

Practice and feedback effects on the confidence-accuracy relation in eyewitness memory

Timothy J. Perfect

University of Plymouth, UK

Tara S. Hollins and Adam L. R. Hunt

University of Bristol, UK

It has been claimed that the lack of a reliable confidence–accuracy relation in eyewitness memory stems from eyewitnesses' lack of knowledge concerning their relative expertise. Two studies tested this idea by contrasting the effects of practice alone with practice with feedback in three successive eyewitness tests. Experiment 1 tested recall for events, and Experiment 2 used recognition of faces as test materials. Both studies showed that practice alone did not increase the confidence–accuracy relation, but practice with feedback on relative performance produced robust increases in the confidence–accuracy relation. This suggests that lack of calibration is one factor that causes the reported lack of association between confidence and accuracy for eyewitness memory.

INTRODUCTION

In the survey of expert opinion conducted over 10 years ago it was widely accepted that the confidence with which a witness makes an identification has little relation to the accuracy of that identification (Kassin, Ellsworth, & Smith, 1989). Recently, however, there is evidence to challenge the generality of this conclusion (Read, Lindsay, & Nicholls, 1998). Work in our own laboratory has repeatedly shown that confidence judgements are predictive across items but not across witnesses (Hollins & Perfect, 1997; Perfect & Hollins, 1997; Perfect, Watson, & Wagstaff, 1993). Kebbell, Wagstaff, and Covey (1996) and Read et al. (1998) have shown that the level of correlation is driven by the range of difficulty of items across which the relation is calculated. Linday, Read, and Sharma (1998) showed that the confidence–accuracy relation is strongest when variability between encoding conditions for different witnesses is greatest. At first glance this appears to contradict

Deffenbacher's optimality hypothesis (Deffenbacher, 1980; Bothwell, Deffenbacher, & Brigham, 1987) which states that the confidence–accuracy relation is driven by the quality of encoding, with stronger correlations found between confidence and accuracy when encoding conditions are better (or more optimal) at the time of witnessing the event. These two positions appear contradictory but they are not. In Deffenbacher's work the correlations are calculated for a group of individuals witnessing under the same conditions. In such circumstances, better encoding conditions probably produce more variability between individuals, as floor effects are avoided. It has also been claimed (Robinson & Johnson, 1996; Robinson, Johnson, & Herndon, 1997) that the confidence–accuracy relation is superior for recall than for recognition, although this claim has not gone unchallenged (Hollins & Perfect, 1997). Other researchers have focused on the witnesses, and shown that the confidence–accuracy relation is reliable if one separates out those who make a

Requests for reprints should be sent to Professor Tim Perfect, Department of Psychology, University of Plymouth, Plymouth PL4 8AA, UK. Email: tperfect@plymouth.ac.uk

This research was in part funded by the Economic and Social Research Council, UK (Grant no. R000234838).

choice in a lineup from those who do not (Sporer, Penrod, Read, & Cutler, 1995).

The research presented here is aimed at testing one particular theoretical account of the lack of the confidence–accuracy relation in eyewitness memory, namely the calibration hypothesis, which originally stems from Wells, Lindsay, and Ferguson (1979) and was elaborated on by Perfect et al. (1993). We outline this theoretical view later. Because this is our focus, we do not seek to rule out the explanations mentioned earlier. In the studies reported here our critical comparisons are for items matched for difficulty and range of difficulty, seen under identical encoding conditions with correlations calculated for the whole sample. Therefore our data do not address item difficulty, encoding conditions, or differences between witnesses, and any effects observed cannot be explained by these factors.

Our previous research (Hollins & Perfect, 1997; Perfect & Hollins, 1997; Perfect et al., 1993) has demonstrated that although there is generally a weak or absent confidence–accuracy relation across individuals for eyewitness memory, there is a robust confidence–accuracy relation for tests of general knowledge for those same individuals, even when general knowledge is matched for difficulty (and range of difficulty) with eyewitness memory. At the same time we have found that within-subject confidence–accuracy correlations are equally accurate for general knowledge and eyewitness memory. How is this pattern to be explained?

We have argued that within-subject correlations are based on the use of various retrieval heuristics, such as speed (Costermans, Lories, & Ansay, 1992) or ease (Kelley & Lindsay, 1993) of retrieval from memory. Such heuristics are likely to be as accurate for eyewitness memory as for general knowledge, hence there is no difference for within-subject correlations. However, for between-subjects correlations to emerge for a population, there must be some consistency in the use of the confidence scale across people. When one person is “fairly sure”, they should be as accurate as another fairly sure person. Clearly, for eyewitness memory this is not the case, as no correlation is found, but it is the case for general knowledge. Why should this be?

We have argued that individuals are well calibrated with respect to each other for general knowledge because there is ample opportunity for individuals to learn their relative areas of expertise and weakness in general knowledge through

reference to textbooks, media, colleagues and so forth (see Juslin, 1994 for a similar view). People who share a culture therefore learn to become calibrated to the same standard and so become comparable. This view has been eloquently described by Allwood and Granhag (1996, p.26):

... the important conditions for the realism of confidence judgements are assumed to be located in the relation between the individual and his or her environment ... A stable relation between the individual and the environment allows relevant learning of the success probabilities for cues as means for finding answers to questions of a specific kind.

The result of the cultural learning described by Allwood and Granhag (1996) is that an individual learns not only their relative expertise from one domain to another, but also (roughly) where they stand in a population. For example a person who watches a lot of sport, but reads little, will know that they are likely to be well above average on sport questions, but below average on literature. Thus any speed or ease of retrieval heuristic can be anchored by knowledge of relative expertise. However, in eyewitness memory there is no such opportunity to revisit events to test one’s relative expertise. By their nature, eyewitness events are episodes that cannot be relived, and so there is no opportunity to learn calibration for these materials. Correspondingly, the between-subjects correlations for eyewitness memory are weak.

There is a clear hypothesis that stems from this view: providing people with an opportunity to learn their relative standing at eyewitness memory should significantly increase the confidence–accuracy relation. In contrast, practice without feedback will have no benefit, as being a witness is a continuous experience; our daily lives consist of a series of episodes and our social lives involve frequent attempts to remember the past. Merely providing participants with a formalised, forensic version of practice at memory tests will not give them the necessary learning opportunity to develop self-awareness about the accuracy of their episodic memory. Thus we believe that feedback on performance is crucial to the development of knowledge about relative ability at episodic memory tasks, and hence the confidence–accuracy relation across individuals.

There have been a number of previous attempts to provide eyewitnesses with feedback in order to improve the accuracy of their confidence judgements. Kassir (1985) demonstrated the ret-

rospective self-awareness effect in lineup decisions. In this paradigm, participants are video recorded when making their lineup decision. Later they make a judgement of the confidence in their choice. Seeing the video recording of themselves, prior to the confidence decision, significantly increased the magnitude of the confidence–accuracy association. Robinson and Johnson (1998) attempted a series of interventions designed to increase the confidence–accuracy relation. They examined the effects of (1) having participants publicly justify their decisions (accountability), (2) contextual reinstatement using photographs of the scene, (3) writing a retrospective narrative prior to their lineup decision, (4) having to circle the three most likely options in a lineup prior to a final choice, and (5) having to give one reason for and three against the lineup choice. They also investigated the personality factor of public self-consciousness. They found no support for any of the interventions, and concluded “... that attempts to enhance awareness of the thoughts and reasoning processes involved in an identification process may frequently have minimal, or even counterproductive, effects” (Robinson & Johnson, 1998, p.409).

More recently, Bornstein and Zickafoose (1999) showed that the degree of overconfidence across two domains (general knowledge and eyewitness memory) was related. They went on to give feedback to individuals about their overconfidence in general knowledge, as an attempt to improve the confidence–accuracy relation in eyewitness memory, measured across items. They found that this feedback, whether specific to the individual, or general about all individuals, reduced confidence overall, but did not improve the association between confidence and accuracy. They did not examine the confidence–accuracy association across individuals.

Thus, previous attempts to improve the confidence–accuracy relation have had mixed results at best. However, our approach differs from all the prior attempts in one crucial way. Whereas the previous work all focuses on the individual, our approach is concerned with how the individual compares themselves to the rest of the population. Given that the focus of most applied research is the relation between confidence and accuracy across individuals, this is an important distinction.

To test our ideas, three groups were contrasted on a video-based eyewitness memory test. A control group who had neither the practice nor feedback completed the critical trial only. Two

other groups received prior experience on two similar eyewitness memory tests: the practice group completed the tests but were not told how well they had done, the feedback group completed the tests, and were told how their performance compared to others.

EXPERIMENT 1

Method

Participants. A total of 57 participants were recruited from around the University of Bristol. Their ages ranged from 18 to 31 years, with a mean of 21.0 years. Of these, 18 participants were allocated to the control group, 19 to the feedback group, and 20 to the feedback group.

Materials. Three different videos depicting non-violent crimes were used. Video 1 lasted 110 seconds, Video 2 lasted 95 seconds, and Video 3 (the critical video) lasted 50 seconds. The videos were projected onto a large (120 cm × 120 cm) screen using a video projector. A cued-recall memory test was devised for each video. The first consisted of 12 questions, the second 13 questions, and the third 15 questions. Each question was accompanied by a 5-point confidence scale ($1 = \text{not at all confident}$ through to $5 = \text{very confident}$).

Control group. Participants were tested in small groups of up to four people at a time in a quiet testing room. They were instructed that they would see a video depicting a crime, and that they should try and remember as many details as possible. After they had seen the video (Video 3), participants filled out a personality test as a filler activity for 10 minutes before being given a written cued-recall test. They were instructed to answer all questions (e.g. “What colour was the car used for the crime?”), guessing where necessary in order to minimise the effects of criterion differences between individuals, and worked through the questions at their own pace.

Practice group. The procedure for the first video was identical to that for the control group, except that the first test taken by the practice group concerned Video 1. Participants were then told that they would see another video, and that they would be tested in the same manner. Testing continued for a total of three trials, with the final

trial (Video 3) being that used in the control condition. Participants completed a different personality test between witnessing the video and being tested, ensuring that a 10-minute delay intervened between presentation and test each time.

Feedback group. The procedure was the same as for the practice group with one exception. After the cued-recall tests had been completed, while they could still see their responses, participants were told the correct answers to each item on the test, together with the number of participants who had previously answered that question correctly. The data on the proportion correct for previous answers were taken from earlier studies using the same video materials in our laboratory.

Results

The material of particular interest in this Experiment was performance on Video 3, and is shown in the final column of the top half of Table 1. The mean proportion correct recall for all three groups was contrasted for this trial. There was an overall group effect, $F(2, 54) = 6.04$, $MSE = 0.01$, $p < .01$, with the practice group performing more poorly than the other two groups. It is hard to understand

why this should be. Because some authors have argued that difficulty is an important factor in the confidence–accuracy relation, we return to this issue in the discussion. There was no group effect on mean confidence in recall attempts ($F < 1$).

Our main focus is on the correlations between mean confidence and accuracy, and these data are shown in the final column of the top half of Table 2. Our prediction was that feedback is necessary for the development of adequate calibration for eyewitness memory, and so only the feedback group would show a confidence–accuracy relation. This is what was found. We also compared the correlation obtained for the feedback group with the correlation obtained for the other two groups combined ($r = 0.11$). This difference was marginally significant ($z = 1.48$, $p < .07$).

Because the practice and feedback groups had completed three trials, it was possible also to determine whether the confidence–accuracy relation showed improvement over the course of practice/feedback. Performance for these two groups only was therefore contrasted over the three trials. However, it should be noted that this analysis confounds practice with the materials under test, and any conclusions drawn can only be tentative.

For proportion correct recall there was no effect of group ($F < 1$) but there was an effect of

TABLE 1

Mean (and SD) proportion of items correctly recognised, and mean (and SD) confidence in choices for the three trials in Experiments 1 and 2

		Trial 1		Trial 2		Trial 3	
		M	SD	M	SD	M	SD
<i>Experiment 1</i>							
Control	Recall	–	–	–	–	0.51	0.08
	Conf	–	–	–	–	2.74	0.41
Practice	Recall	0.79	0.11	0.70	0.16	0.44	0.10
	Conf	3.68	0.51	3.65	0.49	2.59	0.40
F-B: Other	Recall	0.74	0.13	0.68	0.13	0.51	0.08
	Conf	3.68	0.61	3.54	0.55	2.74	0.41
<i>Experiment 2</i>							
Control	Recog	–	–	–	–	0.52	0.35
	Conf	–	–	–	–	2.50	0.89
Practice	Recog	0.59	0.35	0.44	0.31	0.63	0.37
	Conf	2.31	0.64	2.71	0.98	2.56	1.08
F-B: Self	Recog	0.52	0.33	0.58	0.31	0.60	0.32
	Conf	2.67	0.80	2.37	0.91	2.98	0.82
F-B: Other	Recog	0.57	0.34	0.63	0.27	0.59	0.32
	Conf	2.73	0.77	2.69	0.78	2.88	0.71

F-B: Other = Feedback given on own performance in comparison to others. F-B: Self = Feedback given on own performance only. Recog = mean proportion correct on the lineup recognition test. Conf = mean confidence in recognition decision.

trial, $F(2, 74) = 64.2$, $MSE = 0.01$, $p < .001$, and a significant interaction, $F(2, 74) = 4.33$, $MSE = 0.01$, $p < .05$. For mean confidence there was no effect of group ($F < 1$) but once again there was a trial effect, $F(2, 74) = 90.3$, $MSE = 0.14$, $p < .001$. There was no interaction. Mean proportion correct and mean confidence for these materials are shown in the top half of Table 1. The data indicate that performance on the final video was the poorest, and mean confidence reflected this.

The difference in the difficulty levels of the videos slightly complicates the interpretation of the confidence–accuracy relations for the three videos, shown in Table 2. Our expectation that the confidence–accuracy relation in the feedback group would show a steady increase from trial 1 to trial 3 was supported. However, the picture is muddled because the practice group also showed an improvement from trial 1 to trial 2, although this was not sustained for the critical third trial.

DISCUSSION

In line with many previous studies, the correlational data for the control group showed a non-significant association between confidence and accuracy. However, for the very same materials, witnessed in exactly the same manner, there was a confidence–accuracy relation for the feedback group as predicted. Also as predicted, practice did not raise the confidence–accuracy relation above the non-significant level seen for the control group.

TABLE 2
Pearson correlations between confidence and accuracy
across 3 trials for Experiments 1 and 2

	<i>n</i>	<i>Trial 1</i>	<i>Trial 2</i>	<i>Trial 3</i>
<i>Experiment 1</i>				
Control	18	–	–	–.10
Practice	19	.36	.55 ^b	.13
F-B: Other	20	.29	.46 ^a	.50 ^a
<i>Experiment 2</i>				
Control	28	–	–	.20
Practice	24	.10	.19	.07
F-B: Self	26	.04	.09	.19
F-B: Other	24	–.02	.33 ⁺	.50 ^b

F-B: Self = Feedback given on own performance only. F-B: Other = Feedback given on own performance in comparison to others. ⁺ $p < .10$, ^a $p < .05$, ^b $p < .01$, one tailed. Correlations for Experiment 1 are for average confidence against mean performance. Correlations for Experiment 2 are based on average z -transformed r s for male and female lineups separately.

Thus the data are entirely consistent with the calibration hypothesis outlined in the Introduction.

This interpretation is made slightly more complex by the fact that the practice group performed more poorly on the critical test. However, it is hard to see how level of performance on the criterion test relates to the confidence–accuracy relation in any straightforward manner. If higher confidence–accuracy relations stem from higher levels of performance, it is hard to see why the control group did not show the same confidence–accuracy level as the feedback group. It also seems unlikely that variance is the key factor, as variance in both confidence and accuracy was equivalent across the three groups. Thus we are satisfied that the poor performance of the practice group does not alter the conclusions based on the correlations in the critical trial.

Task difficulty also complicates the interpretation of the pattern across the three videos. Our expectation that practice alone would not have an effect appears to be compromised by the significant correlation between confidence and accuracy for trial 2 that is then absent for trial 3. It is hard to explain this pattern. If practice has an effect, one would expect to see it continue to the final trial. If practice has no effect, one would not expect the correlation for trial 2. Task difficulty does not seem to explain the pattern, as performance on trial 2 was higher than trial 3, which had a non-significant confidence–accuracy correlation, but lower than trial 1, which also failed to show a confidence–accuracy association.

In order to determine the robustness of the feedback effect as well as the practice effect in Experiment 1, we ran a second experiment. This differed from the first experiment in a number of key ways. Foremost of these was that we counterbalanced the materials across trials, so that they would be matched for difficulty across the different conditions. Another important change was that the second experiment focused on lineup performance, rather than event memory. We wished to do this to determine the forensic utility of the feedback effect, and to be able to relate our findings to the large number of studies that have examined the confidence–relation for identification tasks. Consequently, our measure of memory performance changed from cued recall, to forced-choice recognition.

We also decided to further explore the nature of the feedback effect observed in Experiment 1. The feedback in the first study consisted of telling participants whether their answers had been cor-

rect or not, together with information on the normative likelihood of success for an item. We were interested to see whether feedback on one's own performance, without learning how this compares to others, would be sufficient to produce the effect seen in the first study. Our expectation was that it would not, because feedback on performance alone does not allow people to learn how they compare with others, and so it is unlikely that they will adjust their confidence ratings appropriately. Consequently Experiment 2 had four conditions; a control and practice condition as before, and two feedback conditions, feedback on self, and feedback on self plus others, which mirrored the feedback condition in Experiment 1. As before, there were three trials in all, and we expected the confidence–accuracy relation to improve only for the group who learned how their performance compared to others.

EXPERIMENT 2

Method

Participants. Volunteers ($n=102$) were recruited from the University of Bristol, and local community. They ranged in age from 18 to 58 and were allocated randomly to one of four conditions: control ($n=28$), practice only ($n=24$), feedback on own performance ($n=26$), and feedback on self and others ($n=24$).

Materials. Photographs were downloaded from the University of Stirling pictures database (<http://pics.psych.stir.ac.uk>). A total of 18 three-quarter profile black-and-white photographs (9 male, 9 female) were downloaded, 6 of which were designated as targets. For these 6 target faces, 6 photographic lineups were constructed, using photographs that portrayed the face frontally, rather than three-quarters profile, and consisted of the target plus 4 similar distractors. The lineups were constructed so that the distractors were matched for ethnicity, and roughly matched for hair colour, hair style, and age.

Procedure. Participants were tested individually in a quiet testing room. Prior to testing the participants were told that the study concerned the attractiveness of male and female faces. Each participant was then shown a total of 18 photographs (9 male, 9 female) and asked to rate each for attractiveness on a 5-point scale (1=very

attractive, 5=not all attractive). Of these photographs, 6 (3 male, 3 female) were designated targets. Participants were asked to make a rapid rating, based on their first impressions. Following this self-paced task, the photographs were removed and the participants were informed that in fact the study concerned eyewitness memory, and that they would be tested on their memory for the people in the photographs. At this point the methodology differed for the different conditions.

Control condition. Participants were shown one male-only lineup and one female-only lineup, and asked to either identify which of the lineup members had been seen before, or indicate whether the lineup did not contain a person presented earlier. Following each decision participants were asked to rate the confidence in their decision on a 5-point scale (1=very confident, 5=not at all confident). The particular lineups used were counterbalanced across participants.

Practice condition. The procedure was identical to the control condition, except that participants conducted two further trials of the male-only and female-only lineups, rating confidence each time. The order of testing of the lineups across the factor of practice was counterbalanced across participants.

Practice plus feedback on self. The procedure was identical to the practice condition, with the addition that after rating their confidence, participants were told whether or not they had been correct in their decision.

Practice plus feedback on others. The procedure was identical to the practice condition except that after rating their confidence, participants were told whether or not they had been correct in their decision, and what proportion of participants normally got that answer correct. The data on previous success-rates were derived from analysis of the performance of the control group.

Results

The first analysis concerned the level of recognition performance across the four conditions. In theory performance should be matched across condition because there was counterbalancing of items, in contrast to Experiment 1. However, it was considered prudent to check that this was the

case. One issue to consider was whether the control condition should be compared to the first or last trial of the other conditions, because counterbalancing meant that in effect this is an arbitrary choice. Because her ultimate interest is in the confidence–accuracy relation after practice and feedback, we compared performance for the final trial for all groups, designating the single trial for the control group to be the final trial. Our counterbalancing had been successful: there was no effect of group for recognition ($F < 1$), nor for mean confidence, $F(3, 98) = 1.83$, $MSE = 1.43$. However, recognition was better and confidence higher for female lineups than male, $F(1, 98) = 11.44$, $MSE = 0.25$, $p < .001$; $F(1, 98) = 10.53$, $MSE = 1.07$, $p < .01$ for recognition and confidence respectively, but there was no interaction between group and gender of lineup for recognition or confidence, $F < 1$ for recognition, and $F(1, 98) = 2.30$, $MSE = 1.07$ for confidence.

Likewise, across the three groups who completed all three trials, there was no effect of Trial, $F < 1$, nor condition, $F < 1$, although female faces were recognised more readily than the male ones, $F(1, 71) = 21.79$, $MSE = 4.73$, $p < .001$. None of the interactions was significant (all F s < 1). The mean recognition performance on Trials 1 to 3 are shown in the bottom half of Table 1.

We decided to collapse the correlational data across male and female lineups before examining the effects of trial. Performance differed greatly for male and female lineups, and so comparison between them would be problematic. In any case we had no theoretical motivation for doing so. We therefore calculated the correlations separately for male and female lineups, and then took the average correlation (based on the average z -transform) for each trial. These data are shown in the bottom half of Table 2.

If we first consider performance on the critical third trial, it is clear that the data replicate Experiment 1. There is no confidence–accuracy relation for the control group, nor for the practice group. For participants who receive feedback on their own performance alone, there is no correlation, as expected. However, as in Experiment 1, the group who learn about their relative ability produced a robust confidence–accuracy relation for the critical final trial. In order to test this statistically, the correlation for the feedback on others was contrasted with the correlation obtained across the three other conditions ($r = 0.16$) using a z -test. The difference was marginally significant ($z = 1.38$, $p < .08$).

Because we have counterbalanced items across trials, it is also easier to examine the pattern across trials. Here, unlike Experiment 1, there is no anomalous correlation for the practice condition on trial 2. Similarly, the feedback on self group show no significant correlations between confidence and accuracy on any trial. Only the feedback on self and others group show the linear increase in the correlation predicted by the calibration hypothesis. We used Steiger's (1980) method to compare the magnitude of correlations in a related sample. Only the feedback with others showed a significant increase in the magnitude of the confidence–accuracy relation from trial 1 to trial 3 ($Z_2^* = 2.10$, $p < .02$).

Discussion

Experiment 2 was designed to overcome the confounding of test order and test materials present in Experiment 1, yet produced the same pattern of results as that study. Confidence–accuracy relations in the critical third trial are absent for the control condition and the practice condition, but are present when participants have received prior practice accompanied by feedback on their performance in comparison to others. Feedback on one's own performance alone is not sufficient to produce this effect. The same conclusions are reached if one compares the correlations across trials as practice/feedback develops. The feedback in comparison with others condition is the only one to show statistically reliable increases in the magnitude of the correlation from trial 1 to trial 3, and the magnitude of the final correlation for that condition is higher than is seen for all the other conditions combined.

Experiment 2 differed from Experiment 1 in a number of forensically relevant ways, and the replication of the effect suggests that the effect is a robust one. Experiment 1 tested memory for events, using video, with a cued-recall test. On the other hand, Experiment 2 tested memory for faces, using photographs, with a recognition test. Experiment 1 used a deliberate study phase in which participants knew they would be tested, whereas Experiment 2 used an incidental exposure technique in which participants did not know they would later be tested. All of these differences make the experimental paradigm closer to the real-world situation of an eyewitness. An eyewitness often does not know that they will be tested later. If they are tested for their memory for

faces, it is most often in the form of viewing photographs or live lineups. It is therefore reassuring that the feedback effect in Experiment 1 was replicated. We argue that this is because the locus of the effect is the way in which participants make judgements of confidence, rather than being a part of the memory process. We make this claim because the only manner in which the groups differ concerns their knowledge of their prior performance; there are no differences between the groups in how they attempt to study or retrieve the material, and no observable differences in how they perform on the memory test.

GENERAL DISCUSSION

The results of the two experiments are consistent. In standard eyewitness conditions, whether measured with cued recall for event memory, or with recognition for faces, there was no association between confidence and accuracy on the criterion test. However, after two trials of practice in which participants received feedback on their performance, together with information as to how their performance compared to others, there were reliable confidence–accuracy relations in both studies. The effects of practice alone were entirely absent in Experiment 2, suggesting that the confidence–accuracy relation for Trial 2 in Experiment 1 was a chance effect. Interestingly, feedback on one's own performance, in the absence of information about the performance of others, did not have a beneficial effect.

What is the mechanism for this effect? The calibration hypothesis that led to the current experiments was based on the premise that witnesses lack insight into their relative expertise. The fact that providing such insight produces a reliable confidence–accuracy association is therefore, on the face of it, strong evidence for the calibration hypothesis. However, further analysis leads us to modify our original calibration hypothesis.

The problem for a simple calibration account is that it contains the assumption that there is a reliable level of skill at eyewitness memory that participants can learn about from feedback. That is, people who are better than average at eyewitness memory will be so across trials, and will learn about this from the feedback they receive. Unfortunately, closer analysis of the data from the two studies suggests that this does not seem to be the case. In Experiment 1, the correlations

between performance on the 3 trials ranged from -0.04 to 0.28 , and for Experiment 2 the range was -0.05 to 0.05 . Clearly, in these experiments there was no consistent level of ability at eyewitness memory across trials. How then can feedback about one trial lead to more accurate metacognitive assessments on a subsequent trial, when performance on those two trials is unrelated? Clearly, the answer is not that people are learning how well they generally do at eyewitness memory, because there is no reliable index of general performance.

One account is as follows. As with the original calibration hypothesis, our starting assumption is that confidence judgements are related both to the ease or fluency with which retrieval was achieved, and to pre-existing beliefs about skill in the domain under test. Where this account differs from the simple calibration hypothesis is in the effect of feedback on the impact that beliefs have on confidence judgements. The original hypothesis was that feedback would make such beliefs more appropriate, but the lack of a stable level of ability renders this account untenable. Instead, it may be the case that participants learn that their beliefs are untenable, and so the relative weighting of the beliefs' impact on confidence judgements is reduced. Under these circumstances, participants' confidence judgements are based on memory strength more directly, uncontaminated by misleading beliefs about ability, and so a confidence–accuracy relation emerges overall for the feedback condition.

Why would this only happen when participants receive feedback about their relative performance compared to others? Practice alone would not reduce the weighting of prior beliefs because participants do not learn whether the beliefs were appropriate or not. People who believe themselves to be good at face identification will believe that they have selected the target face irrespective of their performance. Similarly, feedback on one's own performance in the absence of knowledge about how others perform is unlikely to alter beliefs because unexpected failure (or success) could be attributed to the test itself. Thus people who believe themselves to be good face-identifiers, who have learned that they selected the wrong face, could reason that the test was too difficult to be "fair" and that no-one would have succeeded. However, learning that they failed when most people succeed is difficult to explain away in this manner, and so is most likely to lead to a reduction in the impact of the prior beliefs.

Clearly, this account is speculative and requires further experimental support. It rests on the assumptions that prior beliefs about the domain under test have an impact on confidence judgements, that those beliefs are inappropriate for eyewitness memory, and that the relative impact of those beliefs can be reduced by feedback. However, we do have some data from our laboratory that support the first two assumptions. In two experiments we asked participants to rate their ability in eyewitness memory, prior to taking an eyewitness memory test, and rating their confidence in their answers. Consistent with the first two assumptions of the modified calibration hypothesis, participants' pre-test beliefs were predictive of their subsequent confidence ratings, but were unrelated to how well they perform.

Before discussing the practical implications of these data, it is perhaps worth reiterating that the feedback effect cannot be explained in terms of the item effects discussed in the Introduction. Because performance did not differ on the critical final trial in Experiment 2, arguments based on item difficulty or range of difficulty cannot explain the effect. Nor can the effect be explained in terms of quality of encoding, as the optimality hypothesis requires, or in terms of personality of the witnesses. The only difference across conditions was the nature of the feedback given, and thus any explanation of the effect must focus on the impact that feedback has on confidence judgements.

The technique used in Experiment 2 was cheap, quick, and inexpensive. Practice plus feedback consisted of a total of four photographs, and four photographic lineups, and yet it produced a robust effect on the confidence-accuracy relation for the critical trial. This suggests that the technique may have some forensic utility, although clearly much more work would be required before clear recommendations to investigators could be made with confidence. However, the data do suggest that prior exposure to a couple of simple tests, together with feedback on relative performance, mean that participants' judgements of confidence have much greater utility. There are many questions that remain for future studies: Does further practice produce further improvements? How long prior to the criterion test should practice take place? How well does practice generalise? What is the best sort of feedback to give? These questions, and others like them, form the basis of our ongoing research programme.

REFERENCES

- Allwood, C.M., & Granhag, P.A. (1996). Considering the knowledge you have: Effects on realism in confidence judgements. *European Journal of Cognitive Psychology*, 8, 235-256.
- Bornstein, B.H., & Zickafosse, D.J. (1999). "I know I know it, I know I saw it": The stability of the confidence-accuracy relationship across domains. *Journal of Experimental Psychology: Applied*, 5, 76-88.
- Bothwell, R.K., Deffenbacher, K.A., & Brigham, J.C. (1987). Correlations of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology*, 72, 691-695.
- Costermans, J., Lories, G., & Ansary, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18, 142-150.
- Deffenbacher, K.A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, 4, 243-260.
- 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.
- Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, 57, 225-246.
- Kassin, S.M. (1985). Eyewitness identification: Retrospective self-awareness and the accuracy-confidence correlation. *Journal of Personality and Social Psychology*, 49, 878-893.
- Kassin, S.M., Ellsworth, P.C., & Smith, V.L. (1989). The general acceptance of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist*, 44, 1089-1098.
- 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.
- Kelley, C.M., & Lindsay, D.S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, 32, 1-24.
- Lindsay, D.S., Read, D.J., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science*, 9, 215-218.
- Perfect, T.J., & Hollins, T.S. (1997). The confidence-accuracy relation in eyewitness event memory: The mixed question type effect. *Legal and Criminological Psychology*, 2, 205-218.
- Perfect, T.J., Watson, E.L., & Wagstaff, G.F. (1993). 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. Togliani (Eds.), *Basic*

- and applied aspects of remembering*. Hillsdale, NJ: 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 on the identification process. *Law and Human Behavior, 22*, 409-428.
- Robinson, M.D., Johnson, J.T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*, 416-425.
- 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.
- Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.
- Wells, G.L., Linssay, R.C.L., & Ferguson, T.J. (1979). Accuracy, confidence and juror perceptions in eyewitness identification. *Journal of Applied Psychology, 64*, 440-448.