoples' beliefs in their ability in eyewitness memory tests. Despite the use of a different test format (recall not recognition), a different confidence scale, and testing event knowledge rather than face recognition ability, participants' self-ratings still did not predict their relative performance on the eyewitness test. At the same time, changing the nature of the general knowledge test did not alter the robust association between self-ratings and test performance.

Clearly if it had merely been shown that there is no relation between confidence and accuracy for eyewitness memory this would have been a demonstration of the null hypothesis. However, the inclusion of the general knowledge items is a crucial control that allows us to rule out several uninteresting explanations for this lack of association for eyewitness memory. Firstly, since the same individuals participated for general knowledge

and eyewitness memory tests, there can be no explanation in terms of a general inability to self-rate since clearly, for general knowledge, they can. This makes the failure to show such a relationship in eyewitness memory all the more interesting.

The data from the second experiment also make clear that the findings of Experiment 1 apply to other domains of eyewitness memory, tested by means of a recall test. Thus, the lack of ability to predict relative performance in eyewitness memory is not restricted only to memory for faces, and nor is it restricted to a recognition task with multiple (similar) foils. In the recall task, there were no foils to be considered (beyond the answers generated by the participants) and answering a particular question was optional. Further, since performance was slightly higher for eyewitness memory than for general knowledge, explanations of the relative pattern of C-A relations based upon task difficulty (Kebbell et al., 1996) cannot be sustained here.

As in Experiment 1, it is possible that for some reason the questions selected for the eyewitness test were somehow unrepresentative of the domain of eyewitness memory, while at the same time, the general knowledge questions were representative. Whilst this possibility must be acknowledged, it does not seem entirely plausible. First, the pattern was identical to Experiment 1, even though an entirely different form of eyewitness knowledge was tested (i.e. event details as opposed to face recognition). At the same time, a different set of general knowledge questions were asked. This begs the question of what a representative set of questions for eyewitness memory might be, if not face recognition and event memory. Additionally, the same argument as was advanced for Experiment 1 applies here. Asking witnesses for details about events is standard interviewing practice. Thus, the fact that witnesses' self-ratings lack validity in the present study is likely to apply to any real world situation to the same degree.

As in previous research (e.g. Perfect & Hollins, 1996, 1999) there were clear relations between confidence and accuracy for general knowledge, whether measured within- or between-subjects. Also in line with the same previous research were the robust within-subject relations for eyewitness memory. However, unlike that previous research, we also found a higher C-A relation (assessed by Gamma) for general knowledge than for eyewitness memory. This was unexpected, and we have no immediate explanation for this, given the failure to find a difference on numerous other occasions (see Perfect, 2002 for a review).

The evidence for a role of self-rated ability in post-test confidence was less compelling in this study than in Experiment 1, since neither of the domains showed a significant correlation, although both showed trends in the predicted direction. Two potential explanations of why this relation is reduced in the second study occur to us. One is that the element of choice in responding in this study reduced the association between self-rated ability and post-test confidence. Unlike Experiment 1, participants had a choice whether or not to attempt recall to a question. Thus participants with low self-rated ability in a domain may have been biased not to attempt a question when they were in doubt. This in turn may have led them to attempt questions only when their post-test confidence was high. This is a plausible explanation, but it is apparently not supported by the data. There was no association between self-rated ability and the propensity to give an incorrect answer, which would be expected if there were a bias not to respond when in doubt, $r = 0.16$, $p < 0.32$, for general knowledge, and $r = 0.21$, $p < 0.19$ for eyewitness memory.

A second possibility is that the number of items in the tests, and the nature of the test format, may have moderated this association. In Experiment 1, only three questions in each domain were used. Participants were given targets and distractors to choose between,

and attempted all questions. In Experiment 2, because we were interested in a range of questions, we used 48 questions on each test. We also used recall as our criterion test. This may have impacted on the use of confidence ratings in the following manner. As the test questions progress, participants will have experienced varying degrees of success in generating answers to the questions. This in turn may vary their initial impression of their own ability, so that later questions were answered with more or less confidence. Consider, for example, a person who failed to answer the first 10 questions, but generated an answer to the 11th. As a result of their previous 10 failures, that person may now have become convinced that they have a poor memory, and so give a lower confidence rating in the generated answer than they would have otherwise. Exactly the opposite might occur in someone who has generated answers to all the questions, and so has become more confident by the 11th item. Thus, the correlation between post-test confidence and initial self-rating for later items may become weaker as a result of earlier test performance. Note that the use of a recall test may be crucial here, because failing to recall an answer offers clearer feedback of performance than does a recognition test on which people may fail when they believe they are accurate, or succeed when they think they have failed. Clearly, this proposal is highly speculative. Unfortunately, the present data do not lend themselves to test these speculations, since different individuals failed different items. Thus, examining the consequence of n consecutive failures on the subsequent confidence rating for trial $n + 1$ is inevitably confounded with item differences. In any case, it is hard to know, a priori, how many successes or failures, in what proportion, are required to alter self-belief. Further research is needed to explore the relative roles of the number of items, and the feedback from the use of recall tests in systematic ways. This is beyond the scope of the present work, which has achieved what it set out to do.

Overall, then, the two experiments here demonstrate that self-rated ability in a domain can have an impact on how people rate their confidence on a particular test in that domain. Further, these studies suggest that the use of this heuristic is not useful in eyewitness memory, whilst it is for general knowledge. This is not to say that confidence judgements come only from self-rated ability: psychological research has documented many influences on the absolute level and accuracy of post-decision confidence judgements (e.g. Luus & Wells, 1994; Robinson & Johnson, 1998; Shaw, 1996; Sporer, Penrod, Read, & Cutler, 1995; Wells & Bradfield, 1998; Zackay, 1998). However, these factors were not the study of the present work.

CONCLUSIONS

The two experiments reported here represent the first direct test of the hypothesis that people lack insight into their relative ability at eyewitness memory. Both experiments were clear in demonstrating that whilst individuals were able to predict their relative standing in general knowledge, their self-ratings did not predict their ability in eyewitness memory. This supports the central tenet of the hypothesis that insight into relative ability in a domain provides a heuristic by which individuals can anchor their confidence judgements. The use of this heuristic is therefore likely to be useful in general knowledge, but not eyewitness memory.

A second strong prediction was that confidence judgements would be more predictive for general knowledge than for eyewitness memory. This has been reported on a number of occasions (Perfect et al., 1993; Perfect & Hollins, 1996, 1999; Robinson & Johnson,

1996). In the present studies, this pattern was found once again. This pattern is consistent with the idea that the C-A relation in eyewitness memory is weak, and is consistent with the idea that insight into relative expertise is lacking in eyewitness memory.

Finally, a third prediction was that post-test confidence would be based in part on pre-test beliefs about relative ability in a domain of knowledge. Experiment 1 showed this pattern clearly, but Experiment 2 showed only weak evidence for this. It was suggested that the weaker effects in the second experiment might be due to the use of multiple items in each domain leading participants to alter their self-belief as the tests progressed.

Whilst these data are consistent with the model outlined in the introductory section, it must also be acknowledged that they are correlational only, and so cannot be regarded as providing definitive evidence. What cannot yet be known is the extent to which belief in one's ability correlates with post-decision confidence under different test conditions. Also, these data only provide a partial test of the model presented. Whilst they suggest that belief about ability in a domain does impact on confidence judgements, they provide no evidence about the genesis of those beliefs. Wells et al.'s (1979) original idea, that feedback is lacking in eyewitness memory and this causes poor calibration for the domain of eyewitness memory, remains untested by the present data set.

ACKNOWLEDGEMENTS

The author would like to thank Ellie Shand for collecting the data reported in Experiment 2, and Ian Dennis for useful statistical advice.

REFERENCES

- Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: optimality hypothesis revisited. *Journal of Applied Psychology*, *72*, 691–695.
- Hollins, T. S., & Perfect, T. J. (1997). The confidence-accuracy relation in eyewitness memory: the mixed question type effect. *Legal and Criminological Psychology*, *2*, 205–218.
- Kebbell, M. R., Wagstaff, G. F., & Covey, J. (1996). The influence of item difficulty on the relationship between eyewitness confidence and accuracy. *British Journal of Psychology*, *87*, 653–662.
- Luus, C. A. E., & Wells, G. L. (1994). The malleability of eyewitness confidence: co-witness and perseverance effects. *Journal of Applied Psychology*, *79*, 714–724.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–133.
- Perfect, T. J. (2002). When does eyewitness confidence predict performance? In T. J. Perfect, & B. Schwartz (Eds.), *Applied Metacognition* (pp. 95–120). Cambridge: Cambridge University Press.
- Perfect, T. J., & Hollins, T. S. (1996). Predictive feeling of knowing judgements and postdictive confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, *10*, 371–382.
- Perfect, T. J., & Hollins, T. S. (1999). For eyewitness memory, confidence judgements are predictive, but feeling of knowing judgements are not. *Journal of Experimental Psychology: Applied*, *5*, 250–264.
- Perfect, T. J., Hollins, T. S., & Hunt, A. L. R. (2000). Practice and feedback effects on the confidence-accuracy relation in eyewitness memory. *Memory*, *8*, 235–244.
- Perfect, T. J., Watson, E., & Wagstaff, G. F. (1993). The accuracy of confidence ratings associated with general knowledge and eyewitness memory. *Journal of Applied Psychology*, *78*, 144–147.
- Read, J. D., Lindsay, D. S., & Nicholls, T. (1998). The relationship between accuracy and confidence in eyewitness identification studies: is the conclusion changing? In C. P. Thompson, D. Bruce,

- J. D. Read, D. Herrman, D. Payne, & M. Toglia (Eds.), *Basic and applied aspects of remembering*. Hillsdale, CA: Lawrence Erlbaum Associates.
- Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory and the eyewitness confidence-accuracy correlation. *Journal of Applied Psychology, 81*, 587–594.
- Robinson, M. D., & Johnson, J. T. (1998). How not to enhance the confidence-accuracy relation: the detrimental effects of attention to the identification process. *Law and Human Behavior, 22*, 409–428.
- Shaw, J. S., III. (1996). Increases in eyewitness confidence resulting from postevent questioning. *Journal of Experimental Psychology: Applied, 2*, 126–146.
- 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.
- Trafimow, D., & Sniezek, J. A. (1994). Perceived expertise and its effect on confidence. *Organizational Behavior and Human Decision Processes, 57*, 290–302.
- Wells, G. L., & Bradfield, A. L. (1998). 'Good, you identified the suspect': feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360–376.
- Wells, G. L., Lindsay, R. C. L., & Ferguson, T. J. (1979). Accuracy, confidence and juror perceptions in eyewitness identification. *Journal of Applied Psychology, 64*, 440–448.
- Zackay, D. (1998). Determinants of confidence in accuracy of knowledge retrieval. *European Journal of Cognitive Psychology, 10*, 291–306.