

Accuracy of Confidence Ratings Associated With General Knowledge and Eyewitness Memory

Timothy J. Perfect, Emma L. Watson, and Graham F. Wagstaff

The confidence–accuracy (C–A) relation for general knowledge (GK) and eyewitness memory (EM) was compared in both within- and between-subjects analyses. Researchers in the cognitive tradition tend to use within-subjects designs and to find moderately positive C–A relations, whereas those in the forensic tradition tend to use between-subjects designs and to find no relation. Eighty subjects took part in one of two conditions—EM or GK. No difference between conditions was found on the within-subjects measure of the C–A relation, but there was differentiation with a between-subjects measure. There was a strong positive C–A correlation ($r = .58, p < .01$) for GK but not for EM ($r = -.11, ns$). One source of this difference may be the differing opportunities for calibration offered by the two kinds of memory.

There is a marked degree of agreement in the cognitive literature that there is a moderate yet robust positive relation between subjects' confidence evaluations and their performance. This relation exists over different populations, tasks, and experimental materials (see Nelson, 1988, for a review). However, Wells and Murray (1984), in a review of 28 studies of the relation between eyewitness confidence and eyewitness accuracy, found the average correlation between confidence and accuracy across all the studies to be an unimpressive .07.

Of course, there are many differences between laboratory-based work and the more forensically motivated research that might account for such a difference. Eyewitness memory is based on an event witnessed only once, under nonoptimal conditions, with incidental learning, perhaps even with a strong emotional element. Specific recall of details from episodic memory is required. All these aspects differ from the memory requirements of the typical feeling-of-knowing study, which evokes no emotional involvement and which tests memory for relatively familiar material. As well as differences in the kind of encoding that takes place, there are likely to be differences in the opportunities for confidence calibrations afforded by semantic and episodic memory (Wells, Lindsay, & Ferguson, 1979). For semantic memory, one has many opportunities to self-test one's ability at retrieving particular facts. Moreover, there are specified and agreed upon answers to general knowledge questions, so one not only can gauge how appropriate one's confidence in an answer should be but also can determine one's relative performance compared with others in the same situation. However, for eyewitness memory there is no way of knowing that a specific item retrieved is correct because the event cannot be revisited. There is no agreed upon answer to calibrate one's performance against, nor can one know whether one's ability to recall an event is better or worse than anyone else's

because no two witnesses are likely to have seen the same event from exactly the same perspective.

A further possible distinction is that in general knowledge (GK) tests, one is usually aware of one's strengths and weaknesses. For example, if asked one question about the rules of cricket and another about American presidents, British subjects might feel quite confident about the former but totally unconfident about the latter. The reverse would probably be true for American subjects. However, in episodic tasks like eyewitness memory, one rarely has expertise in one subdomain, such as recollecting what someone was wearing. So there seems to be a qualitative difference between semantic memory, which can be divided into areas of expertise, and episodic memory, which cannot be subdivided. However, there is one difference between the cognitive and forensic literatures that preempts the preceding speculation, which is that the majority of studies in the cognitive literature have assessed metamemory within individuals, whereas the majority of studies in the forensic literature have assessed metamemory accuracy between individuals. The standard cognitive approach involves collecting ratings from an individual for a set of materials and then calculating the relation between confidence and accuracy for that person over a reasonably large number of trials (e.g., Lichtenstein & Fischhoff, 1977). For pragmatic reasons, the forensic approach is somewhat different. Typically, a large number of individuals witness a single event, and then the performance of individuals who are high in confidence is compared with the performance of individuals who are lower in confidence; thus, the confidence–accuracy (C–A) relation is determined for the group as a whole. Therefore, without an adequate empirical demonstration of an inferior C–A relation in eyewitness memory under the same conditions as a cognitive test, speculation about its cause may be premature.

Only one study of the C–A relation in eyewitness memory has looked at this methodological issue directly. Smith, Kassin, and Ellsworth (1989) reported uniformly low ($r < .2$) correlations between confidence and performance for ratings taken both within and between subjects. However, there are problems in accepting Smith et al.'s conclusion that "confidence is not a good predictor of accuracy" (p. 358). In their study, eyewitness

Timothy J. Perfect, Emma L. Watson, and Graham F. Wagstaff, Department of Psychology, University of Liverpool, England.

We would like to thank three anonymous reviewers for their helpful comments on a previous draft of this article.

Correspondence concerning this article should be addressed to Timothy J. Perfect, Department of Psychology, University of Liverpool, P.O. Box 147, Liverpool L69 3BX, England.

memory was assessed with a two-choice recognition procedure, yet the hit rate was only 63%. Because 37% of the guesses were incorrect by chance, this implies that 37% of the guesses would have been correct, leaving the number of items actually remembered at 26% ($63\% - 37\%$), which is less than the number correct by chance. This would substantially reduce the C–A relation because subjects would have been correctly indicating low confidence on the majority of items that they got correct. Therefore, Smith et al. (1989) may be unwarranted in drawing their strong conclusion about the C–A relation from their null effect because this is in part due to the difficulty of the questions and the high guessing rate in their study.

The rationale behind the present study was to develop a test of eyewitness memory that is procedurally similar to the cognitive approach without the shortcomings identified in Smith et al.'s (1989) study. The relative accuracy of the C–A relation for eyewitness memory could then be compared with semantic memory using the same methodology. One further factor was included in the design. A potentially important issue in estimates of the C–A relation is whether people make confidence judgments entirely on the basis of self-generated information or whether they utilize information provided externally. For example, if one were asked to choose an attacker from a lineup, it is presumably a different matter to make a confidence judgment before seeing the candidates than to make a confidence judgment after having seen the full lineup. In the former case, the evaluation can only be based on recollection of the prior episode, whereas in the latter, one can evaluate the plausibility of the distractors as well as the strength of familiarity produced by the identified individual. We made no strong prediction regarding the accuracy of prospective and retrospective judgments in this study. To examine these issues, we tested two groups of 40 subjects; one group was tested on general knowledge, the other on eyewitness memory. Each group was presented with 35 questions, and subjects rated their confidence in their ability to answer before attempting a multiple-choice recognition version of the test. After selecting an answer from five choices, the subjects rated their confidence once again. This enabled the examination of the C–A relation both between and within subjects for the two kinds of memory as well as the examination of the accuracy of confidence decisions made in the absence of external cues (i.e., prior to recognition) and following presentation of the cue embedded in a set of distractors (i.e., after the recognition decision).

Method

Subjects and Materials

Potential university undergraduates visiting at an open university day ($N = 80$) were randomly allocated to two equally sized conditions—a GK condition and an eyewitness memory (EM) condition. Thirty-five experimenter-generated general knowledge questions formed the test material for the GK condition (e.g., “Who wrote *The Mill on the Floss*?”). Thirty-five experimenter-generated questions based on a 30-min video clip of the film *Midnight Express* were used in the EM condition (e.g., “What was the name on the carrier bag Billy was carrying?”).

Procedure

Subjects in each condition were tested as a group in a large lecture room. Subjects in the EM condition were first shown a 30-min clip of

the feature film with the instruction to watch carefully, but no indication was given that there would be a subsequent memory test. All subjects were given a sheet of 35 typewritten questions (general knowledge for the GK condition and questions based on the film for the EM condition) and for each question were asked to rate how sure they were that they knew the answer, on a scale of 1 to 5 (1 = *very confident*, 2 = *fairly confident*, 3 = *moderately confident*, 4 = *not at all confident*, and 5 = *no idea*). This rating formed the prospective judgment. After subjects completed the full 35 items, these responses were removed, and the subjects were given five-alternative multiple-choice versions of the same questions. Subjects were instructed to answer every question, guessing if necessary, and to rate their confidence in the answer they had chosen on the same 5-point scale as before. The confidence ratings that followed selection of the answers were called retrospective judgments.

Results

Between-Group Comparisons

The first analyses investigated whether there were any differences in the overall difficulty of the tests or in the confidence ratings they elicited. The aim was to rule out any scaling effects that could confound differences in the C–A relation between the conditions. The mean number of correct answers on the multiple-choice recognition versions of the tests were 14.45 ($SD = 4.6$) and 14.65 ($SD = 3.5$) out of 35 for the GK and EM subjects, respectively. This difference was not significant ($F < 1$). The mean confidence rating elicited by each test, before and after the recognition test, was analyzed with a two-way (Test Type \times Time of Rating) analysis of variance (ANOVA). There was no effect of test type ($F < 1$), but the time of rating effect was significant, $F(1, 78) = 5.49$, $p < .05$, such that ratings made after the recognition test tended to indicate higher confidence. There was no interaction. The mean ratings elicited by the GK test were 3.36 ($SD = 0.50$) and 3.09 ($SD = 0.61$) for the prospective and retrospective ratings, respectively, compared with ratings in the EM condition of 3.24 ($SD = 0.47$) and 3.13 ($SD = 0.47$) for the same two times.

Having established that the two tests elicited similar overall levels of performance and confidence, we then addressed the main issue of interest, namely, the relation between performance and confidence. This was analyzed in two ways. First, a $2 \times 2 \times 5$ (Condition \times Time of Rating \times Confidence Rating) ANOVA on proportion of correct answers was conducted to compare confidence ratings made before and after recognition for the two kinds of test material. There were significant main effects of confidence rating, $F(4, 248) = 45.77$, $p < .01$, indicating that higher confidence was associated with better performance, but no main effect of time of rating ($F < 1$) or condition, $F(1, 62) = 1.01$, *ns*. No interactions were significant. The proportions of correct answers at each level of confidence are shown in Table 1.

The use of confidence ratings was also evaluated overall by means of a Goodman-Kruskal gamma correlation, as recommended by Nelson (1984). A 2×2 (Group \times Time of Rating) ANOVA on gamma scores revealed no main effect of group ($F < 1$) but a significant main effect of time of rating, $F(1, 78) = 47.23$, $p < .01$, and a significant two-way interaction, $F(1, 78) = 4.04$, $p < .05$. The mean gamma correlations in the GK condition were .29 ($SD = .23$) and .43 ($SD = .22$) for the prospective

Table 1
Proportion of Correct Answers at Each Level of Confidence for Ratings Made Before and After a Five-Alternative Multiple-Choice Recognition Test in the General Knowledge (GK) and Eyewitness Memory (EM) Conditions

Confidence	Prospective rating		Retrospective rating	
	GK	EM	GK	EM
1 (high)	.69	.57	.70	.70
2	.48	.45	.49	.59
3	.43	.43	.40	.40
4	.33	.33	.31	.32
5 (low)	.31	.31	.35	.20

and retrospective ratings, respectively. The corresponding values for the EM condition were .24 ($SD = .23$) and .49 ($SD = .23$). All mean gamma values were significantly above zero. Overall, the C-A relation was stronger after the recognition test. Although the means suggest that this effect was greater for the EM group, post hoc Scheffé tests did not reveal any differences between the groups on either prospective or retrospective judgment accuracy, despite the significant interaction.

Within-Group Comparisons

Three further analyses were conducted, this time within each group, to investigate the relation between recognition performance on a particular test with metamemory performance. However, although the two tests were matched for mean performance, the GK test had a wider range of scores (7–28) than the EM test (8–24). Because this would mean that within-group correlations for the EM group might be attenuated compared with those of the GK group, the two highest scoring and lowest scoring individuals from the GK condition were removed to match for range. Hence, in the following analyses there are only 36 subjects in the GK condition.

Pearson correlation coefficients were calculated between performance on the criterion test and overall metamemory accuracy (as estimated by gamma) to estimate the relation between confidence and accuracy across subjects. As Table 2 indicates, the gamma correlations are unrelated to level of performance on either test.

For each subject, the mean confidence rating was calculated for both prospective and retrospective confidence ratings. Then, for each condition, the correlation between overall performance and mean confidence rating was determined. These correlations are given in Table 2. For the GK group, there was a strong relation between confidence and performance on both prospective and retrospective evaluations. However, there was no relation between confidence and performance for the EM group. For the prospective measures, the GK correlation was reliably higher than the EM correlation ($p < .01$), but the two did not differ reliably on the retrospective measure.

Because confidence was measured before and after the recognition test, it was possible to see whether there was any interindividual stability in confidence levels for both the GK and EM conditions, that is, whether confidence levels before the recognition test were altered when subjects were shown the potential answers. For GK subjects, the mean level of confidence before

the test was highly correlated with mean level of confidence after the test, $r(34) = .58, p < .01$. This means that the presentation of the target and distractors did not radically alter subjects' confidence. However, for EM subjects there was no such stability, $r(38) = -.11, ns$, indicating that subjects' confidence in their answers was completely altered by the presentation of the alternatives. The two correlations were significantly different ($p < .01$). These data perhaps suggest that a witness's confidence that he or she will be able to identify a suspect before seeing a lineup may be unrelated to his or her confidence in the decision made on the basis of the lineup.

Discussion

The primary purpose of this study was to answer the pragmatic question of whether subjects' assessments of confidence in an answer are as accurate for events they have witnessed as for general knowledge if the same procedure is utilized. A further goal was to examine whether subjects' confidence judgments are more accurate before or after they are offered a choice of answers. On this second point, the results were quite clear; the subjects in both conditions were more accurate making retrospective confidence ratings. Although investigation of the basis for each kind of confidence judgment was beyond the scope of the present study, the results indicate that any account of differences between prospective and retrospective judgments of confidence will have to explain the different patterns of performance in the GK and EM conditions. Despite the fact that the two conditions were matched for mean recognition level (and variance), mean confidence rating, and mean accuracy of confidence ratings both prospectively and retrospectively (gamma), in the GK condition there was a strong positive correlation between the mean levels of confidence before and after the recognition test whereas in the EM condition there was no such stability.

In the rest of the discussion, we focus on the overall differences in the C-A relation between the GK and EM conditions. The results at the group level support the view that metamemory for an eyewitnessed event is just as accurate as for general knowledge. However, if one looks at individual differences within the two conditions, two interesting differences are found. As discussed earlier, in the GK condition overall the use

Table 2
Within-Group Pearson Correlations Between Number of Items Correct and Two Measures of Metamemory (Gamma and Mean Confidence Rating) for Ratings Made Before and After the Recognition Test

Metamemory measure	Prospective rating	Retrospective rating
Gamma		
GK	.27	.30
EM	.09	-.03
Mean confidence rating		
GK	-.43*	-.45*
EM	.18	-.16

Note. GK = general knowledge condition; EM = eyewitness memory condition.

* $p < .01$.

of confidence ratings was stable across the two times of measurement, and there was a positive correlation between confidence and memory performance. This was not the case in the EM condition. Individuals' mean level of confidence was unrelated to the overall amount of information they knew, and the use of confidence ratings was less stable across the two times of measurement. Analyses at the level of individual differences support the split observed in the literature; in eyewitness test situations, the overall level of memory performance is unrelated to level of confidence. This is intriguing because the mean scores were virtually the same across conditions. So there is an apparent paradox. At the group level, metamemory was just as accurate in the EM condition as in the GK condition but in the EM condition alone, mean confidence was unrelated to performance. How can this be explained?

One distinction is the difference between relative and absolute ratings of performance. Within-subjects analyses based on gamma are measures of relative confidence; that is, gamma measures the degree to which there is an association between the ordering of the confidence ratings (high-low) and memory performance (high-low). Metamemory accuracy as measured by gamma is logically unrelated to performance (Nelson, 1984). This should be so because it is as valid to claim ignorance as knowledge, as long as the claim is accurate. When gamma is used as a measure of relative metamemory accuracy (as recommended by Nelson, 1984), one finds no difference between the GK and EM conditions. However, at the between-subjects level, the mean level of confidence was unrelated to EM performance. This between-subjects analysis is an absolute measure of confidence because it is based on the relation between confidence and actual recognition scores. Although subjects in the EM condition could discriminate between items (make relative judgments), the lack of a between-subjects correlation indicates that they had no anchor point for this discrimination (i.e., they could not make absolute judgments). This contrasts with the GK tests, for which subjects' relative judgments had both a relative and an absolute basis, as demonstrated by the fact that there was a between- as well as a within-subjects correlation between confidence and performance.

Perhaps one key difference between the two kinds of tasks is the opportunity to receive feedback on one's performance. People continuously attempt to retrieve such information in general life and through this experience learn to calibrate confidence with performance. This process provided the subjects in this study with a yardstick with which they could estimate their likely success on the test. However, people do not often receive feedback about their eyewitness memory. In contrast to general knowledge tests, there is often no specific answer for a question about an event or individual. If one is asked to remember a particular event (e.g., a mugging), it is difficult to determine how good a representation one can generate. There is no objective yardstick to compare recall with, and memory for an event is inherently subjective and personalized. How can people decide if they have generated the right answer? Certain facts may be verifiable, but not all. For some information (e.g., the mugger), the only test is whether recognition is successful at a later date. However, even this is unsatisfactory as feedback because it is a very crude way of distinguishing memory for two individuals. Furthermore, one may see the attacker but fail to recognize him (or her); this would be a recognition miss, and confi-

dence would be unaltered. Hence, the opportunities to calibrate confidence and performance on eyewitness memory tests are sparse, and so there is no correspondence between confidence and absolute performance in eyewitness memory. This argument is similar to that made by Wells et al. (1979), who proposed not only that there is little opportunity for calibration but also that everyday experience may serve to maintain poor calibration: If one believes that one recognizes a passerby and smiles at that person, the person will most often smile back. This is received as confirmation of the initial recognition when in fact the person may be a total stranger.

Ideally, the findings reported here should be replicated in a more applied setting. However, what we believe this work shows is that what needs to be explained is not why there is no C-A relation in eyewitness memory, as previously reported (Smith et al., 1989; Wells & Murray, 1984), but why relative C-A judgments (within subjects) are accurate whereas absolute C-A judgments (between subjects) are not. It is hard to see how the various factors outlined previously could predict a relative C-A relation that was equivalent in EM and GK conditions but an absolute C-A relation that was stronger for a GK condition than for an EM condition. For example, if knowledge of one's expertise in an area were important in determining the strength of the C-A relation, one would predict that the C-A relation would be stronger for general knowledge than for eyewitness memory on both relative and absolute measures because the inability to judge one's area of relative expertise in eyewitness memory would affect both relative and absolute judgments. A similar argument could be made for other distinctions between forensic and cognitive tests, such as emotional content, degree of incidental learning, and so forth. The only explanation that appears to be able to distinguish between relative and absolute judgments is one based on opportunities for calibration.

References

- Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20, 159-183.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95, 109-133.
- Nelson, T. O. (1988). Predictive accuracy of feeling of knowing across different criterion tasks and across different subject populations and individuals. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 190-196). Chichester, England: Wiley.
- Smith, V. L., Kassin, S., & Ellsworth, P. E. (1989). Eyewitness accuracy and confidence: Within- versus between-subjects correlations. *Journal of Applied Psychology*, 74, 356-359.
- 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.
- Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155-170). Cambridge, England: Cambridge University Press.

Received October 8, 1991

Revision received May 18, 1992

Accepted May 21, 1992 ■