*The European Journal of Psychology Applied to Legal Context*, 2010, 2(1): 55-71
www.usc.es/sepjf

# NEW ADVANCES IN THE STUDY OF THE CONFIDENCE-ACCURACY RELATIONSHIP IN THE MEMORY FOR EVENTS

Karlos Luna* and Beatriz Martín-Luengo**

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

** Faculty of Psychology, University of the Basque Country (Spain)

(Received: 23 February 2009; revised 10 September 2009; accepted 11 September 2009)

**Abstract**

It has been suggested that calibration, and not correlation, could be a better analysis to examine the confidence-accuracy relationship. Calibration refers to the degree in which confidence ratings correspond with the objective probability that the answer is correct. However, to date calibration has been calculated only in a few studies that have addressed eyewitness identification, and never in the memory for events. For this reason, four experiments were conducted to examine the calibration between confidence and accuracy in the recognition and recall of a criminal situation. The basic procedure involved the presentation of a crime through slides or video, followed by a questionnaire. Confidence ratings were also required. Results showed in general a good confidence-accuracy calibration, with variations depending on the memory test and the variables manipulated. They also showed that participants are slightly overconfident, and that they do not calibrate confidence well in the low-medium levels of difficulty. The main conclusion of this research is that confidence could help to evaluate the accuracy of a testimony under certain circumstances, although generalising the results to real-life situations should be done with caution.

**Keywords:** Eyewitness memory, confidence-accuracy relationship, calibration, metamemory, memory for events.

**Resumen**

En los últimos tiempos se ha sugerido que la correlación podría no ser el mejor medio para examinar la relación entre la confianza y la exactitud, y se ha propuesto que la calibración podría ser más adecuada. La calibración es el grado en que los juicios de confianza de los participantes corresponden con la probabilidad real u objetiva de que su respuesta sea correcta. Se realizaron cuatro experimentos para examinar la calibración entre la confianza y la exactitud en el recuerdo de una situación delictiva. El procedimiento general incluyó la presentación de un delito mediante diapositivas o vídeo seguido de una serie de preguntas en las que los participantes debían responder e indicar su confianza. Los resultados en general apuntan a una buena calibración entre la confianza y la exactitud, con variaciones en función de la prueba de memoria y las variables manipuladas. También se ha encontrado que los participantes responden con una cierta sobreconfianza y que no calibran su confianza adecuadamente en los niveles medio-bajos de dificultad. La conclusión de la investigación es que la confianza puede ayudar a evaluar la exactitud de una declaración en determinadas circunstancias.

**Palabras clave:** Memoria de testigos, relación confianza-exactitud, calibración, metamemoria, memoria de suces

Correspondence: Karlos Luna. Centre for Research in Psychology (CIPsi), School of Psychology, University of Minho. Campus de Gualtar, 4710-057 Braga (Portugal). E-mail: karlos.luna.ortega@gmail.com

ISSN 1889-1861 

## Introduction

An eyewitness with no reason to lie is powerful evidence in a trial. In the absence of evidence to the contrary, an eyewitness recollection is generally accepted as true by police officers, prosecutors, juries and judges (Wells, Memon, & Penrod, 2006). The testimony is especially relevant if the eyewitness looks confident, because people think that the more confident the eyewitness, the more accurate his/her declaration (Krug, 2007; Wells, Lindsay, & Ferguson, 1979). For this reason, the study of the confidence-accuracy (C-A) relationship has become an important research topic in the field of eyewitness memory. The analysis of the variables that mediate the C-A relationship could help to distinguish between accurate and judicially valid declarations from other more dubious ones.

The C-A relationship in eyewitness memory has typically been studied through correlations between the mean of accuracy and the mean of confidence. The results of such analyses regularly display low or null correlations, with either research of eyewitness identification or memory for events (for instance, Brown, 2003; Perfect, 2004; Perfect, Watson, & Wagstaff, 1993; see also the meta-analyses of Bothwell, Deffenbacher, & Brigham, 1987, and Sporer, Penrod, Read, & Cutler, 1995). The logical conclusion from these studies is that the C-A relationship is null and, therefore, confidence is not a good marker of accuracy. Accordingly, this opinion is supported by the majority of experts in eyewitness memory interviewed by Kassin, Tubb, Hosch and Memon (2001).

Recently, another approach to the study of the C-A relationship has been proposed which questions the aforementioned conclusion: the analysis of the confidence calibration. Calibration is defined as the correspondence between the objective and subjective probabilities that an answer is correct (Juslin, Olsson, & Winman, 1996). For instance, with a confidence scale from 0 (minimum) to 10 (maximum), participants are well calibrated when 30 percent of the answers given a confidence rating of 3 are correct, when 70 percent of the answers with a confidence rating of 7 are correct, etc. If the confidence scale is from 1 to 5, the accuracy associated with confidence level 1 will be 10 percent (the mid point between 0 and 20 percent); with confidence level 2 it will be 30 percent (medium point between 20 and 40 percent) and so forth. Hence, on a graph the perfect calibration curve is a diagonal line with accuracy and confidence in the axes (see below Figure 1).

It has been suggested that metamemory judgements are inferences based on clues more or less indirect from our knowledge about the requested information (Nelson, Gerler, & Narens, 1984). Thus, if participants recover the correct clues or make the correct inferences, their confidence judgements could match the actual probability that the answer is correct, and therefore obtain a good calibration. On the contrary, if the metamemory processes are not correct, calibration will be poor.

Research that has applied calibration analysis to eyewitness memory is scarce, and all of the studies have asked participants to identify the perpetrator of a crime. In general, the results of these studies on eyewitness identification suggest that calibration between confidence and accuracy is fairly good (see the meta-analysis with seven experiments by Olsson, 2000, or the more recent papers by Brewer & Wells, 2006, and Weber & Brewer, 2003). These results have encouraged discussion on whether confidence is a good marker of accuracy (for an in-depth review, see Brewer, 2006, and Krug, 2007).

In order to analyse the goodness of fit of an observed calibration to the theoretically perfect, the Calibration Index was developed (Olsson, 2000). If the Calibration Index equals zero, then calibration is perfect. This index has been calculated in several researches. For instance, Brewer, Keast, and Rishworth (2002), asked participants to identify the thief of a credit card in a restaurant and found Calibration Index between .010 and .050. Similarly, with the same materials and task, Brewer and Wells (2006) found indices between .016 and .031.

The objective of the present research is to study the calibration of confidence when participants are asked to remember a realistic criminal event presented in either video or slides. To date, the authors are not aware of any research on calibration analysis conducted with this kind of ecological materials. To achieve the objective four experiments were carried out, in which either a video (Experiments 1, 3 and 4) or slide show (Experiment 2) was presented. Then participants answered several questions about the event. The answer format changed depending on the experiment: recognition with two or four alternatives in Experiments 1 and 2, respectively, and cued recall in Experiments 3 and 4.

## EXPERIMENT I

### Method

#### Participants

55 voluntary students from the Faculty of Psychology and the Faculty of Computer Sciences of the University of the Basque Country (Spain) took part in this experiment. 18 students were withdrawn from the sample because they did not use all of the different confidence levels with their answers. Thus, data from 37 participants were used (29 females, *M* age= 21.05 years; *SD*= 3.24).

#### Materials and procedure

The experiment consisted of two sessions on consecutive days. The first day participants saw a 3-minute excerpt from the film *The Stick-Up* (Herrington, 2002). The video depicted two security guards getting off an armoured vehicle and depositing several sacks of money in a bank's safe deposit room. When they drive away, a robber cuts off the power supply to the building and walks into the bank in disguise carrying a sawed-off shotgun. After threatening customers and tellers, he takes the money and drives away. The incident unfolds with no explicit violence. After watching the video, the participants completed several tasks unrelated and were dismissed.

The following day the participants completed a True/False recognition memory test consisting of 24 statements (e.g., The robber had a gun holstered to his belt). The participants' task was to indicate whether the information in the statement had appeared in the video or not. They were also asked to rate on a scale of 1 (no certainty) to 5 (absolute certainty) how confident they felt that the selected answer was correct. Finally, participants were thoroughly debriefed and dismissed.

### Results and discussion

First of all, the accuracy in the different levels of confidence was computed, i.e., the accuracy of the answers assigned a confidence rating of 1, 2 and so on. All the scores and significances are shown in Table 1.

**Table 1**. Accuracy as a function of confidence level in Experiments 1 and 2.

| | Confidence 1 | | Confidence 2 | | Confidence 3 | | Confidence 4 | | Confidence 5 |
|---|---|---|---|---|---|---|---|---|---|
| Exp. 1 | .51 | = | .44 | = | .51 | < | .67 | < | .82 |
| Exp. 2 | .43 | = | .36 | = | .41 | = | .39 | = | .43 |

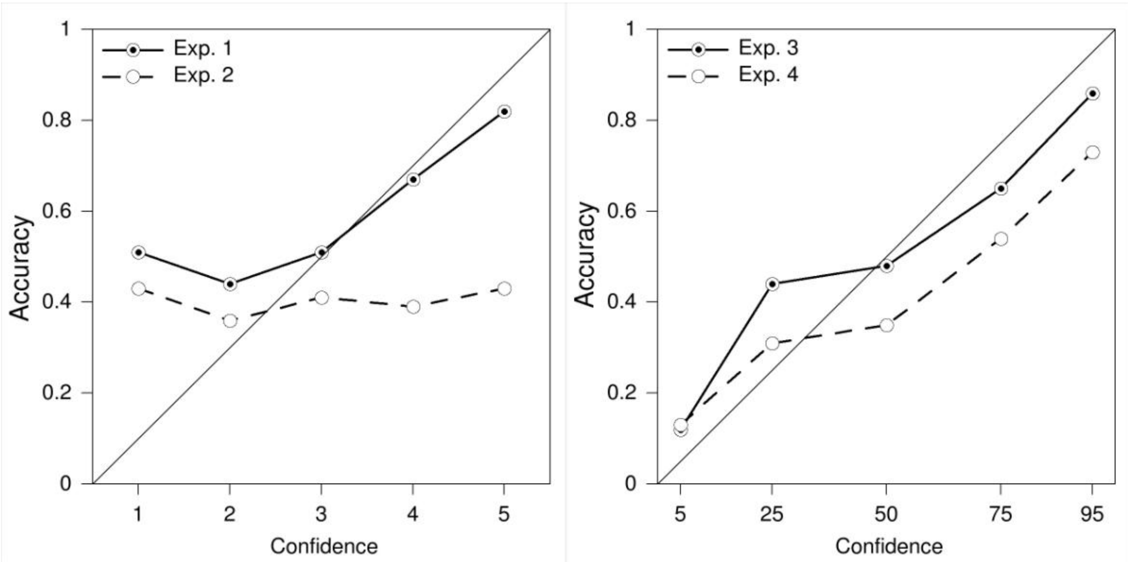


**Figure 1.** Confidence-accuracy calibration in Experiments 1 and 2 (left panel) and Experiments 3 and 4 (right panel). Diagonal line depicts perfect calibration.

To study the calibration of accuracy as a function of confidence a within-subjects Analysis of Variance (ANOVA) 5 (Confidence: 1, 2, 3, 4, and 5) with accuracy as a dependent measure was conducted. Accuracy was significantly different depending on the confidence level, $F(4, 144)= 15.06$; $p< .001$; $\eta^2= .29$. The differences were analysed with pairwise comparison using the Student's t-test. The accuracy of answers given a confidence rating of 5 was higher than the rest (all $p< .001$), and the accuracy of answers rated with confidence level 4 was higher than with confidence level 3, 2 or 1 (all $p < .003$). No other differences were significant. As shown in Figure 1 (left panel), participants calibrate relatively well in the upper part of the calibration curve.

This result can be explained because in the recognition memory test with two alternatives the answer is correct just by chance half of the times. When participants do not know the answer and randomly choose one of the alternatives, theoretically they should rate their answer with confidence level 1. Any random answer will be correct 50 percent of the time, although according to the calibration curve the expected result will have a very low accuracy rate. The same thing will happen when a participant has (or thinks s/he has) some knowledge about the subject and rates the answer with confidence 2 or 3. The accuracy theoretically associated with these confidence levels is 30 and 50 percent, but, again, the participants answered 50 percent correctly by chance. When the theoretical accuracy exceeds the accuracy by chance, i.e., with a confidence rating of 4, actual accuracy significantly increases to 64 percent. Accordingly, when participants are completely sure about their answer (confidence level 5; theoretical accuracy 90 percent), accuracy also increases significantly (82 percent).

In order to know how well the actual curve fits with the theoretically perfect, the calibration index developed by Olsson (2000) was computed. The index was different from 0 ($M$= .096, $SD$= .06; $t(54) = 12.05$; $p < .001$) meaning calibration was not perfect. We will return to this result in the General Discussion to comment it in relation with the indices from Experiments 2, 3, and 4.

In summary, when the theoretical calibration curve predicts a below chance performance, observed calibration is very poor because this performance cannot be obtained. However, when performance is above chance calibration seems to be quite good. This result led us to design another experiment using a different memory test with more alternative answers. With more alternatives the number of correct answers by chance will be lower, and the study of the calibration of the confidence in the lower levels of accuracy will be possible.

## EXPERIMENT II

The results of Experiment 1 did not allow us to examine the calibration in the lower levels of accuracy. As a consequence, another experiment was conducted using a memory test with more alternative answers. In this case, a source monitoring memory test (Lindsay & Johnson, 1989) was selected. In this test participants have to report the origin of their memories in a format that usually has four alternatives (e.g. Hornstein & Mulligan, 2004; Zaragoza & Lane, 1994). With this memory test there is a 25 percent probability that an answer is correct by chance. Another reason to use a source monitoring memory test was to explore calibration between confidence and accuracy when participants were asked to focus on information on the origin of the memories.

## Method

### Participants

60 students from the Faculty of Psychology of the University of the Basque Country took part in this experiment as a course requirement. Three participants were withdrawn because they did not come to the second session and 17 because they did not use all of the confidence levels. Thus, the analyses were conducted with 40 participants (34 female; age $M$ = 20.18 years; $SD$ = 1.32). None of them participated in the previous experiment.

### Materials and procedure

Two major changes were introduced compared to Experiment 1. The first was the way in which the event was presented, switching from video to slides. This change allowed us to study variables not included in this paper. Preliminary analyses showed that the variables did not affect calibration, and therefore the data were collapsed across conditions. The second change was the introduction of a questionnaire between the presentation of the event and the memory test. The purpose of this change was to introduce a new source of information to create the necessary conditions for the source monitoring memory test.

The experiment included two sessions. In the first one participants watched a slide show consisting of 22 slides about a mobile phone stolen in an outdoor area on the university campus. After the slides, participants completed a filler task and immediately afterwards answered the questionnaire containing 32 open questions about the event (e.g., When the robber approaches the girl, what did he ask for?). Eight questions included false information and another 8, true information previously presented in the slides. In addition to answering the questions, the participants had to rate their confidence on a scale of 1 (no certainty) to 5 (absolute certainty).

Two days later the participants completed the source monitoring memory test in which they had to select the origin of some pieces of information. The memory test included 32 statements (e.g. The thief was wearing a dark blue cagoule), 8 presented only in the slides, 8 presented only in the questionnaire, 8 presented in both slides and questionnaire, and 8 in neither. Accordingly, the answer format included four options: Slides, Questions, Both, and Neither. A confidence rating was also required. Finally, the objectives of the experiment were briefly mentioned and questions about it were answered.

### Results and discussion

As in Experiment 1, accuracy was computed separately for each confidence level. The data were analysed with a within-subjects ANOVA 5 (Confidence: 1, 2, 3, 4, and 5) with accuracy as dependent measure. The scores can be seen in Table 1.

There was no difference in accuracy as a function of confidence, $F(4, 156) = 0.37$; *ns*. Accuracy was very similar in the five confidence levels, as shown in Figure 1 (left panel). Congruently, the calibration index was high and significantly different from 0 ($M = 0.167$, $SD = 0.09$; $t(56) = 14.80$; $p < .001$). We will turn our attention to this result later on.

The main conclusion of the experiment is that with a source monitoring memory test participants do not correctly calibrate confidence. Although the rationale for this experiment was to examine calibration when accuracy was low, the results showed that calibration was null with both high and low accuracy. A possible explanation for the poor calibration is that in a source monitoring memory test participants have to recover the origin of their memories. If confidence is an inference extracted from clues as, for instance, the fluency or easiness of the retrieval of the information (Brewer, Sampaio, & Barlow, 2005; Perfect, Hollins, & Hunt, 2000), the active search of information about the origin of the memory provoked by the source monitoring test (Johnson, Hashtroudi, & Lindsay, 1993) could interfere or even block the recovery and analysis of these clues. This interference seems to happen in spite of the fact that both source monitoring and confidence tasks are performed sequentially and not simultaneously. As a result, the source monitoring is well performed (i.e. with the habitual proportion of errors), but the calibration between confidence and accuracy is impaired.

## EXPERIMENT III

In Experiment 2 it was not possible to explore calibration in the lower levels of accuracy. Thus, a third experiment was conducted with a different memory test. Instead of a recognition, a cued recall memory test was used in which participants were directly asked about several facts surrounding the event and with no alternative answers. The probability of guessing correctly when the answer is not known is almost zero with this test. Moreover, some results suggest that C-A correlation is high with recall test (Odinot & Wolters, 2006; Robinson & Johnson, 1996). Therefore, it was not expected a null calibration as in Experiment 2.

## Method

### Participants

22 participants took part in this experiment. Five did not use all of the confidence levels with their answers and were withdrawn, so that data from 17 participants were included in the analyses (14 female, age $M$= 27.24 years; $SD$= 8.19). Participants included students from the Faculty of Philosophy and Education Sciences of the University of the Basque Country, who participated as a requirement for a course, and voluntary researchers from the Basque Institute of Criminology. None of them took part in previous experiments.

### Materials and procedure

Participants completed the experiment in small groups and in a single session. The same video as in Experiment 1 was used. After watching the video about the bank robbery, participants completed a filler task, immediately followed by a booklet with 40 cued recall questions about the event (e.g. When the robber was in the car, what did he have in his hand?). All the questions had a brief correct answer that could be written in few words. A confidence rating was also required on a scale from 0 to 100 percent. The scale was changed to a percentage so that participants would be more familiar with it. If they did not know the answer, they were asked to guess and write whatever they wanted, rating their guess with a confidence level of 0. Finally, the objectives were discussed and questions about the experiment were answered. The session lasted approximately one hour.

## Results and discussion

Firstly, the confidence data was recoded to reduce the 11 levels to a more adequate number for the calibration analysis. In this recodification the balance of the intervals took priority over the equidistance between their medium points, which served as markers of the level. Then, the confidence levels 0-10 were collapsed (medium point 5 percent), 20-30 (25 percent), 40-50-60 (50 percent), 70-80 (75 percent) y 90-100 (95 percent).

In order to determine if accuracy changed as a function of confidence, a 5 (Confidence: 5, 25, 50, 75 and 95) within-subjects ANOVA was conducted. Accuracy was significantly different as a function of confidence, $F(4, 64)= 30.66$; $p< .001$; $\eta^2= .68$. Scores and differences are shown in Table 2, and calibration curve in Figure 1 (right panel).

To further examine differences in accuracy, several pairwise comparisons with the Student's t-test were conducted. All the differences were significant except for confidence levels 25 and 50, $t(17)= -.70$; *ns*. This result suggests that participants are unable to calibrate confidence correctly with medium level accuracy, and that at these levels they rated confidence in a wide interval between 20 and 60 percent. When confidence is 50 percent calibration is very good because accuracy is very close to .50, but with this same accuracy answers could be also rated with a much lower confidence. In line with the good calibration, the calibration index was low, although still different from zero ($M= .043$, $SD= .02$; $t(21)= 9.75$; $p< .001$).

**Table 2**. Accuracy as a function of confidence level in Experiments 3 and 4.

| | Conf. 5 | | Conf. 25 | | Conf. 50 | | Conf. 75 | | Conf. 95 |
|---|---|---|---|---|---|---|---|---|---|
| Exp. 3 | .12 | < | .44 | = | .48 | < | .65 | < | .86 |
| Exp. 4 (Overall) | .14 | < | .31 | = | .35 | < | .53 | < | .74 |
| Encoding | .11 | < | .31 | = | .38 | < | .58 | < | .74 |
| Metamemory | .16 | < | .31 | = | .32 | < | .50 | < | .72 |

In summary, in this experiment it was possible to examine calibration in the lower levels of accuracy, and the results point towards a very good calibration with both high and low accuracy. However, calibration was not good in all the confidence levels.

Moreover, participants showed slight overconfidence, i.e. higher observed than theoretical confidence, with easy questions (high accuracy) and a slight underconfidence, i.e. lower observed than theoretical confidence, with hard questions (low accuracy). In order to diminish or eliminate these calibration errors, the next experiment was conducted.

## EXPERIMENT IV

As a consequence of the results from Experiment 3, a new experiment was planned in order to improve calibration in the levels of accuracy where the difference between the observed and perfect calibration was higher. Some research has found that calibration can improve if certain instructions are used. For instance, Brewer et al. (2002) asked participants to identify a credit card thief in a line-up and asked for confidence ratings. One group was instructed to think about the conditions in which information was encoded, and another was asked to think about possible reasons why their identification could be wrong. The authors found better C-A correlation in both groups compared with a control without instructions.

In this experiment the instructions were also manipulated. One group was instructed to focus on the information encoding conditions. If participants are aware of the conditions, they might realize that sometimes they are not the best possible for encoding the information. As a result, confidence could decrease. The opposite could be true in conditions in which the encoding of information was likely to be good. The second group was asked to focus on their own metamemory, i.e. on the process of evaluating confidence. This way, it was expected a change in the pattern of underconfidence in the lower levels of accuracy and overconfidence in the higher levels.

### Method

#### Participants

54 volunteer students from the Faculty of Psychology of the University of the Basque Country took part in this experiment, of which 14 did not use all the confidence

levels with their answers. Thus, the data from 40 participants were used (32 female; age *M*= 19.46 years; *SD*= 3.49). None of them took part in previous experiments.

**Materials and procedure**

The materials and procedure were the same as in Experiment 3. The participants watched a video about a bank robbery, completed a filler task and answered 40 cued recall questions. Confidence was also measured on a 0 to 100 percent scale. The participants were randomly assigned to one of two conditions. The group instructed to focus on the encoding conditions included 19 participants, and the group with metamemory instructions included 21 participants. The instructions for the encoding condition were as follows:

> When answering, please bear in mind the visibility of characters, objects and actions in the video and the amount of times they appeared. Also consider your own attention when watching the video, whether you lost concentration or looked anyplace other than the screen at any time.

The instructions for the metamemory condition were as follows:

> When rating confidence, please bear in mind that when faced with a question we consider easy we tend to rate the answer with higher confidence than we should. Conversely, faced with a hard question we tend to rate the answer with lower confidence than we should.

## Results and discussion

Just like Experiment 3, confidence ratings were collapsed in 5 levels. To examine accuracy as a function of confidence and instructions, a mixed factor ANOVA 5 (Confidence: 5, 25, 50, 75 and 95) x 2 (Instructions: encoding and metamemory) was conducted with confidence as a within-subject and instructions as between-subject factor.

The analysis showed significant differences for accuracy as a function of confidence, $F(4, 152)= 56.92$; $p< .001$; $\eta^2= .60$. The pattern of results matched that of

Experiment 3: all the differences were significant (all $p < .001$), except between confidence levels 25 and 50, $t(44)= -.92$; *ns*. However, there were no differences as a function of the instructions, $F(1,38)= .31$; *ns*. All the scores and differences can be seen in the Table 2. Furthermore, participants' accuracy did not vary with the instructions. The performance was very similar in both encoding ($M= .41$) and metamemory ($M= .44$) conditions. This result was expected because the manipulation of the instructions was not intended to affect accuracy, but rather the ability to correctly calibrate confidence. The interaction between confidence and instructions was also non significant.

The objective of this experiment was to improve calibration by means of instructions. However, the result was the opposite. The calibration indices for the whole sample ($M= .072$, $SD= .05$), or for either participants with the encoding ($M= .062$, $SD= .05$) or metamemory instructions ($M= .081$, $SD= .06$) were worse than the index from Experiment 3 (all $p> .05$) and significantly different from 0 (all $p> .001$). Not only did the instructions used not improve calibration, but actually impaired it, as shown in Figure 1 (right panel).

There are several reasons to explain why calibration was worse when instructions were included. One such reason could be that confidence ratings are selected by means of an automatic process, and that when participants tried to control it the only consequence is that they performed worse. Similarly, when an ability has been automated, for instance riding a bicycle, when trying to consciously control the movements involved some difficulties can be experienced. However, the results reported by Brewer et al. (2002) suggest that calibration can conceivably improve with certain instructions. The main difference between the Brewer et al. (2002) experiment and our Experiment 4 is the task, eyewitness identification in the first and recall of a complex event in the second. There are no theoretical reasons to think that the metamemory processes are different depending on the task, but this is something that may deserve future consideration and research. Another reason that might explain the poor calibration here is that the instructions may not have been the right ones, or that participants did not follow them. However, in these cases calibration should be similar to that of Experiment 3 without instructions, and this was not the case.

## General Discussion

In this series of experiments the calibration between confidence and accuracy in the recall and recognition of a complex criminal event was examined. The main conclusion is that the C-A relationship may not be as bad as inferred from studies using correlations. Our results, in line with eyewitness identification research (Brewer & Wells, 2006; Olsson, Juslin, & Winman, 1998), suggest some kind of relationship between confidence and accuracy, but that this relationship cannot be adequately captured by correlations (Juslin et al., 1996). Furthermore, correlations give limited information from a practical point of view. The knowledge that the C-A correlation is, for instance, about .15 does not help to evaluate the accuracy of a specific piece of information rated with 80 percent. However, the knowledge that such a piece of information has nearly an 80 percent probability of being correct can be more helpful and informative in a trial (Weber & Brewer, 2003).

At this point it is important to mention that confidence showed during a trial is not a useful measure to evaluate the accuracy of the information provided. Before a trial eyewitnesses are interrogated several times and asked repeatedly to remember the same event. As a consequence, confidence tends to increase (see the imagination inflation phenomenon, Garry, Manning, Loftus, & Sherman, 1996). However, it has been suggested that confidence ratings may be a useful marker of accuracy when measured at the very first interrogation and when the interval from the event is minimum (Brewer, 2008, referred to face identification). For this reason, the results from this and similar laboratory research have to be interpreted with care, due to the significant differences between experimental and real situations.

The good calibration curves also point to a different conclusion from that of experts in eyewitness memory, who claim that confidence is not a good marker of accuracy (Kassin et al., 2001). Even though in Experiment 4 calibration worsened with instructions, calibration curves were very close to the perfect calibration. However, it is worth noting that in Experiment 2 calibration was null. This result should be examined in the future in order to study the potential causes.

To obtain a measure of the differences between the observed and the perfect calibration the Olsson (2000) calibration index was calculated. This index varied between .167 (Exp. 2) and .043 (Exp. 3). However, in Experiment 2 calibration was almost null because accuracy was very similar in all the confidence levels. The calibration index, as formulated, may not prove to be a sensible measure. Further research regarding the limits of the calibration index, sensibility and adequateness will be necessary.

In Spain today, eyewitness memory experts are not commonly asked their opinion in trials about the validity of a testimony. However, a few years ago the jury system was introduced in the judicial system, in which several lay persons have to deliver a verdict of guilty or not guilty. It is very important to help the members of the jury to determine the extent to which they can believe a testimony, or to show them which variables can affect the credibility of an eyewitness. By studying memory and the relationship between confidence and accuracy, experts could help juries to determine when an eyewitness can be believed or not. Nevertheless, psychologists are still far from achieving this objective. For this reason, it is very important to develop lines of research in these subjects to improve both theoretical and empirical knowledge. This way, it will be finally possible to pass this knowledge on to the judicial system

## Acknowledgements

We wish to thank Dr. Marlen Migueles for her collaboration in some of the experiments.

## References

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology*, *72*, 691-695.

Brewer, N. (2006). Uses and abuses of eyewitness identification confidence. *Legal and Criminological Psychology*, *11*, 3-23.

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.

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

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.

Brown, J. M. (2003). Eyewitness memory for arousing events: Putting things into context. *Applied Cognitive Psychology*, *17*, 93-106.

Garry, M., Manning, C., Loftus, E. y Sherman, S. (1996). Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin & Review, 3,* 208-214.

Herrington, R. (Director) (2002). *The Stick-Up*. USA.: Universal Pictures Video.

Hornstein, S. L., & Mulligan, N. W. (2004). Memory for actions: Enactment and source memory. *Psychonomic Bulletin & Review*, *11*, 367-372.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3-28.

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.

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research. *American Psychologist*, *56*, 405-416.

Krug, K. (2007). The relationship between confidence and accuracy: Current thoughts of the literature and a new area of research. *Applied Psychology in Criminal Justice, 3*, 7-41.

Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition*, *17,* 349-358.

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General*, *113*, 282-300.

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., & Winman, A. (1998). Realism of confidence in earwitness versus eyewitness identification. *Journal of Experimental Psychology: Applied*, *4*, 101-118.

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., & Hunt, A. L. (2000). Practice and feedback effects on the confidence-accuracy relation in eyewitness memory. *Memory*, *8*, 235-244.

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.

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence-accuracy correlation. *Journal of Applied Psychology*, *81*, 587-594.

Sporer, S. L., Penrod, S., Read, J. D., & Cutler, B. L. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*, 315-327.

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.

Wells, G. L., Lindsay, R. C., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology*, *64*, 440-448.

Wells, G. L., Memon, A., & Penrod, S. (2006). Eyewitness evidence. Improving its probative value. *Psychological Science in the Public Interest*, *18*, 427-443.

Zaragoza, M. S., & Lane, S. M. (1994). Source misattributions and the suggestibility of eyewitness memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 934-945.