Applied Cognitive Psychology, *Appl. Cognit. Psychol.* 26: 289–295 (2012)

Confidence–Accuracy Calibration with General Knowledge and Eyewitness Memory Cued Recall Questions

KARLOS LUNA[1]* and BEATRIZ MARTÍN-LUENGO[2]

[1]*Centre for Research in Psychology (CIPsi), University of Minho, Portugal*

[2]*Faculty of Psychology, University of the Basque Country, Spain*

Summary: The confidence–accuracy relationship has primarily been studied through recognition tests and correlation analysis. However, cued recall is more ecological from a forensic perspective. Moreover, there may be more informative ways of analysing the confidence–accuracy relationship than correlations. In the present study, participants viewed a video of a bank robbery and were asked cued recall questions covering general knowledge and the video itself. Confidence ratings were collected, and correla- tions, calibration and discrimination measures were calculated. All measures indicated a strong confidence–accuracy relationship that was better for general knowledge than eyewitness memory questions. However, there were no differences in confidence ratings for correct answers, suggesting that the differences could be limited to the evaluation of incorrect answers. We concluded that confidence may be a good marker for accuracy with cued recall, but that further research using ecological tests and more informative data analysis techniques is needed.

Confidence in answers has been studied as a possible predictor of actual performance. Research on the confidence–accuracy relationship has been conducted using both general knowl- edge and eyewitness memory questions. Studies involving general knowledge questions have primarily been aimed at testing hypotheses and theories about how metamemory judgments work, whereas studies posing eyewitness memory questions have focused on application, based on the hypoth- esis that in a forensic context, confidence in answers helps distinguish between correct and incorrect information in a police interrogation.

However, research conducted to date on the confidence– accuracy relationship has some important shortcomings. First, it has primarily made use of recognition memory tests (e.g. Baranski & Petrusic, 1995; Bornstein & Zickafoose, 1999; W.F. Brewer & Sampaio, 2006; Mengelkamp & Bannert, 2010; Migueles & García-Bajos, 2001; Perfect, Watson, & Wagstaff, 1993; Schneider & Laurion, 1993), thus limiting the generalization of results to other contexts, such as an initial police interrogation in which the goal is to determine what happened during the offense or crime and in which no alternative answers are provided. The cued recall test is more commonly used in these cases than the recognition test. A second limitation is that historically the confidence– accuracy relationship has been studied by means of correla- tion, a technique that may not be the most informative. Finally, in reference to eyewitness memory research, most of the stud- ies on the confidence–accuracy relationship have been in the area of eyewitness identification (e.g. N. Brewer, Keast, & Rishworth, 2002; N. Brewer & Wells, 2006; Juslin, Olsson, & Winman, 1996; Weber & Brewer, 2006; for an extensive review, see Leippe & Eisenstadt, 2007), with little interest in what happens with confidence in the recall of complex events. Referring to the confidence–accuracy relationship, Hollins and Perfect (1997) suggested that the results of iden- tification studies may not be generalized to what they called eyewitness event memory, i.e. the memory of what happened. To address these limitations, we conducted an experiment with the explicit aim of studying the confidence– accuracy

relationship with general knowledge and eyewitness memory questions using a cued recall memory test.

Several authors have pointed out that metamemory judg- ments, and therefore confidence ratings, are mainly based on by-products of the retrieval process, such as the amount and intensity of retrieved information (Koriat, 1993), the fluency of retrieval (Hertzog, Dunlosky, Robinson & Kidder, 2003; Kelley & Lindsay, 1993), the vividness or completeness of the memory (W. F. Brewer, Sampaio, & Barlow, 2005; Robinson, Johnson, & Robertson, 2000) or the response la- tency (Weber & Brewer, 2006). These by-products could be more influential using a cued recall test in which partici- pants must generate their own response alternatives, than a recognition test in which they are only required to select one of several alternatives provided. The process of generat- ing alternatives may make the by-products stronger or more accessible, thus leading to a better estimation of the confi- dence ratings with cued recall than recognition. In this line of research, Robinson and Johnson (1996) and Robinson, Johnson and Herndon (1997) found higher confidence– accuracy correlation for recall than for recognition tests.

Another variable known to affect the confidence–accuracy relationship is the study material used, i.e. general knowl- edge or eyewitness memory questions. Most of the research that has compared the confidence–accuracy relationship us- ing general knowledge and eyewitness memory questions has found that the correlation between confidence and accu- racy is better with the former than with the latter (Hollins & Perfect, 1997; Perfect, 2004, Experiment 2; Perfect & Hollins, 1996). This result is explained by participants' lack of insight at eyewitness memory contents. However, all of these studies used correlations, a data analysis technique that has not gone unchallenged. Several authors have pointed out that correlations may not be the best way to study the confidence–accuracy relationship, proposing an alter- native technique: calibration (Baranski & Petrusic, 1995; Juslin et al., 1996; Olsson, 2000; Weber & Brewer, 2003; Weingardt, Leonesio, & Loftus, 1994, Wells & Lindsay, 1985). Calibration refers to the extent to which confidence judgments match the actual probability that an answer is cor- rect, i.e. when 100% of the answers with a confidence rating of 100 (on a 0–100 scale) are correct, when 90% of the answers with a confidence rating of 90 are correct and so on. Calibration has several advantages over correlation. Spe- cifically, correlation has been criticized for not being sensi- tive and informative enough to be useful to measure the confidence–accuracy relationship. With regard to sensitivity, Olsson and Juslin (2002) showed that a low correlation was compatible with both overconfidence (when the subjective probability that the answer is correct, i.e. confidence, is higher than the objective probability, i.e. accuracy) and underconfidence (when the subjective probability is lower than the objective probability). Regarding informativeness, calibration allows easy visualization of the confidence– accuracy relationship by means of calibration curves, i.e. the graphical plotting of confidence and accuracy in X and Y axes, respectively (for an example, see Figure 1 below). By analysing calibration, it can also be determined whether a given confidence level is a good predictor of accuracy when over/underconfidence is low. This information might be of more use in a trial to evaluate the likelihood that a specific piece of information provided by a witness is correct. Only one study, to our knowledge, compared calibration using general knowledge and eyewitness memory questions, although based on a recognition memory test. Bornstein and Zickafoose (1999, Experiment 2) found similar calibrations with general knowledge and eyewitness memory questions, most probably because before answering the eyewitness memory questions, participants were told that there was overconfidence in the previous general knowledge task. This instruction reduced the participants' overconfidence on

the eyewitness memory questions, thus improving calibration With cued recall test and eyewitness memory questions, only one study calculated calibration. Odinot and Wolters (2006) showed a video that ended in a traffic accident, and asked 22 open-ended questions several times. Observed calibration was apparently similar to the theoretically perfect, although the authors simply presented the calibration curve without quantifying it, thus missing some of the relevant information that calibration may have offered.

Calibration analysis allows a detailed examination of the confidence–accuracy relationship with general knowledge and eyewitness memory questions. Following the hypothesis of the lack of insight into one's ability with eyewitness memory questions, we would expect a better overall calibration with general knowledge than with eyewitness memory questions. However, in some cases participants may have good insight into their performance on a specific question, regard- less of the material. For example, it might be easier to evaluate the confidence in very easy or very difficult questions in a cued recall test because the answers may be accompanied by several or none of the aforementioned by-products of the retrieval process (e.g. fluency, vividness or response latency). In support of the idea that participants may very well know when they do not know the answer, Glucksberg and McCloskey (1981) found that answers were faster and more accurate when participants did not know the answer than when they were explicitly told that the answer was unknown. Thus, with eyewitness memory, we expected low underconfidence at the lower end and low overconfidence at the higher end of the calibration curve when accuracy is either very low or very high. Over/underconfidence with general knowledge questions on those levels, which is expected to be close to the perfect calibration, will serve as a comparison of the over/underconfidence with eyewitness memory questions. To test these hypotheses, we conducted an experiment in which participants watched a video of a bank robbery, and were then asked to answer and rate their confidence for 40 cued recall general knowledge questions, followed by 40 eyewitness memory questions about the video.

METHOD

Participants and design

Fifty-three students (eight male) from the Faculty of Psychology of the University of the Basque Country, with age range of 18–39years (*M*=18.34, *SD* =3.11), completed this experiment as a course requirement. The only variable was the type of question: either general knowledge or eyewitness memory, and was manipulated within subjects.

Materials and procedure

In order to select general knowledge questions over the entire range of difficulty, we conducted a preliminary normative study. Eleven participants (five males), with age range of 21–46years (*M*=32, *SD* =7.46), completed this study, none of whom took part in the main experiment. We selected 57 cued recall questions based on encyclopaedias, the Trivial Pursuit Genus Edition board game and the authors' general knowledge—three sources commonly used to create general

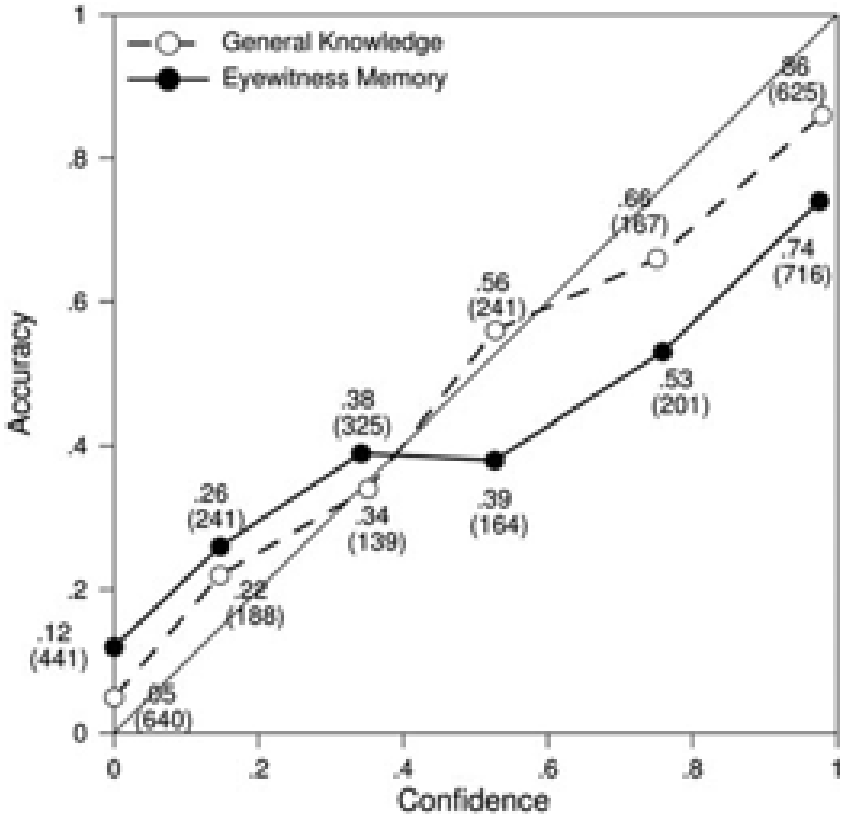

Figure 1. Calibration curves for general knowledge and eyewitness memory questions. Diagonal line depicts perfect calibration. The data above the diagonal show underconfidence, and the data below the diagonal show overconfidence. The accuracy mean (number of observations) in each category appears beside each data point

knowledge questions (Baranski, 2007; Migueles & García- Bajos, 2001). Some of the questions that were too easy (accuracy=1) or too difficult (accuracy=0) were removed, although some were included in the final list of 40 questions used in the experiment. This was done to include questions covering the entire range of difficulty. The decision whether or not to remove a question was determined by chance. The eyewitness memory questions were drawn from a pool of questions about a bank robbery developed in our laboratory (Luna & Migueles, 2007, 2008, 2009). By examining the ac- curacy of the answers, we selected 40 questions that covered the entire range of difficulty for the main experiment. All 80 questions involved short one- or two-word answers.

All of the participants were tested at the same time. They were told they were about to watch a video and would then be asked to answer some questions. In the first phase, a 3-minute excerpt from the film 'The Stick-Up' (Herrington, 2002) depicting a bank robbery was projected. In the video, two security guards unload sacks of money from an armoured vehicle, place them in a safe deposit room and drive away. A bank robber stationed nearby cuts off the power supply to the building, walks into the bank in dis- guise, threatens the people inside and makes off with the money. The incident unfolds with no explicit violence.

After the video, the participants received a questionnaire with the 40 general knowledge questions (e.g. What family would be your enemy if you were a Capulet? Answer: the Montagues). Apart from the answer, they were asked to rate their confidence that their answer was correct on an 11-point percentage scale ranging from 0 to 100 in intervals of 10. If participants did not know the answer, they were instructed to write something down and assign their answer with a confidence rating of 0. This forced participants to make an extra effort to retrieve the answer and prevented the tendency to leave questions for which the response did not come easily to mind unanswered.

No time limit was imposed for completing the general knowledge questionnaire. After about 20minutes, when all of the participants had finished, they received a second questionnaire containing 40 questions about the bank robbery video (e.g. When the robber is in the car, what is he holding? Answer: a watch). As for the general knowledge questions, we asked participants to rate their confidence on a scale of 0%– 100% with no time limit. Finally, the experimenters explained the main objectives of the research and answered the participants' questions. The session lasted approximately 1hour.

RESULTS

Several analyses were conducted to provide a comprehensive examination of the confidence–accuracy relationship. Preliminary analyses are first conducted to help characterize the data. Correlations are then presented, followed by calibration and associated indices of calibration and discrimination. All the analyses were performed with the R software (R Development Core Team, 2008). Unless otherwise indicated, significance level for all analyses was α=.05.

Preliminary analyses

Accuracy was higher with general knowledge (M=0.48, SD =0.11) than eyewitness memory questions (M=0.43, SD =0.10), t(52)=2.40, p=.020, d=0.42. Accuracy rates vary depending on the difficulty of the task, and difficulty affects metamemory (Baranski & Petrusic, 1995; Kebbell, Wagstaff, & Covey, 1996). Indices such as resolution, calibration or over/underconfidence worsen as the difficulty of the task increases and, as a consequence, accuracy decreases (Baranski & Petrusic, 1995). To avoid any explanation of our data in terms of differences in the difficulty of the materials, we removed two very easy general knowledge questions to match the accuracy with both materials. Thus, with 38 general knowledge and 40 eyewitness memory questions, accuracy was similar (M=0.46, SD =0.12 and M=0.43, SD =0.10, respectively), t(52) =1.06, p=.29, d=0.19. Confidence, however, was higher with eyewitness memory (M=53.37, SD =13.40) than general knowledge (M=47.09, SD =13.84), t(52)=—3.32, p=.002, d=0.47.

Two pairwise comparisons with the Student's t-test on confidence for correct and incorrect answers showed that confidence for correct answers for general knowledge and eyewitness memory questions was similar, t(52)=1.42, p=.163, d=0.19 (see Table 1).[1] However, confidence was higher for incorrect answers with eyewitness memory than with general knowledge questions, t(52) =8.36, p<.001, d=1.10.

Correlations

Two correlations were calculated: between-subjects (Pearson) and within-subject (Goodman-Kruskal gamma). Both correlations ranged from +1 to —1, reflecting, respectively, a perfect positive and perfect negative relationship. If correlation equals 0, then the relationship is null.

All the correlations are shown in Table 1. The Pearson's correlation was computed using the accuracy and confidence means for each participant. Both Pear- son correlations were significantly different from zero, $t(51)=6.02$, $p<.001$, 95% confidence interval (CI), [0.45, 0.78] for general knowledge and $t(51)=2.48$, $p=.016$, 95% CI [0.06, 0.55] for eyewitness memory, and higher for the former than the later, $z=2.14$, $p=.034$. Gamma was computed for each participant and material and then averaged. Gammas were high and significantly different from 0, $t(52)=54.67$, $p<.001$, 95% CI [0.80, 0.86] for general knowledge and $t(52)=27.32$, $p<.001$, 95% CI [0.58, 0.67] for eyewitness memory, respectively, and higher for general knowledge than eyewitness memory questions, $t(52)=8.21$, $p<.001$, $d=1.43$.

Calibration between confidence and accuracy

In order to be sufficiently informative, calibration must be based on a large set of data points, approximately 200 in each confidence level (Juslin et al., 1996). To increase this number of observations, we reduced the 11 confidence categories to six. Thus, we collapsed confidence levels 10–20, 30–40, 50–60, 70–80 and 90–100. Confidence level 0 was not collapsed with any other level because of the instructions given to participants to write something and to use this

[1] However, when the confidence mean for correct answers for the 40 general knowledge questions was computed ($M=77.49$, $SD=10.36$), the difference was significant, $t(52)=2.72$, $p=.009$, $d=0.36$. This result is consistent with the better monitoring of the correctness of an answer when accuracy is higher and, therefore, difficulty is lower.

Table 1. Means (standard deviations) from the main measures

	General knowledge	Eyewitness memory
Confidence in correct answers	75.45 (10.96)	73.05 (14.02)
Confidence in incorrect answers	22.95 (13.01)	37.80 (14.04)
Pearson's correlation	.64	.33
Gamma correlation	.83 (.11)	.63 (.17)
Calibration index (*C*)	0.041 (0.03)	0.070 (0.06)
ANDI (discrimination)	0.448 (0.17)	0.210 (0.14)

ANDI, Adjusted-Normalized Discrimination Index

confidence level when they did not know the answer. Both general knowledge and eyewitness memory calibration curves can be seen in Figure 1. At a descriptive level, calibration with general knowledge questions is closer to the diagonal than with eyewitness memory questions. For both general knowledge and eyewitness memory questions, there is slight underconfidence in the lower levels of confidence. In the higher levels, there is slight overconfidence for general knowledge and substantial overconfidence for eyewitness memory questions.

To quantify calibration and examine the differences be- tween general knowledge and eyewitness memory questions, the Calibration index (*C*; see N. Brewer et al., 2002 for the calculations), and a simple measure of over/underconfidence (confidence minus accuracy) were computed. *C* was used to compare overall general knowledge and eyewitness memory calibration. A *C* of 0 indicates perfect calibration, and higher values indicate worse calibration. The main statistics are pre- sented in Table 1. The results showed better calibration with

general knowledge than eyewitness memory questions, $t(52)=—3.93$, $p<.001$, $d=0.68$. Both *C*s were significantly different from 0, $t(52)=11.22$, $p<.001$, $d=2.20$ for general knowledge and $t(52)=9.16$, $p<.001$, $d=1.80$ for eyewitness memory.

The over/underconfidence measure was computed for each participant on each of the six confidence levels and then averaged. It ranged from —1 to +1, showing, respectively, underconfidence and overconfidence. The main statistics are presented in Table 2. The comparison of over/underconfidence for general knowledge and eyewitness memory on each confidence level is pertinent because it allows the statis- tical analysis of what would otherwise be a descriptive anal- ysis of the calibration. Thus, six comparisons, one for each confidence level, were conducted to compare the over/under- confidence with general knowledge and eyewitness memory questions. Bonferroni correction to α was applied for these analyses ($\alpha=.008$). Over/underconfidence was similar for SD =0.10), $t(52)=2.40$, $p=.020$, $d=0.42$. Accuracy rates vary depending on the difficulty of the task, and difficulty affects metamemory (Baranski & Petrusic, 1995; Kebbell, Wagstaff, & Covey, 1996). Indices such as resolution, calibration or over/underconfidence worsen as the difficulty of the task increases and, as a consequence, accuracy decreases (Baranski & Petrusic, 1995). To avoid any explanation of our data in terms of differences in the difficulty of the materials, we removed two very easy general knowledge questions to match the accuracy with both materials. Thus, with 38 general knowledge and 40 eyewitness memory questions, accuracy was similar ($M=0.46$, SD =0.12 and $M=0.43$, SD =0.10, respectively), $t(52)$ =1.06, $p=.29$, $d=0.19$. Confidence, however, was higher with eyewitness memory ($M=53.37$, SD =13.40) than general knowledge ($M=47.09$, SD =13.84), $t(52)=—3.32$, $p=.002$, $d=0.47$.

Two pairwise comparisons with the Student's *t*-test on confidence for correct and incorrect answers showed that confidence for correct answers for general knowledge and eyewitness memory questions was similar, $t(52)=1.42$, $p=.163$, $d=0.19$ (see Table 1).[1] However, confidence was higher for incorrect answers with eyewitness memory than with general knowledge questions, $t(52)$ =8.36, $p<.001$, $d=1.10$.

Correlations

Two correlations were calculated: between-subjects (Pearson) and within-subject (Goodman-Kruskal gamma). Both correlations ranged from +1 to —1, reflecting, respectively, a perfect positive and perfect negative relationship. If correlation equals 0, then the relationship is null. All the correlations are shown in Table 1.

The Pearson's correlation was computed using the accuracy and confidence means for each participant. Both Pear- son correlations were significantly different from zero, $t(51)=6.02$, $p<.001$, 95% confidence interval (CI), [0.45, 0.78] for general knowledge and $t(51)=2.48$, $p=.016$, 95% CI [0.06, 0.55] for eyewitness memory, and higher for the former than the later, z =2.14, $p=.034$. Gamma was computed for each participant and material and then averaged. Gammas were high and significantly different from 0, $t(52)$ =54.67, $p<.001$, 95% CI [0.80, 0.86] for general knowledge and $t(52)=27.32$, $p<.001$, 95% CI [0.58, 0.67] for eyewitness memory, respectively, and higher for general knowledge than eyewitness memory questions, $t(52)=8.21$, $p<.001$, $d=1.43$.

Calibration between confidence and accuracy

In order to be sufficiently informative, calibration must be based on a large set of data points, approximately 200 in each confidence level (Juslin et al., 1996). To increase this number of observations, we reduced the 11 confidence categories to six. Thus, we collapsed confidence levels 10–20, 30–40, 50–60, 70–80 and 90–100. Confidence level 0 was not collapsed with any other level because of the instructions fidence level when they did not know the answer. Both gen- eral knowledge and eyewitness memory calibration curves can be seen in Figure 1. At a descriptive level, calibration with general knowledge questions is closer to the diagonal than with eyewitness memory questions. For both general knowledge and eyewitness memory questions, there is slight underconfidence in the lower levels of confidence. In the higher levels, there is slight overconfidence for general knowledge and substantial overconfidence for eyewitness memory questions.

To quantify calibration and examine the differences be- tween general knowledge and eyewitness memory questions, the Calibration index (*C*; see N. Brewer et al., 2002 for the calculations), and a simple measure of over/underconfidence (confidence minus accuracy) were computed. *C* was used to compare overall general knowledge and eyewitness memory calibration. A *C* of 0 indicates perfect calibration, and higher values indicate worse calibration. The main statistics are pre- sented in Table 1. The results showed better calibration with general knowledge than eyewitness memory questions, $t(52)=—3.93$, $p<.001$, $d=0.68$. Both *C*s were significantly different from 0, $t(52)=11.22$, $p<.001$, $d=2.20$ for general knowledge and $t(52)=9.16$, $p<.001$, $d=1.80$ for eyewitness memory.

The over/underconfidence measure was computed for each participant on each of the six confidence levels and then averaged. It ranged from —1 to +1, showing, respectively, underconfidence and overconfidence. The main statistics are presented in Table 2. The comparison of over/underconfidence for general knowledge and eyewitness memory on each confidence level is pertinent because it allows the statis- tical analysis of what would otherwise be a descriptive anal- ysis of the calibration. Thus, six comparisons, one for each confidence level, were conducted to compare the over/under- confidence with general knowledge and eyewitness memory questions. Bonferroni correction to α was applied for these analyses ($\alpha=.008$). Over/underconfidence was similar for in the confidence levels 0 ($p=.014$), 10–20 ($p=.533$) and 30–40 ($p=.463$). For the remaining collapsed confidence levels, overconfidence was greater for eyewitness memory than for general knowledge questions, $t(52)=—3.71$, $p<.001$, $d=0.66$ for confidence levels 50–60, $t(52) = —3.19$, $p=.002$, $d=0.45$ for confidence levels 70–80, and $t(52)=—6.55$, $p<.001$, $d=0.81$ for confidence levels 90–100. Finally, the Adjusted-Normalized Discrimination Index (ANDI, for the formulae, see Yaniv, Yates, & Smith, 1991) was computed. The ANDI is a measure of discrimination (also called resolution), i.e. the ability of participants to dis- criminate between correct and incorrect answers. ANDI ranges from 0 (null discrimination) to 1 (perfect discrimination) and can be seen in Table 1. Discrimination was better with general knowledge than eyewitness memory questions, $t(52)=10.26$, $p<.001$, $d=1.53$. Both discrimination indices for general knowledge and eyewitness memory questions were significantly different from 0, $t(52)=18.86$, $p<.001$, $d=3.70$ and $t(52)=11.03$, $p<.001$, $d=2.16$, respectively.

DISCUSSION

In this experiment, we examined the relationship between confidence and accuracy with general knowledge and eye- witness memory questions using a cued recall memory test. We

conducted several analyses to study the confidence– accuracy relationship, which provided highly detailed information. There are two main outcomes from this study. First, the confidence–accuracy relationship was analysed with several measures (correlations, confidence means, calibration and discrimination), and all of them showed a better relationship with general knowledge than eyewitness memory questions. Second, with a more ecological memory task, i.e. a cued recall test, and a complex event, the confidence–accuracy relationship with eyewitness memory questions is positive and strong. The original contribution here is that the correspondence between confidence ratings and accuracy, i

Table 2. Over and underconfidence means (standard deviations) for general knowledge and eyewitness memory as a function of the confidence level

	Confidence level					
	0	10–20	30–40	50–60	70–80	90–100
General knowledge	—0.05 (0.07)	—0.07 (0.26)	0.01 (0.33)	—0.03 (0.30)	0.07 (0.26)	0.12 (0.13)
Eyewitness memory	—0.12 (0.17)	—0.10 (0.21)	—0.04 (0.28)	0.14 (0.23)	0.19 (0.27)	0.24 (0.17)

Note: Negative numbers indicate underconfidence and positive numbers indicate overconfidence.

e. calibration, is high in certain levels, whereas not so high in others.

Several studies have shown that the confidence–accuracy relationship is better with general knowledge than eyewit- ness memory questions, but almost all of them have used recognition tests and correlations (Perfect, 2004; Perfect & Hollins, 1996). At first glance, our results also support the view that in general, the confidence–accuracy relationship is better with general knowledge questions. This result is consistent with the hypothesis that the lack of insight into our own knowledge on the topic could impair the confi- dence–accuracy relationship with eyewitness memory ques- tions (Perfect, 2004).

However, we also found that there are no differences in confidence with correct answers.[2] This result suggests that similar factors may affect the evaluations of confidence for correct answers with both general knowledge and eyewitness memory, thus leading to similar confidence ratings when ac- curacy is matched. In fact, there is no theory about meta- memory that predicts different confidence ratings depending on the piece of information, whether semantic or episodic (see Gigerenzer, Hoffrage, & Kleinbölting, 1991; Koriat & Goldsmith, 1996; Nelson & Narens, 1990). At the same time, other factors might increase confidence in incorrect answers with eyewitness memory questions without affecting the incorrect answers with general knowledge questions. For example, the lack of insight into the relative ability of participants with eyewitness memory questions may affect only certain incorrect answers and not all of them. To help rate the likelihood that an answer is correct, participants in a cued recall test may have access to a number of by-products of the retrieval process. For certain incorrect answers, these by-products could be weak or few in number. Because of the lack of insight into the relative ability in the eyewitness memory domain, the weak or few-in-number by-products could be wrongly interpreted as sufficient to guarantee a correct answer, resulting in the incorrect answers being reported with higher confidence than they should. Another factor that might explain why overconfidence is limited to incorrect answers in eyewitness memory questions is the effect of the underlying schema. A schema is a structure of knowledge that represents all of the actions, objects and people typically involved in a given situation, such as going out to dinner, or a bank robbery (Luna & Migueles, 2008). Among other functions, schemata serve to fill the gaps left by uncoded or unretrieved

information, using the features that most typically appear in similar situations. For example, García-Bajos and Migueles (2003) presented a mugging account and found more errors of high- than low- typicality information in both a free recall and a recognition test. Furthermore, in the recognition test, confidence for false alarms was higher for high-typicality than for low-typicality contents. High-typicality errors are likely to be rated with

_____

[2] Note that this result could be the artificial consequence of equated accuracy. The higher confidence for correct answers to general knowledge than eyewitness memory questions, taking into account all questions, could be because the general knowledge questions were easier; however, it has been proposed that the daily exposure to general knowledge questions could im- prove metamemory ratings compared with the scarce experience most people have with eyewitness memory materials (Perfect et al., 1993).

_____

high confidence because of their greater likelihood to be present in most of similar situations. Thus, even when the correct answer is completely unknown, by answering with a high-typicality content, one can have a reasonable chance of being right. However, schemata only affect episodic information, as there is no clear schema for semantic information. Thus, certain incorrect answers that include typical information may have received a boost of confidence because they were part of the schema.[3]

Turning our attention to the eyewitness memory questions and their implication in a forensic context, it is important to acknowledge a limitation in the present research. Unlike most, if not all, real interrogatories, participants here were re- quired to answer, even with just a wild guess, and were asked to assign their wild guesses a confidence rating of 0. This procedure considerably reduces the ecological validity of our re- search. However, if we had allowed participants to withhold answers, probably few of them would have been reported with low confidence. In this vein, Koriat and Goldsmith (1996) found a response criterion (also called Prc) of 0.50 in a recall test, i.e. that participants tend to report answers with confidence 50 or higher, and to withhold answers rated with lower confidence. That would have made the examination of the calibration at the lowest levels of confidence difficult because of the high number of data points needed for the calibration.

The main conclusion with eyewitness memory questions is that confidence seems to be a strong indicator of the accuracy of the answer with a cued recall test. The relatively low *C*, the high ANDI and correlations, and the large difference in confidence ratings for correct and incorrect answers allows us to draw an optimistic scenario about the possibilities of confidence as a predictor of performance in the recall of a complex event. However, a closer look at calibration shows that it was far from perfect and that there is room from improvement. In particular, in the applied setting, the over- confidence at the higher end of the scale is of great importance. This result shows the limitations of confidence as a predictor of accuracy in the higher confidence levels. This is especially relevant because people tend to believe information stated with high confidence, most probably because they assume that high confidence implies high accuracy. Equally important is that 12% of the answers given a 0 confidence rating were correct, despite the zero probability of reporting the correct answer by sheer chance. In other words, participants thought that they were guessing, but actually, some of their memories were correct. This result supports interviewing techniques that encourage eyewitnesses to re-

port everything they can remember, even if they are unsure or consider the information irrelevant (e.g. the cognitive interview, Fisher & Geiselman, 1992).

_____

[3] There are also structures that serve to semantically or conceptually organize related information. For example, in the general knowledge question 'What family would be your enemy if you were a Capulet?', the category 'characters in a Shakespeare play' might be accessed. It seems reasonable that if the correct answer does not come to mind, then one would answer with another character from a Shakespeare play (e.g. Othello), especially when participants must provide an answer as in our procedure. However, it is unlikely that this answer will be rated with high confidence because Othello does not appear in most of Shakespeare plays.

It is also important to highlight the practical implications of the greater informativeness of the calibration. This analysis can help pass information on to other groups, such as judges or police officers, about the relevance of confidence as a marker of accuracy and the credibility attributable to a specific answer. These groups may find it easier to under- stand a sentence that reads '74 percent of the answers rated with confidence 90 or 100 are correct' than 'the correlation between confidence and accuracy is .64'. This last sentence cannot be understood without a good command of statistics; moreover, it does not actually provide information on the likelihood that a given answer is correct or incorrect. An- other implication of the results is that confidence could also help reduce the amount of information that police must take into account when deciding which line of investigation to pursue. There is no point in considering information evaluated with low or even medium confidence, because it is not likely to be correct. Thus, investigators should focus on the information rated with high confidence, because this is the information that ensures a certain level of accuracy.

In summary, little is yet known about the factors that influence the confidence–accuracy relationship with cued recall tests. Determining the variables that affect the confidence–accuracy relationship in an ecological setting is a matter of great importance in eyewitness memory. Once research demonstrates the effect of these as yet unknown variables, we will be in a position to transmit this knowledge to legal system professionals and help prevent costly mistakes from believing a wrong yet confident witness.

ACKNOWLEDGEMENTS

The authors thank Malen Migueles, Teresa Bajo, Phil Higham, Jason Hicks and two anonymous reviewers for their helpful comments on earlier versions of this paper.

REFERENCES

Baranski, J. V. (2007). Fatigue, sleep loss, and confidence in judgment.*Journal of Experimental Psychology: Applied*, *13*, 182–196.

Baranski, J. V., & Petrusic, W. M. (1995). On the calibration of knowledge and perception. *Canadian Journal of Experimental Psychology*, *49*, 397–407. Bornstein, B. H., & Zickafoose, 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.

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence–accuracy relationship in eyewitness identification: The effects of reflection and dis- confirmation on correlation and calibration. *Journal of Experimental Psychology: Applied*, *8*, 44–56.

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory*, *14*, 540–552.

Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language*, *52*, 618–627.

Brewer, N., & Wells, G. L. (2006). The confidence–accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*, 11–30.

Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing techniques for investigative interviewing: The cognitive interview*. Springfield, IL: Charles C. Thomas Publisher Ltd.

García-Bajos, E., & Migueles, M. (2003). False memories for script actions in a mugging account. *European Journal of Cognitive Psychology*, *15*, 195–208.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 311–325.

Herrington, R. (Writer/Director) (2002). *The stick-up [Motion picture]*. United States: Universal Pictures Video.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 22–34.

Hollins, T. S., & Perfect, T. J. (1997). The confidence–accuracy relation in eyewitness event memory: The mixed question type effect. *Legal and Criminological Psychology*, *2*, 205–218.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology. Learning: Memory, and Cognition*, *22*, 1304–1316.

Kebbell, M. R., Wagstaff, G. F., & Covey, J. A. (1996). The influence of item difficulty on the relationship between eyewitness confidence and ac- curacy. *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.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*, 609–639.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory. *Psychological Review*, *103*, 490–517.

Leippe, M. R., & Eisenstadt, D. (2007). Eyewitness confidence and the confidence–accuracy relationship in memory for people. In F. Ross, D. Read& M. P. Toglia (Eds.), *The handbook of eyewitness psychology*(Vol. 2, pp. 377–425). Mahwah, NJ: Erlbaum.

Luna, K., & Migueles, M. (2007). Acciones y detalles en la aceptación de información postsuceso falsa y en la confianza [Actions and details in the acceptance of false information and in confidence]. *Estudios de Psicología*, *28*, 69–81.

Luna, K., & Migueles, M. (2008). Typicality and misinformation: Two sources of distortion. *Psicológica*, *29*, 171–188.

Luna, K., & Migueles, M. (2009). Acceptance of central and peripheral misinformation and confidence. *Spanish Journal of Psychology*, *12*, 405–413.

Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory & Cognition*, *38*, 441–451.

Migueles, M., & García-Bajos, E. (2001). Confianza y exactitud en la memoria de testigos vs. conocimientos generales [Confidence and accuracy in eyewitness memory vs. general knowledge]. *Estudios de Psicología*, *22*, 259–271.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory*, (vol. 26, pp. 125–169). San Diego, CA: Academic Press.

Odinot, G., & Wolters, G. (2006). Repeated recall, retention interval and the accuracy–confidence relation in eyewitness memory. *Applied Cognitive Psychology*, *20*, 973–985.

Olsson, N. (2000). A comparison of correlation, calibration, and diagnosticity as measures of the confidence–accuracy relationship in witness identification. *Journal of Applied Psychology*, *85*, 504–511.

Olsson, N., & Juslin, P. (2002). Calibration of confidence among eyewitnesses and earwitnesses. In P. Chambres, M. Izaute, & P. J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 203–218). Boston, MA: Kluwer.

Perfect, T. J. (2004). The role of self-rated ability in the accuracy of confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, *18*, 157–168.

Perfect, T. J., & Hollins, T. S. (1996). Predictive feeling of knowing judge- ments and postdictive confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, *10*, 371–382.

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.

R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. Retrieved from http://www.R-project.org.

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., & 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.

Robinson, M. D., Johnson, J. T., & Robertson, D. A. (2000). Process versus content in eyewitness metamemory monitoring. *Journal of Experimental Psychology: Applied*, *6*, 207–221.

Schneider, S. L., & Laurion, S. K. (1993). Do we know what we've learned from listening to the news? *Memory & Cognition*, *21*, 198–209.

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence–accuracy calibration in face recognition. *Journal of Applied Psychology*, *88*, 490–499.

Weber, N., & Brewer, N. (2006). Positive versus negative face recognition decisions: Confidence, accuracy and response latency. *Applied Cognitive Psychology*, *20*, 17–31.

Weingardt, K. R., Leonesio, R. J., & Loftus, E. F. (1994). Viewing eye- witness research from a metacognitive perspective. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 157–184). Cambridge, MA: The MIT Press.

Wells, G. L., & Lindsay, R. C. L. (1985). Methodological notes on the ac- curacy–confidence relation in eyewitness identifications. *Journal of Applied Psychology*, *70*, 413–419.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*, 611–617