Journal of Experimental Psychology: Applied
2004, Vol. 10, No. 3, 156–172

1076-898X/04/$12.00 DOI: 10.1037/1076-898X.10.3.156

Confidence–Accuracy Calibration in Absolute and Relative Face Recognition Judgments

Nathan Weber and Neil Brewer
Flinders University

Confidence–accuracy (CA) calibration was examined for absolute and relative face recognition judgments as well as for recognition judgments from groups of stimuli presented simultaneously or sequentially (i.e., simultaneous or sequential mini-lineups). When the effect of difficulty was controlled, absolute and relative judgments produced negligibly different CA calibration, whereas no significant difference was observed for simultaneous and sequential mini-lineups. Further, the effect of difficulty on CA calibration was equivalent across judgment and mini-lineup types. It is interesting to note that positive (i.e., old) recognition judgments demonstrated strong CA calibration whereas negative (i.e., new) judgments evidenced little or no CA association. Implications for eyewitness identification are discussed.

The problem of mistaken eyewitness identifications has long been known to forensic psychologists but has been emphasized in recent years by the cases that highlight mistaken eyewitness identifications as a major cause of the conviction of innocent people (Wells et al., 1998). A large body of research in the psychology–law domain has addressed this issue by searching for potential independent markers that will help shed light on whether an identification was accurate. The majority of this research has focused on eyewitness identification confidence as just such a marker. Historically, the data on the confidence–accuracy (CA) relationship in eyewitness identification have not been encouraging. A large body of research has investigated the CA relationship in eyewitness identification using the point-biserial correlation as the index of this relationship. Various reviews and meta-analyses have concluded that confidence and accuracy are, at best, only weakly correlated (Bothwell, Deffenbacher, & Brigham, 1987; Sporer, Penrod, Read, & Cutler, 1995; Wells & Murray, 1984). The average coefficients identified in these reviews range from .07 to .28. Not surprisingly, on the basis of this evidence the conclusion of the field has been that confidence is not a useful indicator of accuracy in the eyewitness identification context, a conclusion that has been offered by many expert witnesses in the courtroom.

However, a growing body of recent research, which we discuss in detail a little later in this section, has contradicted this conclusion and highlighted the importance of further research on the CA relationship in eyewitness identification. A compelling stream of this research has focused on the use of calibration, instead of the point-biserial correlation, as an index of the CA relationship. Two major questions regarding the utility of confidence as a marker of identification accuracy are of immediate applied interest. How does CA calibration, and hence the utility of confidence as an indicator of identification accuracy, differ between simultaneous and sequential lineups? Similarly, how does CA calibration differ between choosers (i.e., witnesses who made a positive identification from target-present or target-absent lineups) and nonchoosers (i.e., witnesses who rejected target-present or target-absent lineups)? These questions are important not only in outlining the situations in which confidence is a reliable indicator of accuracy, but they are also relevant to the development of an identification procedure that maximizes conviction of guilty suspects while minimizing the conviction of innocent suspects.

We believe, for three main reasons, that one of the best approaches to understanding CA calibration in eyewitness identifications is to start, not only with investigation of CA calibration in identification tasks, but also with a comparison of the basic face recognition processes underlying these identification tasks. Specifically, this involves investigation of calibration in the most basic face recognition judgment tasks thought to underlie decisions from simultaneous (i.e., relative judgments) and sequential (i.e., absolute judgments) lineups. One advantage of this approach, previously discussed by Weber and Brewer (2003), relates to the fact that calibration analysis requires a large number of observations per condition. In an eyewitness identification paradigm in which each participant typically makes only a single identification decision, an extraordinarily large sample is needed to assess calibration adequately. For example, in excess of 900 participants completed Brewer, Keast, and Rishworth's (2002) experiment that compared calibration in three experimental conditions, and confidence categories still had to be collapsed to achieve a stable calibration curve. In contrast, a face recognition paradigm allows each participant to make many decisions and, thus, needs only a relatively modest sample. For example, Weber and Brewer used a face recognition task to examine CA calibration in four conditions and required only 48 participants. A second advantage of the face recognition approach is that this paradigm allows a large number of stimuli to be used. Thus, through random grouping of stimuli and counterbalanced, or randomized, assignment of these stimuli groups to experimental conditions, we can negate the effects of the nature of individual stimuli. In contrast, the cost of eyewitness identification stimulus development and the time required for a participant to

Nathan Weber and Neil Brewer, School of Psychology, Flinders University.

This research was supported by Grant A00104516 from the Australian Research Council.

Correspondence concerning this article should be addressed to Neil Brewer, School of Psychology, Flinders University, GPO Box 2100, Adelaide, South Australia 5001, Australia. E-mail: neil.brewer@flinders.edu.au

view each stimulus presentation (either live or a video) and make a subsequent identification decision severely limit the number of different stimuli that it is practical to use in any one experiment. Thus, without repeated replication with varying stimuli, the effect of the nature of the stimuli cannot be ruled out when using an eyewitness identification paradigm. Finally, by initially studying CA calibration in the most basic tasks and gradually increasing the complexity of the task in successively closer approximations to a lineup decision, understanding can be advanced in two ways. First, the nature of the processes underlying identification judgments can be more thoroughly explored. Second, the impact of different aspects of the identification task on performance and processing can be described, thus potentially exposing modifications to lineup procedures that may encourage use of the most adaptive cognitive processing mechanisms and hence improve performance.

Returning then to the issue of the CA relation, challenges to the conclusion that confidence and accuracy in eyewitness identifications are, at best, weakly related are evident in three separate streams of research. Sporer et al. (1995) meta-analyzed 30 identification experiments and found overall results consistent with those described above ($r = .28$). However, in addition to investigating the overall relationship, Sporer et al. also explored the confidence–accuracy correlation for choosers and nonchoosers separately. They found a stronger association between confidence and accuracy for choosers ($r = .37$) than for nonchoosers ($r = .12$) and concluded that, when investigation is confined only to witnesses who make positive identifications, confidence is a stronger predictor of accuracy than originally thought.

Much stronger evidence for the efficacy of confidence as a predictor of accuracy has been reported by D. S. Lindsay, Read, and colleagues (D. S. Lindsay, Nilsen, & Read, 2000; D. S. Lindsay, Read, & Sharma, 1998; Read, D. S. Lindsay, & Nicholls, 1998). They argued that in typical eyewitness identification experiments all participants are (a) exposed to very similar stimuli under the same viewing conditions and (b) make their identification decisions from the same lineup in identical testing conditions. Such conditions constrain the variability in confidence judgments, leading to the low confidence–accuracy correlations typically observed. Data from three experiments supported this argument with CA correlations (e.g., $r = .72$ and $r = .69$; Read et al., 1998) observed under encoding and retrieval conditions producing variable performance but more typical correlations (e.g., $r = .18$ and $r = .26$) when performance was more homogeneous.

The most compelling evidence for the CA relationship, however, comes from studies using calibration, rather than the point-biserial correlation, as the index of the CA association. Calibration refers to the association between the objective (accuracy) and subjective (confidence) probabilities of the occurrence of an event. Calibration is typically assessed in four ways: calibration curves, the calibration statistic (C), the over/underconfidence statistic (O/U), and resolution. Calibration curves are created by plotting the proportion of accurate decisions for each level (or range, for a continuous scale) of confidence judgments against the mean confidence for that level or range. Perfect calibration occurs when all of the decisions made with 100% confidence are correct, 90% of the decisions made at 90% confidence are correct, and so on. In other words, the calibration curve is a straight line with slope one and *y*-intercept zero. The calibration statistic is a measure of deviation from perfect calibration ranging from 0 (*perfect calibration*) to 1 (*worst possible calibration*). It is computed as the weighted mean of the squared difference between confidence and proportion correct for each confidence level. Over/underconfidence is a gross measure of a participant's tendency to respond, on average, with more or less confidence than the accuracy of their decisions warrants. It ranges from –1 (*complete underconfidence*) to +1 (*complete overconfidence*) and is calculated as the difference between mean confidence and mean accuracy. Finally, resolution is an index of the extent to which confidence judgments discriminate correct from incorrect decisions. The normalized resolution index ranges from 0 (*no resolution*) to 1 (*perfect discrimination*). For a detailed discussion and formal development of these indices see, for example, Baranski and Petrusic (1994) or Yaniv, Yates, and Smith (1991).

In a number of articles Juslin, Olsson, and colleagues have demonstrated that confidence and accuracy can be well calibrated in an eyewitness identification context (Juslin, Olsson, & Winman, 1996; Olsson, 2000; Olsson & Juslin, 1999; Olsson, Juslin, & Winman, 1998). The utility of confidence was further supported by Brewer et al. (2002) who demonstrated that CA calibration could be improved by experimental manipulations. Importantly, all of the existing evidence of good CA calibration in the eyewitness identification context comes from studies that use simultaneous lineups. A growing body of research, though, suggests that superior overall identification accuracy is produced by the sequential lineup procedure (e.g., R. C. L. Lindsay & Wells, 1985; Steblay, Dysart, Fulero, & R. C. L. Lindsay, 2001) and is, thus, being used as the basis for the adoption of sequential, rather than simultaneous, lineups by the police. In the standard simultaneous lineup procedure, witnesses are exposed to all of the lineup members at once and are asked to identify the offender from this array or to reject the lineup if the offender is not present. In contrast, the sequential procedure has witnesses view the lineup members one at a time in sequence. Importantly, the witness must make a decision, which cannot be revisited, about each lineup member before seeing the next. Once the witness has identified a lineup member, the procedure is ended. The lineup is rejected when the witness responds negatively to all of the lineup members. Thus, these procedures differ in the extent to which witnesses can use relative judgment strategies. Specifically, the simultaneous procedure more readily allows comparison of the lineup members and, therefore, relative judgment processes to be used.

An important issue in the understanding of CA calibration in eyewitness identification is the extent to which the simultaneous and sequential lineups differ. This problem is important for two reasons. First, a small body of evidence is establishing the utility of confidence as a marker of identification accuracy for simultaneous lineups, but no data on CA calibration in sequential lineups are currently available. Thus, examination of CA calibration for both of these types of decision is important in establishing the utility of confidence as a marker of identification accuracy. Additionally, such an investigation could play an important role in the formation of recommendations about the superior lineup procedure, as a procedure that allows reliable diagnosis of identification errors would have obvious practical advantages. Although sequential lineups may produce fewer errors than the simultaneous procedure, if these errors are only distinguishable in simultaneous lineups, then the proportion of convictions based on erroneous identifications can be reduced after the fact for simultaneous but

not sequential lineups. If simultaneous lineup errors are sufficiently distinguishable, this could reduce, or even eliminate, the benefits of using sequential lineups. Thus, an understanding of CA calibration, in addition to data on the accuracy of lineup identifications, for simultaneous and sequential lineups could be pivotal in determining the most appropriate police practices.

To examine the CA calibration difference between absolute and relative judgments, we require more than a simple comparison. Possibly, the most robust finding in the calibration literature is the effect of difficulty on calibration (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, Winman, & Olsson, 2000). This effect, known as the *hard–easy effect*, produces more overconfidence as the difficulty of the task increases and more underconfidence with decreasing difficulty. Whether the hard–easy effect is the result of the nature of the judgment process (e.g., Gigerenzer et al., 1991) or artifacts such as scale end effects, regression to the mean, and the linear dependence between difficulty and O/U (Juslin et al., 2000), it is obvious that task difficulty must be considered when comparing calibration across conditions. Comparison of face recognition performance suggests that absolute and relative judgments do indeed differ in task difficulty (Deffenbacher, Leu, & Brown, 1981; Weber & Brewer, 2003). Therefore, to develop a comprehensive understanding of the difference in CA calibration between absolute and relative judgments, we must answer two questions: Does CA calibration differ when difficulty is equated across the two judgment types? Does difficulty affect absolute and relative judgments equivalently?

Another issue regarding CA calibration of considerable applied importance is the difference between choosers and nonchoosers. Attention is often focused on the value of understanding positive identifications because of the obvious import of both correct identifications of guilty suspects and, also, the incorrect identification of innocent suspects. However, as demonstrated by Wells and Olsson (2002), lineup rejections can provide valuable information regarding the likely innocence of a suspect. In other words, a marker of the accuracy of negative decisions allows the discrimination of correct rejections, which should lead to the release of innocent suspects, from incorrect rejections which, undetected, could lead to the erroneous release of a guilty suspect. In other words, both categories of identification response are important.

Of course, the equal importance of lineup rejections and positive identifications does not necessitate or imply equivalent CA relationships for the two types of decisions. Evidence from a number of studies using point-biserial correlation (Sporer et al., 1995) reported higher CA correlations for choosers (i.e., witnesses making positive identifications) than for nonchoosers (i.e., witnesses who rejected the lineup). Only one study has compared CA calibration for choosers and nonchoosers, and this comparison (Brewer et al., 2002), as a result of a small proportion of target-absent lineups, was based on an extrapolation from the observed data. The nonchoosers in Brewer et al.'s (2002) control condition displayed a weak, at best, CA association.

The presence of a strong CA association for positive, but not negative, decisions has obvious implications for confidence as an indicator of accuracy in the eyewitness identification domain. Specifically, these findings would suggest that confidence could be used as an indicator of accuracy for positive identifications but not for lineup rejections. Importantly, comparison of calibration for positive and negative decisions, like the comparison for absolute and relative judgments, has implications not only for the use of confidence as an indicator of accuracy but also for decisions regarding the superior lineup procedure. In comparison with the simultaneous procedure, the overall superiority of the sequential procedure is the result of a reduction in the proportion of incorrect identifications from target-absent lineups, a reduction which is accompanied by an albeit smaller reduction in the proportion of correct identifications from target-present lineups (Steblay et al., 2001). In other words, less positive identifications are made from sequential lineups. If confidence is found to be a reliable indicator of the accuracy of positive identifications, then the simultaneous procedure may be less dangerous in comparison with the sequential lineup than has been suggested by the correct and false identification data alone. Whether any advantage in detection of errors using simultaneous lineups is sufficient to offset the lower rate of errors from sequential lineups is an empirical issue that depends both on the degree of calibration using both procedures and the size of the difference in initial error rates. Therefore, examination of CA calibration for positive and negative recognition decisions was another important focus of this work.

The first two experiments used different manipulations of task difficulty to provide converging evidence on the effect of difficulty on CA calibration for absolute and relative judgments and on the CA calibration difference between judgment types for equivalent difficulty levels. Further, the absolute judgment conditions allowed examination of CA calibration for positive and negative responses. Experiment 3 replicated the findings of these experiments using complex judgment tasks that more closely approximate the task of making an identification decision about a lineup.

Experiment 1

Experiment 1 used a within-subjects design to compare CA calibration for absolute and relative judgments at three different levels of difficulty. Manipulation of exposure duration has been demonstrated to influence the difficulty of recognition memory judgments for faces (e.g., Deffenbacher et al., 1981). Thus, we created the levels of difficulty by using three different exposure durations (200 ms, 500 ms, and 1,000 ms) in the study phase. Following Weber and Brewer (2003), who found superior calibration with a half-range (i.e., 50%–100%) rather than a full-range (i.e., 0%–100%) confidence scale, we used a half-range confidence scale.

Method

Participants. Forty-eight first year psychology students (9 male, 39 female), all with normal or corrected-to-normal vision, participated as part of a research participation exercise.

Materials. Three hundred color photographs of faces were used as stimuli. The photographs depicted individuals ranging in age from young adult to middle-age. The majority of the photographs were of people of Caucasian descent, with a small proportion of apparently Middle-Eastern origin. The photographs were collected from various sources. Some of the photos were taken by us, whereas the others were downloaded from the *Psychological Image Collection* at the University of Stirling (2001; http://pics.psych.stir.ac.uk/) and from the AR Face Database (Martinez & Benavente, 1998). The photographs were digitally edited so only the neck and face were shown and no clothing was visible. At both study and test the photographs were presented at a size of 200 pixels × 200 pixels on a monitor set at a resolution of 1,024 pixels × 768 pixels.

Design and procedure. The photographs were divided into six groups of 50 such that the ratio of male to female faces was consistent across groups. In each group, half of the faces of each gender were randomly selected as targets to be displayed in the study phase. For the relative (i.e., forced-choice) conditions each target was randomly paired with a distractor face of the same gender. The assignment of the six groups of stimuli to the six experimental conditions was counterbalanced across participants, as was the order of presentation of the stimuli groups.

Each participant completed six blocks of trials, each corresponding to one of the experimental conditions. Blocks were made up of a study phase, retention interval, and a test phase, which were identical across blocks except where noted. Before the study phase, participants were told that they would be shown a series of 25 photographs of faces and that their memory for them would be tested later. The photographs were presented for varying exposure durations (200 ms, 500 ms, and 1,000 ms) to create three levels of difficulty. These exposure durations were selected, based on pilot tests, to create three levels of performance without producing chance or perfect performance. The exposure duration was varied across, not within, blocks of trials, and a constant interstimulus interval of 500 ms was used. The retention interval lasted for 7 min, during which time participants completed a visuospatial memory task, the *visual matrix span task* (Wilson, Scott, & Power, 1987). The presentation of trials for this task was controlled by the computer to ensure that all participants completed the same number of trials in every retention interval.

Before the test phase, participants were given a description of the task so that they were aware of which type of judgment (absolute or relative) would be required and that they would need to report the confidence in their decision before the next trial began. The test phase consisted of either a series of 50 absolute judgment (i.e., yes–no) or 25 relative judgment (i.e., forced-choice) trials. It is important to note that for both judgment types participants were unaware of the number of trials and, consequently, were unaware of the proportion of old trials in the absolute conditions. In the absolute trials a single photograph was presented in the center of the screen, and two response buttons were displayed (equidistant from the vertical midline of the screen) below it. The left button was labeled *Seen* and the right *Not seen*. The participant clicked the appropriate button with the mouse to indicate whether the photograph was shown in the most recently presented series of photographs. In relative trials two photographs were displayed centered horizontally on the screen and equidistant from the vertical midline. A response button was presented under each photograph. The button under the left photograph was labeled *Face 1*, and the button under the right photograph was labeled *Face 2*. The participant clicked the appropriate button with the mouse to indicate which of the two faces was shown in the most recently presented series of photographs. Regardless of the recognition memory judgment required, immediately after making their response participants were required to indicate their confidence in the accuracy of their decision on a half-range scale (i.e., 50%–100%) with decile response options provided. They indicated their confidence by clicking on the appropriately labeled button on the screen. Participants were given no instruction on the use of the confidence scale, and no verbal anchors were displayed.

Results and Discussion

An alpha level of .05 was used for all inferential analyses, and Cohen's *f* was used as the effect size measure throughout. Values of Cohen's *f* greater than .40 are considered large effects, whereas the cut-off values for small and medium effects are .10 and .25, respectively. No evidence of biased responding was found in the relative judgment conditions (the same holds for Experiment 2). When examining the effect of the manipulation on the dependent measures for each condition, we calculated a single score for each participant separately: The proportion of correct responses was used as an indicator of accuracy, and the mean confidence as an index of confidence. Descriptive statistics for these scores are displayed in Table 1. For each dependent measure, we examined the effects of judgment type and difficulty using 2 × 3 repeated measures analyses of variance (ANOVAs).

For accuracy (i.e., proportion correct per participant) the ANOVA (see Table 2) revealed significant main effects for both judgment type and difficulty. Accuracy was greater in the relative rather than absolute conditions and was consistent with the expected effect of the manipulation. Examination of means and confidence intervals showed that accuracy decreased significantly with reduction in exposure duration for both judgment types. The Judgment Type × Difficulty interaction was nonsignificant.

For mean confidence (Table 2), significant main effects were identified for both judgment type and difficulty. Consistent with the accuracy data, relative judgments were made with more confidence than were absolute judgments, and examination of the means and confidence intervals suggested that confidence decreased significantly with each increase in task difficulty. However, in contrast with proportion correct, a significant interaction was also identified. Examination of the means suggests that the

Table 1
Descriptive Statistics for Accuracy and Confidence by Judgment Type and Difficulty for Experiment 1

	Judgment type		
Measure and difficulty	Absolute	Relative	Overall
Accuracy			
Easy			
M	0.75	0.87	0.81
SD	0.07	0.08	0.07
95% CI	0.73–0.77	0.84–0.89	0.79–0.82
Moderate			
M	0.70	0.81	0.76
SD	0.09	0.10	0.08
95% CI	0.68–0.73	0.78–0.84	0.73–0.78
Hard			
M	0.63	0.73	0.68
SD	0.08	0.10	0.07
95% CI	0.60–0.65	0.70–0.76	0.66–0.70
Overall			
M	0.69	0.80	0.75
SD	0.06	0.07	0.06
95% CI	0.68–0.71	0.78–0.82	0.73–0.76
Confidence			
Easy			
M	77.41	81.72	79.56
SD	9.11	8.95	8.55
95% CI	74.76–80.05	79.12–84.31	77.08–82.05
Moderate			
M	75.54	78.29	76.91
SD	9.14	9.15	8.57
95% CI	72.88–78.19	75.63–80.95	74.42–79.40
Hard			
M	71.95	73.30	72.63
SD	9.53	8.73	8.80
95% CI	69.19–74.72	70.76–75.84	70.07–75.18
Overall			
M	74.97	77.77	76.37
SD	8.71	8.27	8.23
95% CI	72.44–77.50	75.37–80.17	73.98–78.76

Note. CI = confidence interval.

Table 2
Repeated Measures ANOVAs for Confidence and Accuracy for Experiment 1

Dependent measure and source	*df*	*F*	*f*
Accuracy			
Judgment type (J)	1	157.64*	0.85
J error	47	(0.01)	
Difficulty (D)	2	65.41*	0.73
D error	94	(0.01)	
J × D	2	0.16	0.07
J × D error	94	(0.00)	
Confidence			
J	1	21.66*	0.16
J error	47	(26.12)	
D	2	56.58*	0.33
D error	94	(20.78)	
J × D	2	4.58*	0.07
J × D error	94	(11.50)	

Note. Values in parentheses represent mean-square errors. ANOVA = analysis of variance.
* $p < .05$.

interaction was the result of a decreasing difference in mean confidence between relative and absolute judgment conditions with increasing difficulty.

In sum, the exposure duration manipulation produced three levels of difficulty for each judgment type and significantly influenced confidence, but examination of the means suggests that decreasing the exposure duration had a greater impact on accuracy scores than on confidence. Such a result would be indicative of increasing overconfidence (or decreasing underconfidence) with increasing task difficulty, an issue that is explored fully in the following section.

Confidence–accuracy calibration. Figure 1 shows the confidence–accuracy calibration curves for the collapsed data in each of the six experimental conditions. All six curves display a generally similar shape, specifically, a positive relationship indicative of confidence and accuracy being calibrated. However, consistent with our expectations, the conditions obviously differed in over/underconfidence. Specifically, within each judgment type, overconfidence increased (or underconfidence decreased) with increasing difficulty. This pattern is also evident in the O/U and C statistics (see Table 3), calculated with data collapsed across participants. The insensitivity of the slope of the calibration curves to the effect of difficulty was also evident in the resolution statistics presented in Table 3, which, apart from the absolute–hard condition, were similar within each judgment type. The most striking result in the statistics, as with the calibration curves, is the impact of difficulty on overconfidence.

To confirm this effect of difficulty on overconfidence a 2 × 3 repeated measures ANOVA was conducted using within-subjects O/U (i.e., in each condition an O/U score was calculated separately for each participant) as the dependent measure. A significant main effect was found for judgment type, $F(1, 47) = 74.56$, $MSE = 0.01$, $f = 0.39$, indicating that judgments in the absolute conditions were more overconfident than those in the relative conditions. Similarly, the main effect of difficulty on O/U was significant, $F(2, 94) = 12.59$, $MSE = 0.01$, $f = 0.24$. No significant interaction was identified, $F(2, 94) = 0.44$, $MSE = 0.01$, $f = 0.03$, suggesting that difficulty did not influence O/U differently depending on judgment type. Power calculations, based on Cohen's (1988) power tables, suggest that this test was sufficiently powerful to detect a moderate effect (power = .97).

In addition to the effect of difficulty on over/underconfidence, two results are of particular interest. First, the effect of difficulty on C was different for absolute and relative judgments. Specifically, increasing difficulty improved calibration in the relative judgment conditions but impaired calibration in the absolute conditions. This is due to the different levels of over/underconfidence in the absolute and relative conditions. Absolute judgments were slightly overconfident in the easy condition, and therefore, increasing overconfidence (by increasing difficulty) caused judgments to become more poorly calibrated. In contrast, judgments in the

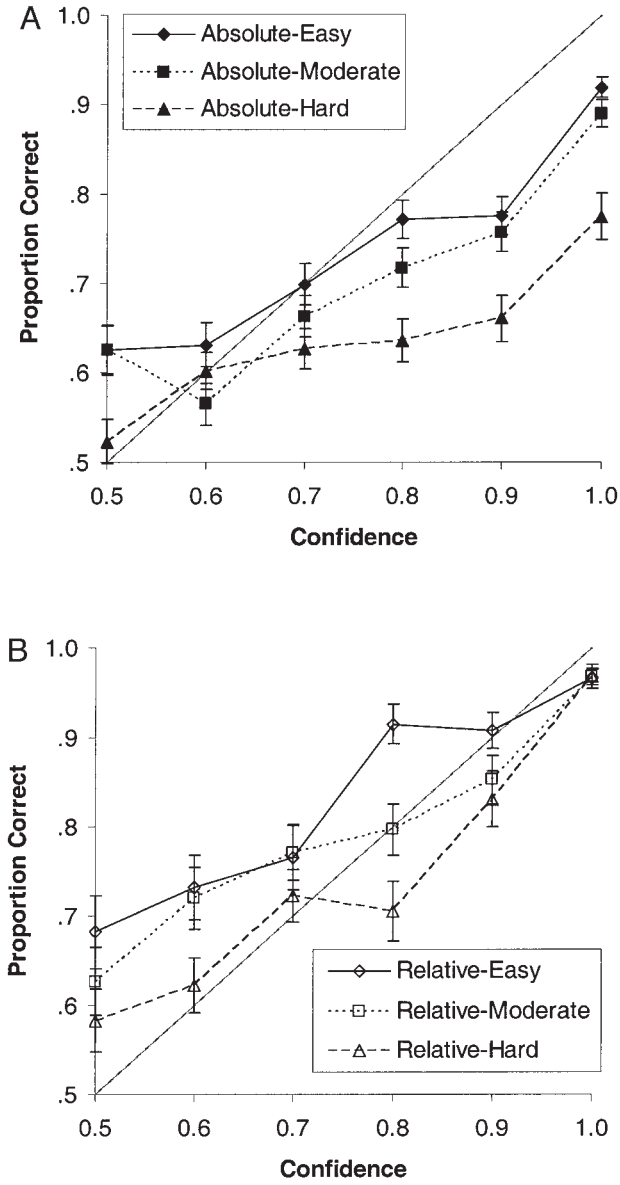

Figure 1. Calibration curves for absolute (A) and relative (B) judgments at each level of difficulty (Experiment 1).

Table 3
C, O/U, and Resolution Statistics by Judgment Type and Difficulty for Collapsed Data for Experiment 1

	Easy		Moderate		Hard	
Statistic	Absolute	Relative	Absolute	Relative	Absolute	Relative
C	.006	.009	.009	.006	.019	.003
Resolution	.062	.097	.054	.084	.019	.080
O/U	.021	−.045	.052	−.030	.094	.006
$SE_{O/U}$	.009	.010	.009	.011	.010	.012

Note. C = calibration statistic; O/U = over/underconfidence.

relative–easy condition were underconfident. Therefore, the increase in difficulty actually produced a decrease in underconfidence and, consequently, better calibration.

Second, although both judgment types displayed similar patterns of change in over/underconfidence with difficulty, the two still appear to differ in over/underconfidence when the difficulty of the task is taken into account. The mean accuracy data suggest that the relative–hard condition was easier than the absolute-moderate, but harder than the absolute–easy condition. If over/underconfidence does not differ between the two judgment types after task difficulty is accounted for, one would expect the O/U value for the relative–hard condition to be less than that observed in the absolute-moderate condition, but greater than that observed in the absolute–easy condition. However, examination of Table 3 indicates that the O/U in the relative–hard condition was greater than that in both the absolute–easy and –moderate conditions. This result suggests that although difficulty has a large impact on over/underconfidence, relative judgments are likely to be less overconfident than absolute judgments of an equivalent difficulty.

In sum, these data provide clear support for the prediction that increasing difficulty leads to greater levels of overconfidence and no evidence was found to suggest that the effect of difficulty on O/U differs between absolute and relative judgments. Further, calibration appears to be influenced by judgment type as well as difficulty. When the effect of difficulty is taken into account, absolute judgments appear to be made with more overconfidence than are relative judgments. One possible interpretation is that both judgment types share fundamentally similar decision and confidence scaling mechanisms, but differ sufficiently to produce the O/U difference for equivalent difficulties suggested by these data. An interesting alternative explanation is that despite a similarity in the underlying cognitive mechanisms, confidence responses are systematically distorted by the application of a conscious or unconscious heuristic. Koriat's (1997) accessibility hypothesis outlines a model whereby judgments of learning are affected, in some situations, by analytic heuristics relevant to the specific judgment. In this instance, a heuristic about the relative ease of absolute or relative judgments could differentially bias the confidence judgments produced by the same cognitive mechanism.

Positive versus negative decisions. Confidence–accuracy calibration was also compared for positive (i.e., old) and negative (i.e., new) recognition decisions in the absolute judgment conditions. Figure 2 displays the calibration curves for positive and negative recognition decisions at the three difficulty levels. The C, O/U, and resolution statistics for the collapsed data are displayed

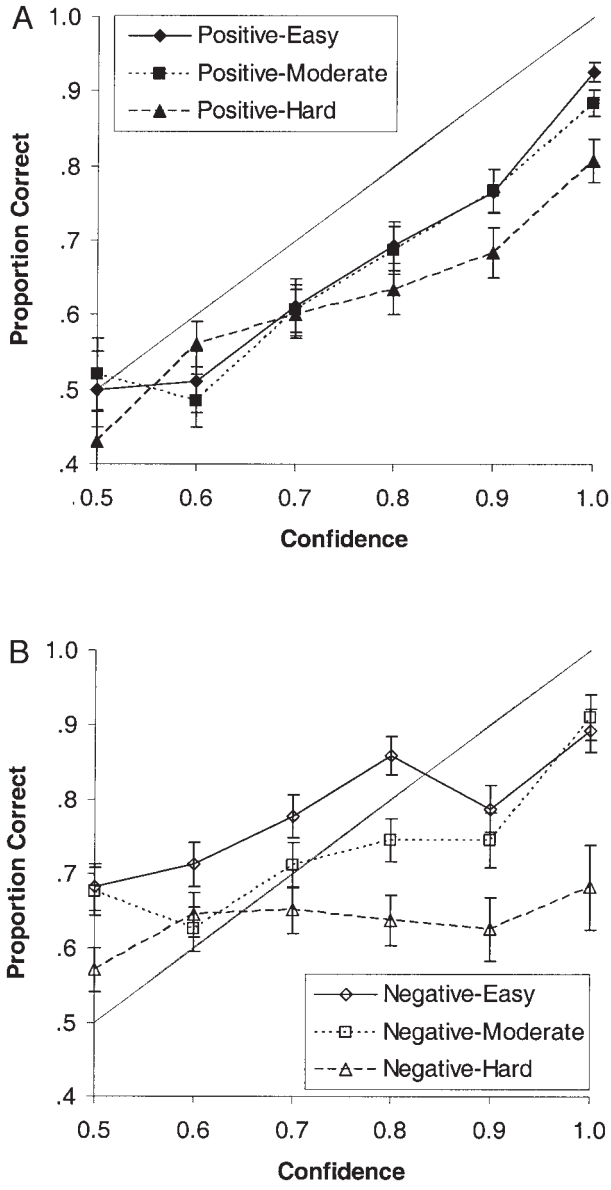

Figure 2. Calibration curves for positive (A) and negative (B) decisions from absolute judgment conditions at each level of difficulty (Experiment 1).

in Table 4. Two features of these data are striking. First, examination of both the calibration curves and O/U statistics indicates that the effect of difficulty on O/U was far more pronounced for negative recognition decisions. This observation was further supported by an analysis of simple main effects on within-subjects O/U that revealed a marginal impact of difficulty on O/U for positive recognition decisions, $F(2, 94) = 2.55$, $MSE = 0.01$, *ns*, $f = 0.13$, compared with an obvious effect on negative decisions, $F(2,94) = 25.67$, $MSE = 0.01$, $f = 0.34$. Second, an obvious difference between the slopes of the calibration curves for the two decision types is evident. Specifically, the calibration curves for positive decisions all display slopes of greater than .50 (easy: $b = 0.85$, $SE_b = 0.10$; moderate: $b = 0.79$, $SE_b = 0.10$; hard: $b = 0.65$, $SE_b = 0.08$), whereas the curves from negative decisions all have slopes less than .50, with the calibration curve for the negative–hard decisions being almost flat (easy: $b = 0.39$, $SE_b = 0.10$; moderate: $b = 0.45$, $SE_b = 0.13$; hard: $b = 0.14$, $SE_b = 0.07$). A similar pattern is observed in the resolution statistics, with the negative decisions displaying uniformly poor resolution at all difficulty levels. Interestingly, the resolution of positive decisions appears to have decreased with increasing difficulty suggesting a reduced ability to differentiate between correct and incorrect decisions at the hard-difficulty level, a finding consistent with the results of Baranski and Petrusic (1994). The combination of these two results suggests that, although confidence is a useful indicator of accuracy for positive recognition decisions, a much weaker association exists for negative decisions.

The relatively poor utility of confidence as an indicator of accuracy in negative decisions may be the result of the use of a recognition memory paradigm. When a face is identified as old, the degree of match between the stimulus face and a specific face in memory (whether the result of an information accumulation judgment process or some other) could be used as the basis of confidence. In contrast, no such comparison can be made for negative decisions as the information would have to reflect the degree of match with the entire set of study faces or perhaps the most similar face in the study set. This asymmetry between the nature of information available both to and as a result of the judgment process could be the cause of the CA calibration discrepancy between the two decision types. If this were the case, a recognition memory task in which each stimulus was compared with a specific study item in memory may not produce such a difference. It is interesting to note that an eyewitness identification would appear to be such a task as the witness is required to decide whether a member of the lineup is a specific individual, that is, the culprit of the witnessed crime. Therefore, this explanation would suggest that it is still possible that confidence could be a useful predictor of accuracy of negative decisions in an eyewitness identification task.

Another possible explanation is that the exposure duration manipulation itself may have been responsible for the interaction of recognition decision type and difficulty. Increasing the exposure duration of the study faces would have increased the discriminability of the old and new faces. As exposure duration was manipulated between blocks, participants would have been able to adjust their decision and confidence criteria for each exposure duration to keep their confidence judgments about old faces consistently calibrated. However, as the manipulation would not influence the familiarity of nonstudied faces, such an adjustment of confidence criteria would cause differences in the calibration of negative judgments. Thus, adjusting confidence criteria in response to a manipulation of encoding conditions could have produced consistent calibration for positive decisions but variable calibration for negative decisions. The results of Experiment 2 speak to these issues.

Experiment 2

To demonstrate that the effect of difficulty on CA calibration for absolute and relative judgments observed in Experiment 1 was the result of manipulating the difficulty of the task and not peculiar to the exposure duration manipulation, we replicated the experiment with a different difficulty manipulation. The exposure duration manipulation presumably influenced the strength or quality of the memory representations of the studied faces. Thus, the strongest converging evidence would be produced by a manipulation that affected difficulty through a different mechanism. Therefore, we manipulated a variable that would influence the decision process itself rather than alter the encoding of the study stimuli: specifically, the exposure duration of the faces at test.

Method

Participants. Fifty-eight first year psychology students (19 male, 39 female), all with normal or corrected-to-normal vision, participated.

Materials. A subset of 288 photographs of faces was selected pseudorandomly (gender ratio was preserved) from those used in Experiment 1 and presented at the same resolution and size.

Design and procedure. Except where noted, the design and procedure were identical to Experiment 1. As difficulty was manipulated within blocks in this experiment six groups of 48, rather than 50, photographs

Table 4
C, O/U, and Resolution Statistics by Decision Type and Difficulty for Collapsed Data for Experiment 1 From Absolute Conditions

	Easy		Moderate		Hard	
Statistic	Positive	Negative	Positive	Negative	Positive	Negative
C	.009	.014	.012	.010	.021	.021
Resolution	.131	.029	.099	.026	.048	.005
O/U	.087	−.055	.104	−.003	.127	.060
$SE_{O/U}$	.013	.012	.014	.012	.015	.014

Note. C = calibration statistic; O/U = over/underconfidence.

were used to allow an even number of trials at each of three difficulty levels per block. In the study phase, a constant exposure duration of 500 ms was used. Before the test phase participants were given a description of the task as in Experiment 1, with the additional warning that the time that the photographs would be visible would vary from trial to trial. They were also specifically instructed that they could take as long as they wanted to make their decision, regardless of the length of time for which the faces were visible. The test phase stimuli (either a single photograph or a pair of photographs) were displayed for either 200 ms, 500 ms, or 1,500 ms. These exposure durations were selected on the basis of pilot testing to create different levels of difficulty without producing chance or perfect performance and, more importantly, to produce absolute and relative conditions of equivalent difficulty. It is important to note that the response buttons were still visible and active after the stimuli disappeared. An equal number of trials of each difficulty level were used in each block and, in the absolute judgment conditions, half of the trials at each difficulty level presented an old photograph. Again, participants were unaware of the proportion of old trials. The order of exposure durations and the assignment of these to test stimuli were counterbalanced across participants. Again, the order of presentation of the groups of stimuli was counterbalanced across participants, as was the assignment of these groups to experimental condition.

Results and Discussion

For each dependent measure, we examined the effects of judgment type and difficulty using a 2 × 3 repeated measures ANOVA (Table 5), with relevant descriptive statistics displayed in Table 6. Accuracy displayed a significant main effect for difficulty but demonstrated no significant difference between judgment types. Planned comparisons were conducted to test the difference in accuracy between absolute and relative judgments at each exposure duration. They revealed that, for both the hard, $t(114) = 1.04$, $d = 0.16$, and moderate, $t(114) = 0.05$, $d = 0.01$, difficulty levels, accuracy did not significantly differ between the two types of judgment. In contrast, relative–easy judgments were found to be significantly more accurate than absolute–easy judgments, $t(114) = -4.09$, $d = 0.66$. Further, examination of the means and confidence intervals suggests that the manipulation produced three levels of difficulty in the relative condition but only two in the absolute condition. Specifically, the absolute–moderate and –easy conditions did not appear to differ in accuracy. Despite this similarity in accuracy, to aid comparison with the relative conditions we used the labels *easy* and *moderate* to identify these two conditions.

Table 5
Repeated Measures ANOVAs for Confidence and Accuracy for Experiment 2

Dependent measure and source	*df*	*F*	*f*
Accuracy			
Judgment type (J)	1	2.80	0.14
J error	57	(0.01)	
Difficulty (D)	2	24.83*	0.50
D error	114	(0.01)	
Confidence			
J	1	8.06*	0.07
J error	57	(19.86)	
D	2	31.74*	0.24
D error	114	(29.93)	
J × D	2	12.79*	0.11
J × D error	114	(15.20)	

Note. Values in parentheses represent mean-square errors. ANOVA = analysis of variance.
* $p < .05$.

Table 6
Descriptive Statistics for Accuracy and Confidence by Judgment Type and Difficulty for Experiment 2

	Judgment type		
Measure and difficulty	Absolute	Relative	Overall
Accuracy			
Easy			
M	0.71	0.78	0.75
SD	0.09	0.10	0.07
95% CI	0.69–0.74	0.75–0.80	0.73–0.77
Moderate			
M	0.70	0.70	0.70
SD	0.08	0.10	0.07
95% CI	0.68–0.72	0.67–0.73	0.68–0.72
Hard			
M	0.67	0.65	0.66
SD	0.08	0.12	0.08
95% CI	0.65–0.69	0.62–0.68	0.64–0.68
Overall			
M	0.69	0.71	0.70
SD	0.07	0.07	0.06
95% CI	0.68–0.71	0.69–0.73	0.69–0.72
Confidence			
Easy			
M	76.48	77.98	77.22
SD	9.66	9.48	9.21
95% CI	73.94–79.02	75.47–80.46	74.80–79.64
Moderate			
M	74.82	72.83	73.83
SD	9.65	10.10	9.58
95% CI	72.28–77.36	70.17–75.49	71.31–76.34
Hard			
M	73.32	69.75	71.53
SD	9.23	12.35	10.32
95% CI	70.89–75.75	66.50–73.00	68.82–74.25
Overall			
M	74.87	73.51	74.19
SD	9.32	9.41	9.18
95% CI	72.42–77.32	71.04–75.99	71.78–76.61

Note. CI = confidence interval.

The ANOVA on mean confidence revealed significant main effects for both judgment type and difficulty. Mean confidence in the easy conditions was significantly greater than that in the moderate conditions which was, in turn, greater than mean confidence in the hard conditions. Thus, relative judgments were made with less confidence than absolute judgments, and confidence decreased with increasing task difficulty. Further, a significant interaction was also identified, with difficulty having a larger impact on the confidence of relative judgments than on the confidence of absolute judgments. In sum, the data were similar to those in Experiment 1. The manipulation affected both confidence and accuracy, but examination of the means and effect size measures suggests confidence was not as sensitive as accuracy. Again, these results suggest a likely difference in O/U between different levels of difficulty.

Confidence–accuracy calibration. CA calibration curves for collapsed data from each of the six conditions are displayed in

Figure 3. The curves are consistent with the observations from Experiment 1. Specifically, the curves for both conditions display a positive CA relationship indicative of calibration. It is important to note that the curves display effects of difficulty consistent with those observed in Experiment 1. Specifically, as difficulty increased the curves display increasing overconfidence. Importantly, the curves from the absolute conditions with similar accuracy (i.e., the easy and moderate conditions) are almost identical. Again, these patterns are also evident in the C, O/U, and resolution statistics displayed in Table 7.

The effect of difficulty on overconfidence was examined using a 2 × 3 repeated measures ANOVA with within-subjects O/U as the dependent measure (Table 8). A significant main effect was found for difficulty on O/U. Similarly, a significant main effect of judgment type was identified, indicating that judgments in the absolute conditions were more overconfident than those in the relative conditions. The interaction was not significant. Again, power calculations suggest that this test was sufficiently powerful to detect a moderate effect (power = .99).

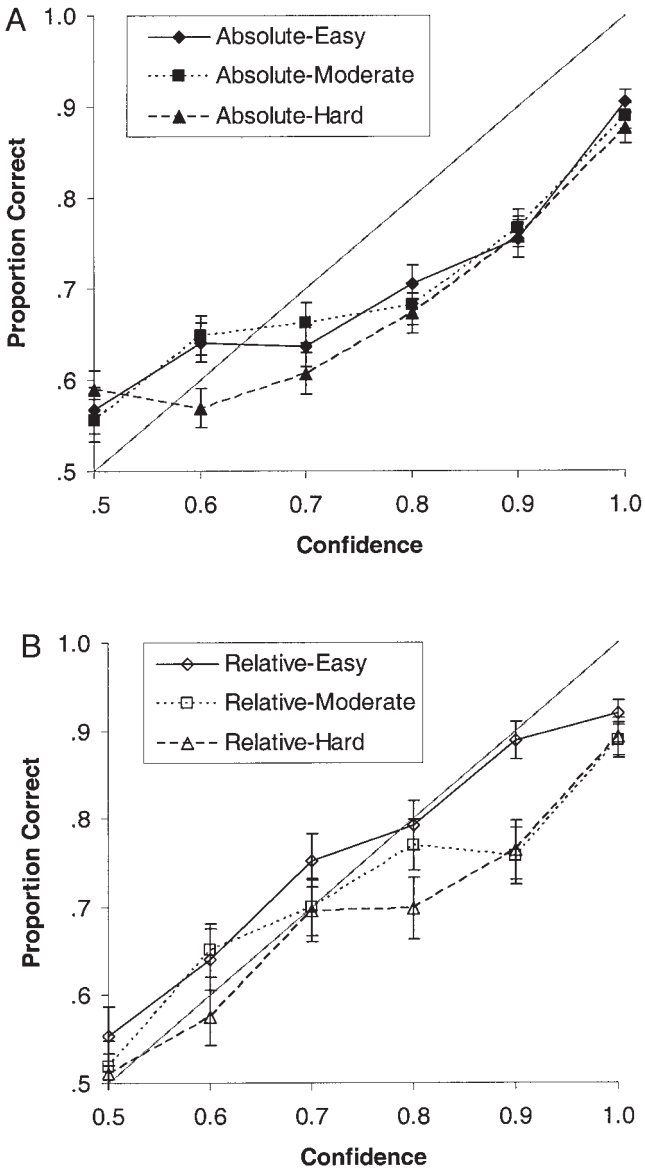

Figure 3. Calibration curves for absolute (A) and relative (B) judgments at each level of difficulty (Experiment 2).

The finding that the absolute– and relative–hard as well as the absolute– and relative–moderate conditions did not differ in accuracy allows us to specifically test some of the observations made in Experiment 1. First, to test for a difference in the effect of difficulty on O/U between judgment types, we compared the difference in O/U between the hard and moderate conditions for the two judgment types. A paired-samples t test suggested that mean O/U difference for absolute judgments ($M = 0.02$, $SD = 0.09$) did not differ significantly from the mean difference for relative judgments ($M = 0.02$, $SD = 0.12$), $t(57) = .02$, $f = 0.00$. Power calculations, based on a linear interpolation from the power tables, suggest that this test was sufficiently powerful to detect a moderate effect (power ≈ .80). Therefore, the test phase exposure duration manipulation produced an equivalent change in accuracy for both absolute and relative judgments and also in O/U, suggesting that the effect of the manipulation on O/U did not depend on judgment type.

The second and most important test that the equivalent accuracies allowed was a comparison of O/U for absolute and relative judgments at the same level of accuracy. We used a 2 × 2 ANOVA with the conditions of equivalent difficulty (i.e., hard and moderate) and within-subjects O/U as the dependent measure. The ANOVA revealed a significant difference in O/U between the absolute and relative judgment conditions, $F(1, 57) = 4.24$, $MSE = 0.01$, $f = .09$. Thus, in accord with the observations in Experiment 1, absolute judgments were made with more overconfidence than were relative judgments of the same difficulty, though the effect size was relatively small.

Positive versus negative decisions. As in Experiment 1, we compared CA calibration for positive and negative recognition decisions. Examination of the calibration curves (Figure 4) suggests a pattern very similar to that observed in the first experiment. Little difference is evident in the curves for positive decisions, whereas the curves for negative decisions show a clear effect of difficulty on O/U. Additionally, the positive curves display a slope of approximately 1.00 (easy: $b = 0.89$, $SE_b = 0.10$; moderate: $b = 0.93$, $SE_b = 0.10$; hard: $b = 0.82$, $SE_b = 0.11$), whereas the negative curves all display slopes less than .50 (easy: $b = 0.41$, $SE_b = 0.07$; moderate: $b = 0.29$, $SE_b = 0.07$; hard: $b = 0.39$, $SE_b = 0.08$). These features are also apparent in the O/U statistics (Table 9), which demonstrate a clear distinction in overconfidence between the negative–hard condition and the negative–moderate and negative–easy conditions, while displaying no meaningful difference between the conditions for positive decisions. Additionally, examination of the resolution statistics shows that, for negative recognition decisions, confidence was virtually useless as a marker of correct and incorrect responses. In contrast, reasonable resolution was displayed for positive recognition decisions. The overconfidence difference was further explored with an analysis of simple main effects ($\alpha = .025$) using within-subjects O/U as the dependent measure. It demonstrated that, as in the first experiment, the interaction was the result of a difficulty effect on O/U for negative decisions, $F(2, 114) = 3.98$, $MSE = 0.01$, $f = 0.13$, but not for positive decisions, $F(2, 114) < 1$, $MSE < 0.01$, $f = 0.02$.

Table 7
C, O/U, and Resolution Statistics by Judgment Type and Difficulty for Collapsed Data for Experiment 2

	Easy		Moderate		Hard	
Statistic	Absolute	Relative	Absolute	Relative	Absolute	Relative
C	.008	.003	.008	.005	.011	.005
Resolution	.064	.102	.052	.071	.051	.079
O/U	.051	.002	.047	.038	.065	.045
$SE_{O/U}$	.008	.011	.009	.012	.009	.012

Note. C = calibration statistic; O/U = over/underconfidence.

Interestingly, this result rules out the role of the encoding manipulation in producing this interaction described earlier. If the interaction was a result of participants adjusting their confidence criteria in response to different encoding conditions, the interaction should not be produced by a manipulation that does not influence encoding conditions and does not provide participants with any a priori knowledge of the test conditions.

Experiment 3

Experiments 1 and 2 provided a clear demonstration that (a) after controlling for difficulty, the difference in CA calibration for absolute and relative judgments is negligible, and (b) there is no difference between absolute and relative judgments in the effect of difficulty on CA calibration. The face recognition paradigm used in these experiments, however, differs from the eyewitness identification task in many ways. For example, in an eyewitness identification witnesses are required to make a more complex judgment about multiple stimuli not a single face or pair of faces, to encode an image from a moving three-dimensional stimulus rather than a still photograph, and to recognize an individual whose appearance may have changed rather than the same image that was encoded at study. Further, the use of multiple recognition judgments may lead participants to adopt a different decision criterion in this task than they would use when making a single judgment from a lineup. This experiment was designed to examine the CA calibration for two types of judgment that more closely resemble those required in an eyewitness identification context while still controlling for the effects of difficulty. Specifically, in one condition, instead of a purely relative judgment, participants were faced

Table 8
Repeated Measures ANOVAs for Over/Underconfidence for Experiment 2

Source	*df*	*F*	*f*
Judgment type (J)	1	13.65*	0.14
J error	57	(0.01)	
Difficulty (D)	2	3.80*	0.11
D error	114	(0.01)	
J × D	2	1.66	0.06
J × D error	114	(0.01)	

Note. Values in parentheses represent mean-square errors. ANOVA = analysis of variance.
* $p < .05$.

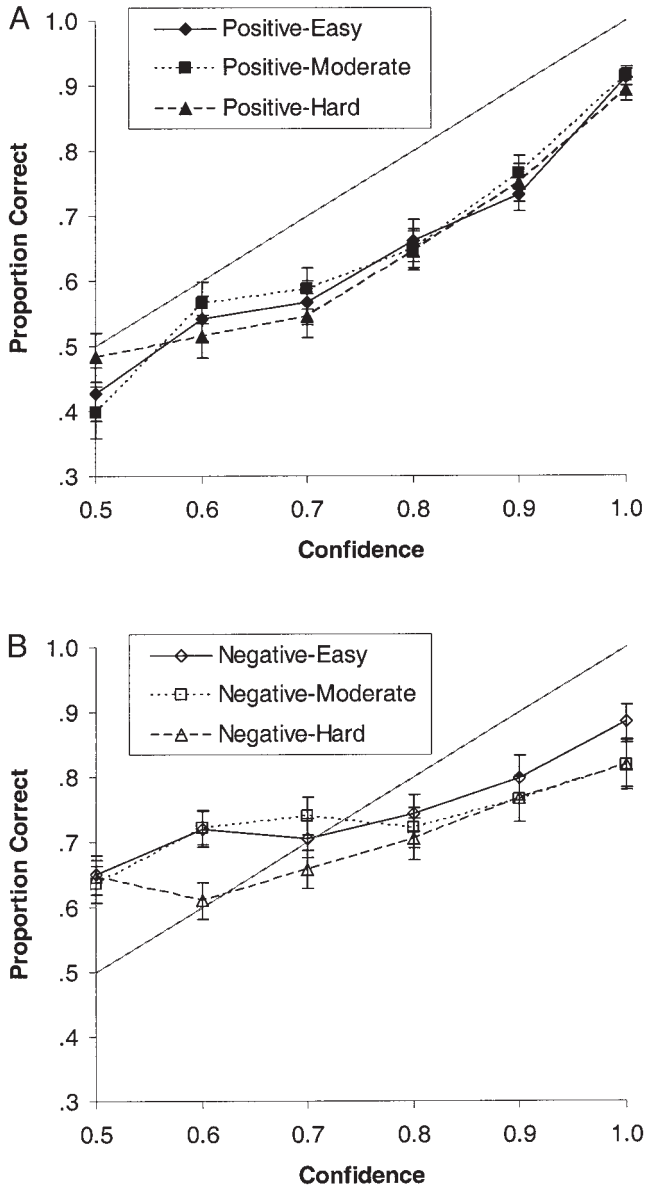

Figure 4. Calibration curves for positive (A) and negative (B) decisions from absolute judgment conditions at each level of difficulty (Experiment 2).

Table 9
C, O/U, and Resolution Statistics by Decision Type and Difficulty for Collapsed Data From Absolute Conditions for Experiment 2

	Easy		Moderate		Hard	
Statistic	Positive	Negative	Positive	Negative	Positive	Negative
C	.014	.011	.012	.014	.015	.011
Resolution	.126	.033	.120	.020	.103	.017
O/U	.111	−.021	.102	−.019	.114	.012
$SE_{O/U}$	.011	.013	.011	.013	.012	.013

Note. C = calibration statistic; O/U = over/underconfidence.

with the possibility that none of the stimuli were presented earlier, thus requiring some degree of absolute processing as in a simultaneous lineup. Similarly, in the other condition, instead of a single absolute judgment, participants were required to make judgments about each of a group of stimuli, presented one at a time. Although this task still required an absolute judgment about each face in the group, for all stimuli after the first some degree of relative processing was possible just as in a sequential lineup. These judgment tasks were operationalized as two types of mini-lineup. In the simultaneous mini-lineup, four faces, one or none of which were presented in the study phase, were presented to the participant who was required to identify the face that was shown previously or to indicate that none of the faces were presented in the study phase. The sequential mini-lineups also required participants to identify which, if any, of a group of four faces had been shown in the study phase. However, the stimuli were presented one at a time, and a judgment (i.e., seen or not seen) was required for each face before the next was shown.

This approach was chosen, rather than progressing immediately to an identification experiment, for a number of reasons. First, if a different pattern of results was observed in an identification experiment that difference could be the result of any of a number of differences between the identification and face recognition paradigms. But, by systematically increasing the complexity of the experimental paradigm in incremental steps that produce systematically better approximations of the identification task, the origin of an inconsistent result can be identified more easily. Thus, such a research program may help to highlight the differences between the processes underlying the standard face recognition judgment and those underlying an eyewitness identification decision. Second, the difficulty of an identification task is thought to be determined, at least partly, by characteristics of the offender and also by the structure of the lineup, but these effects are poorly understood. Therefore, to control for the impact of the stimuli on task difficulty, a large (at least in the context of eyewitness identification experiments) number of different stimuli, counterbalanced across conditions, would be required. Consequently, such an experiment would prove a huge, and at this stage unwarranted, practical undertaking. In contrast, the use of a modified face recognition paradigm allows for the practical use of a large number of stimuli counterbalanced across the experimental conditions. Finally, as discussed in Weber and Brewer (2003), examination of calibration requires a large number of observations. In an identification experiment, this leads to the requirement for huge sample sizes (e.g., $N = 900+$ in Brewer et al., 2002), or an unrealistic number of stimuli. By efficiently allowing collection of many observations per participant, the modified face recognition paradigm overcomes this difficulty.

To guard against the possible influence of a change in confidence scale, as observed by Weber and Brewer (2003), the half-range confidence scale used in Experiments 1 and 2 was also used in this experiment. The use of this scale creates an issue in the interpretation of the magnitude of the O/U value observed. Specifically, as participants chose from more than two response options, the probability of a correct guess was less than 50%, thus the limited range of the confidence scale may have artificially inflated confidence judgments and consequently the value of O/U observed. As we were interested in the difference in O/U observed in two conditions, the distorted absolute value of overconfidence was not problematic.

Method

Participants. Sixty-four first year psychology students (17 male, 47 female), all with normal or corrected-to-normal vision, participated.

Materials. This experiment used 320 photographs, the 300 used in Experiment 1 and 20 more collected from the same sources. All photographs were presented with the same on-screen size and resolution as in Experiment 1.

Design and procedure. To ensure that the mini-lineups consisted of similar looking faces, 80 groups of four similar faces were created, and one of the faces in each group was randomly selected as the target for that mini-lineup. Faces were assigned to these groups on the basis of ratings from four individuals on six dimensions: gender, face shape, hair color and style, eye color, skin color and complexion, and age. To allow four blocks of trials, the 80 mini-lineups were then assigned to one of four groups with an equal gender ratio in each. Of the 20 targets in each block, only 10 were shown in the study phase for any given participant, thus creating 10 target-present and 10 target-absent mini-lineups in the test phase. The selection of these targets for presentation in the study phase was counterbalanced across participants and, thus, each mini-lineup was used as target-present and target-absent an equal number of times.

The study phase was similar to Experiments 1 and 2, with participants shown a series of 10 photographs of faces with an exposure duration of 250 ms or 1,000 ms in each block. As in Experiments 1 and 2, exposure durations were selected to create a difference in difficulty but to avoid floor and ceiling performance. Exposure duration was constant within blocks. The retention interval was identical to those in Experiments 1 and 2. Before the test phase, the task was described to participants and they were explicitly told that, in each group of four faces shown to them, either one or none of the faces was presented to them in the preceding study phase. In the test phase, participants were presented with a series of target-present

or target-absent mini-lineups. Two types of mini-lineup were used, but only one type was used in any given block of trials.

In the simultaneous mini-lineup task, four photographs were presented in a single row, vertically centered and evenly spaced across the width of the screen. Below the photographs a single button labeled *Not present* was displayed. Participants were instructed to click on the photograph of the face that they thought was shown to them in the immediately preceding study phase or, if they thought none of the faces were shown in the study phase, click on the *Not present* button. After indicating their decision, participants rated their confidence in the accuracy of their decision on the same scale used in the other two experiments. Participants were instructed that after making a positive response their confidence estimate should reflect how confident they were that the photograph they indicated had been presented in the study phase and that the other three photographs had not been presented.

In contrast, the sequential mini-lineup required the participants to make a decision about each photograph in the mini-lineup separately. Specifically, the photographs were presented one at a time in the same manner as the absolute judgments in the preceding experiments: that is, in the center of the screen with two response buttons (*Seen* and *Not seen*) below the photograph. Thus, for the first photograph in a mini-lineup the participant was instructed to click the *Seen* button if the face was shown in the immediately preceding study phase and the *Not seen* button otherwise. If the participant clicked the *Not seen* button the next photograph in the mini-lineup was presented, and they were required to make the same seen–not-seen judgment. This continued until either the participant made a positive response about one of the photographs (i.e., clicked the *Seen* button) or all four photographs in the mini-lineup had been presented. At this point participants were required to indicate their confidence in the accuracy of their decision in the same manner as for the simultaneous mini-lineup. As a result of the complicated nature of the decision, or decisions, involved in a sequential mini-lineup, specific instructions were given to participants about their confidence estimates. They were told that their response should indicate how confident they were that (a) none of the photographs were presented in the preceding study phase (if they clicked *Not seen* for each photograph), or (b) the photograph they selected as seen was shown in the preceding study phase and none of the other photographs was shown (if they clicked *Seen* for one of the photographs).

The exposure duration used in the study phase and the mini-lineup type used in the test phase were crossed to produce four experimental conditions. Each was used in a separate block of trials. The order of presentation of both the groups of photographs and the experimental conditions, as well as the assignment of the groups to experimental conditions, was counterbalanced across participants.

Results and Discussion

We examined the effect of difficulty and mini-lineup type on both confidence and accuracy using 2 × 2 repeated measures ANOVAs (see Table 10 for ANOVA statistics and Table 11 for descriptive statistics). A significant main effect of difficulty was identified for both confidence and accuracy, indicating that recognition decisions in the hard condition were made less accurately and with less confidence than those in the easy conditions, confirming the predicted effect of the manipulation. For confidence and accuracy, both the mini-lineup type main effect and the Difficulty × Mini-Lineup Type interaction were nonsignificant. Therefore, decision confidence and accuracy did not differ for the two judgment types. Further, the effect of difficulty on confidence and accuracy was not moderated by mini-lineup type.

Confidence–accuracy calibration. The CA calibration curves are displayed in Figure 5. As in the first two experiments, a clear effect of difficulty on O/U is present in both the calibration curves and O/U statistics (Table 12) and is apparently equivalent for both judgment types. To confirm the presence of the hard–easy effect for the mini-lineup judgment tasks, we conducted a 2 × 2 ANOVA with within-subjects O/U as the dependent measure (Table 13). Although the main effect of mini-lineup type and the Mini-Lineup Type × Difficulty interaction were nonsignificant, a significant main effect of difficulty was identified. Power calculations suggest that this test was sufficiently powerful to detect a moderate sized interaction effect (power = .80).

Table 10
Repeated Measures ANOVAs for Confidence and Accuracy for Experiment 3

Dependent measure and source	*df*	*F*	*f*
Accuracy			
Mini-lineup type (M)	1	1.86	0.08
M error	63	(0.01)	
Difficulty (D)	1	47.04*	0.46
D error	63	(0.02)	
M × D	1	0.69	0.03
M × D error	63	(0.01)	
Confidence			
M	1	0.75	0.03
M error	63	(21.73)	
D	1	32.98*	0.21
D error	63	(26.44)	
M × D	1	0.04	0.03
M × D error	63	(30.14)	

Note. Values in parentheses represent mean-square errors. ANOVA = analysis of variance.
* $p < .05$.

In contrast with Experiments 1 and 2, all of the curves are nonlinear. This was confirmed by simultaneous regressions, predicting proportion correct from confidence and confidence squared. For each condition, they indicated that the quadratic term (i.e., a nonlinear term) significantly predicted proportion correct (simultaneous–easy: $b = 2.51$, $t(4) = 3.28$; simultaneous–hard: $b = 3.46$, $t(4) = 5.64$; sequential–easy: $b = 6.34$, $t(4) = 4.48$; sequential-hard: $b = 5.05$, $t(4) = 5.51$). Specifically, for each condition the two or three lowest confidence categories appear to have been associated with approximately equivalent proportion correct, despite an obvious CA association in the top half of the confidence categories. This nonlinearity is reflected in the relatively poor resolution statistics for all of the conditions. This poor resolution could be the result of an inappropriate calculation of an internal index of confidence from the accumulated information used to make the decision, the inappropriate scaling of such information, or possibly some combination of the two.

Importantly, the nonlinearity of the calibration curves could be the result of the confidence scale on which participants were required to respond. In Experiments 1 and 2, the probability of the correct decision being made by guessing was 50%. Thus, the lowest value of the half-range confidence scale (i.e., 50%) corresponded with the likely accuracy of a guess. In contrast, the probability of a correct guess in Experiment 3 was much less than 50% (20% if all five response options are considered equivalent in a guess response) and therefore did not correspond with the lowest value on the half-range confidence scale. Such a conclusion is suggested by the results of Weber and Brewer (2003), which

Table 11
Descriptive Statistics for Accuracy and Confidence by Mini-Lineup Type and Difficulty for Experiment 3

	Mini-lineup type		
Measure and difficulty	Simultaneous	Sequential	Overall
Accuracy			
Easy			
M	0.62	0.61	0.61
SD	0.15	0.16	0.14
95% CI	0.58–0.65	0.57–0.65	0.58–0.65
Hard			
M	0.50	0.48	0.49
SD	0.14	0.14	0.12
95% CI	0.47–0.54	0.44–0.51	0.46–0.52
Overall			
M	0.56	0.54	0.55
SD	0.12	0.12	0.11
95% CI	0.53–0.59	0.51–0.57	0.52–0.58
Confidence			
Easy			
M	74.27	74.63	74.45
SD	9.74	9.02	8.53
95% CI	71.83–76.70	72.37–76.88	72.31–76.58
Hard			
M	70.43	71.98	70.75
SD	9.07	10.27	9.13
95% CI	68.16–72.70	68.51–73.64	68.47–73.03
Overall			
M	72.35	72.85	72.60
SD	8.65	8.89	8.45
95% CI	70.19–74.51	70.63–75.07	70.49–74.71

CI = confidence interval.

demonstrated no CA association in the lower half of a full-range (i.e., 0%–100%) confidence scale but good calibration in the upper half (i.e., the 50%–100% range). Thus, it appeared that participants were unable to scale their confidence judgments to fit the range of

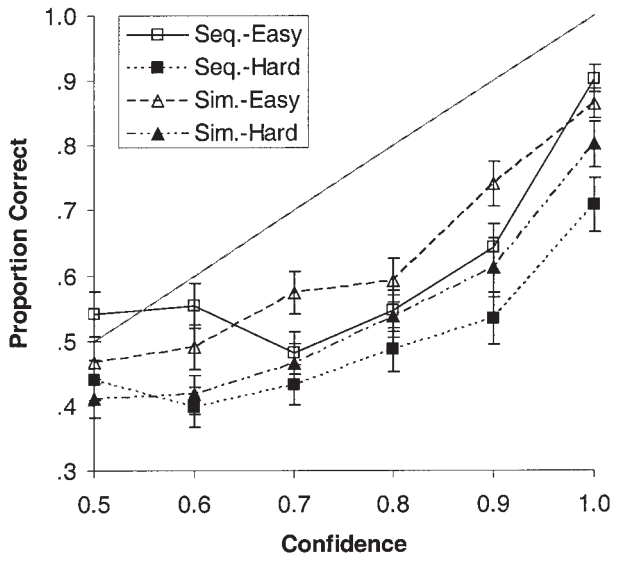

Figure 5. Calibration curves for simultaneous (Sim.) and sequential (Seq.) mini-lineup conditions at each level of difficulty (Experiment 3).

Table 12
C, O/U, and Resolution Statistics by Mini-Lineup Type and Difficulty for Collapsed Data for Experiment 3

	Easy		Hard	
Statistic	Simultaneous	Sequential	Simultaneous	Sequential
C	.019	.032	.044	.066
Resolution	.083	.083	.056	.030
O/U	.126	.139	.200	.236
$SE_{O/U}$	.013	.013	.014	.014

Note. C = calibration statistic; O/U = over/underconfidence.

a scale that was not bounded by the probability of an accurate guess and certainty. In this experiment, participants were faced with a similar problem although, instead of being required to spread their confidence judgments across a scale with a range greater than the guess probability to certainty, they were required to fit their confidence judgments into a restricted range, with the minimum response having a value higher than the guess probability. Thus, the problem could have arisen as the higher end of the scale was used in a numerically appropriate way by participants, but the lower end of the scale was forced to accommodate confidence values from outside the range of the scale in addition to numerically appropriate responses. Recent work in our laboratory (Weber & Brewer, 2004), however, suggests that this explanation is not correct, as the same pattern was observed with a full-range confidence scale.

Positive versus negative decisions. CA calibration curves for positive and negative recognition decisions are presented in Figure 6. Examination of these curves reveals a pattern consistent with those observed in Experiments 1 and 2. Specifically, good resolution is evident for positive decisions and virtually no resolution for negative decisions. This is reflected in the steeply sloping calibration curves for positive decisions (simultaneous–easy: $b = 1.48$, $SE_b = 0.17$; simultaneous–hard: $b = 1.25$, $SE_b = 0.20$; sequential–easy: $b = 1.39$, $SE_b = 0.32$; sequential–hard: $b = 0.95$, $SE_b = 0.20$), the almost flat (i.e., zero slope) curves for negative decisions (simultaneous–easy: $b = -0.18$, $SE_b = 0.21$; simultaneous–hard: $b = 0.28$, $SE_b = 0.09$; sequential–easy: $b = -0.09$, $SE_b = 0.10$; sequential–hard: $b = 0.18$, $SE_b = 0.04$), and the resolution statistics displayed in Table 14. In contrast with the first two experiments, though, difficulty appears to have affected O/U

Table 13
Repeated Measures ANOVAs for Over/Underconfidence for Experiment 3

Source	*df*	*F*	*f*
Mini-lineup type (M)	1	3.12	0.08
M error	63	(0.01)	
Difficulty (D)	1	25.44*	0.28
D error	63	(0.01)	
M × D	1	0.81	0.03
M × D error	63	(0.01)	

Note. Values in parentheses represent mean-square errors. ANOVA = analysis of variance.
* $p < .05$.

for both types of judgments. This was confirmed by a 2 (mini-lineup type) × 2 (difficulty) × 2 (decision type) ANOVA[1] on O/U that identified no significant Difficulty × Decision Type interaction, $F(1, 61) = 3.11$, $MSE = 0.01$, $f = 0.00$.

Interestingly, the ANOVA also revealed a significant Mini-Lineup Type × Decision Type interaction, $F(1, 61) = 10.09$, $MSE = 0.02$, $f = 0.13$. Examination of the O/U statistics suggests that the source of the interaction was a difference in O/U between the lineup types for positive decisions but not for negative decisions. Thus, positive decisions from a sequential mini-lineup were more overconfident than those from a simultaneous mini-lineup, whereas negative decisions from the two mini-lineups produced equivalent overconfidence. This difference between the mini-lineup types may be due to the extent to which each allows relative judgments to be made. Experiments 1 and 2 suggest that absolute judgments are made with slightly more overconfidence than are relative judgments. When a positive decision is made, the sequential mini-lineup obviously provides less scope for the participant to use a relative judgment process than does the simultaneous mini-lineup, so the difference is consistent with the earlier results. Further, for negative decisions one would predict no O/U difference resulting from the extent to which absolute or relative judgment processes could be engaged, as negative decisions cannot be made using a relative judgment process.

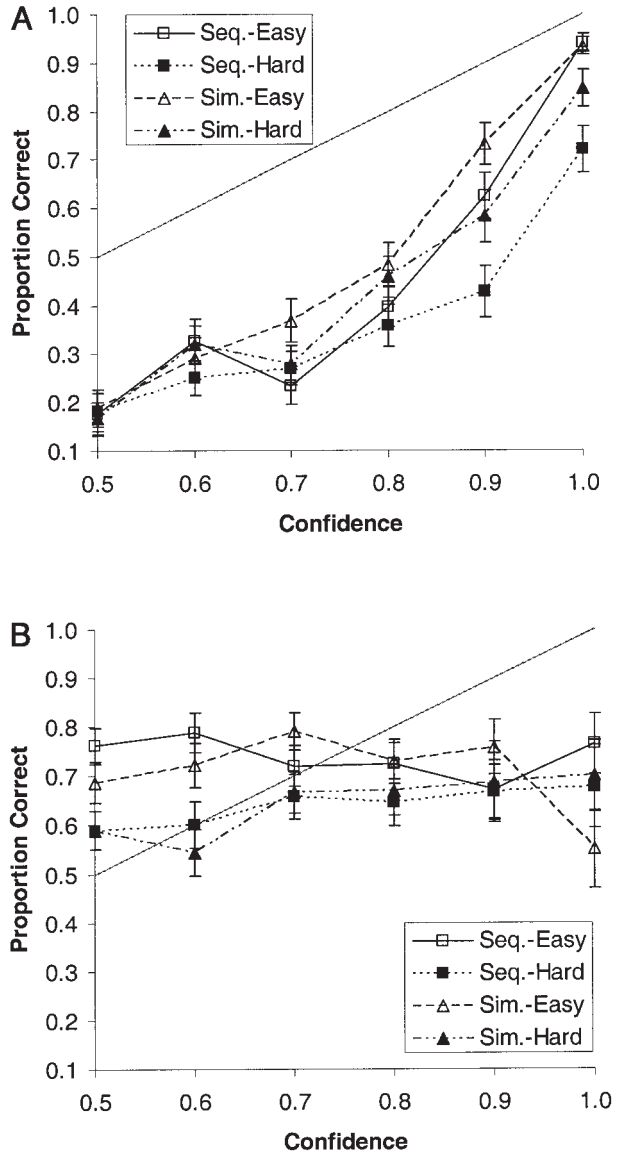

Figure 6. Calibration curves for positive (A) and negative (B) decisions from the simultaneous (Sim.) and sequential (Seq.) mini-lineup conditions at each level of difficulty (Experiment 3).

Table 14
C, O/U, and Resolution Statistics by Decision Type, Mini-Lineup Type, and Difficulty for Collapsed Data From Absolute Conditions for Experiment 3

Mini-lineup type and statistic	Easy		Hard	
	Positive	Negative	Positive	Negative
Simultaneous				
C	.067	.031	.106	.015
Resolution	.300	.018	.174	.103
O/U	.234	−.025	.317	.047
$SE_{O/U}$	.016	.021	.017	.021
Sequential				
C	.101	.033	.154	.019
Resolution	.313	.006	.113	.005
O/U	.284	−.040	.387	.055
$SE_{O/U}$	.016	.020	.017	.021

Note. C = calibration statistic; O/U = over/underconfidence.

General Discussion

The purpose of this study was to examine CA calibration for (a) absolute and relative face recognition judgments under conditions producing different task difficulty levels and (b) positive and negative recognition decisions. We used two different manipulations of task difficulty, one affecting encoding conditions and the other affecting test conditions, to produce converging evidence in the basic judgment tasks. Additionally, the use of judgment tasks that more closely resemble real world eyewitness identification tests than does the face recognition paradigm provides strong evidence for the generality of the observed results. Three striking features were observed in all three experiments. First, at equivalent levels of difficulty a very small but reliable difference in calibration between the judgment types was demonstrated, with absolute judgments made with marginally more overconfidence than relative judgments of equivalent difficulty. Second, the effect of difficulty on overconfidence did not interact with judgment type. Finally, in conditions that required some type of absolute judgment (i.e., the absolute judgment conditions in the first two experiments and both types of mini-lineup used in Experiment 3), a striking difference in calibration for positive and negative recognition decisions was observed. Specifically, for negative decisions CA

[1] Two of the 64 participants made all positive responses in one of the four conditions; therefore, this analysis was conducted only on the data of those 62 participants who made at least one negative decision in each condition.

calibration curves were almost flat indicating that confidence was useless as an indicator of accuracy for those decisions. In contrast, CA calibration curves for positive decisions showed a clear monotonic, positive relationship between confidence and accuracy suggesting confidence was a useful marker of accuracy for positive decisions. Additionally, and not surprisingly, the presence of a consistent hard–easy effect, observed in a number of other research areas, was clearly confirmed. That is, the degree of overconfidence with which recognition judgments were made increased with task difficulty.

These results have a number of important implications for the eyewitness identification domain. Much of the recent literature in that domain has centered on a push for the adoption of sequential versus simultaneous lineups as, relative to simultaneous lineups, sequential procedures appear to produce markedly fewer incorrect identifications in target-absent lineups, typically with a smaller reduction in the proportion of correct identifications in a target-present lineup. In deciding which lineup procedure is superior, though, accuracy is not the only variable that should be considered. Recent evidence has suggested that, at least in some circumstances, good CA calibration can be obtained for positive eyewitness identification decisions (Brewer et al., 2002; Juslin et al., 1996; Olsson, 2000; Olsson & Juslin, 1999; Olsson et al., 1998). Therefore, the extent to which confidence is a useful indicator of accuracy for these two types of lineups is also an issue that should be considered in selecting the superior procedure.

The negligible calibration difference between absolute and relative judgments and the consistency of the two types of complex judgment task, when the effects of task difficulty are accounted for, provides some preliminary evidence that suggests that CA calibration for sequential lineups may match that for simultaneous lineups. Of course, there are a number of important issues, which these experiments do not address, that mean that this suggestion is only advanced tentatively. Obviously, these studies have used a recognition memory paradigm that requires participants to decide only that they had seen the test stimulus previously and not to determine the precise context in which it had been seen. Further, the potential impact of the number of decision options (i.e., lineup size) on calibration has not been addressed here. It could be the case that, as a decision becomes more removed from the pure absolute or relative judgment task, the nature of the CA calibration changes. As we used only four stimuli in our complex judgment tasks, we cannot rule out changes in calibration as a result of a change in the number of test stimuli. Perhaps more importantly, we cannot rule out an interaction between the type of complex decision task (i.e., favoring absolute or relative judgments) and the number of stimuli. Another obvious difference between this paradigm and the identification paradigm is that these experiments used the same stimuli at study and test, thus requiring participants to make picture recognition judgments. Therefore, an important avenue for future research is to replicate these findings for a person recognition task, that is, one where different photographs of the same individual are used at study and test. Nevertheless, the similarity between the calibration observed for positive recognition decisions in these experiments and the results obtained in eyewitness identification calibration studies (Brewer et al., 2002; Brewer & Wells, 2004; Juslin et al., 1996; Olsson, 2000; Olsson & Juslin, 1999; Olsson et al., 1998) suggests that these results provide a potentially useful guide as to the likely results of more ecologically valid experiments.

The identification of an effect of difficulty on over/underconfidence also has important ramifications for the eyewitness identification domain. First, if this effect cannot be controlled then the practical use of confidence may be limited to an indicator of the relative likelihood of accuracy of multiple identification decisions from the same lineup. However, Brewer et al. (2002) have already demonstrated that experimental manipulations that target over/underconfidence can improve calibration. Specifically, participants were asked to reflect on elements of the stimulus and lineup decision that are thought to influence accuracy or to generate reasons why they may have made an incorrect decision about the lineup. Thus, these findings together suggest that an important area for future research in CA calibration in eyewitness identification is the development of techniques, such as Brewer et al.'s reflection and hypothesis disconfirmation manipulations, which nullify the impact of difficulty on calibration.

The lack of a hard–easy effect for positive recognition decisions in the absolute conditions of Experiments 1 and 2 suggests a new lineup procedure that has the potential to produce well-calibrated confidence judgments. Specifically, if the sequential lineup procedure was modified so that a confidence estimate was elicited for every judgment (i.e., for the witness's decision about each lineup member), rather than a single estimate about their ultimate lineup decision, and these findings generalized to the identification context, then this confidence estimate would provide a reasonable estimate of the probability of a positive identification decision being correct. The development of a procedure that produces well-calibrated confidence judgments that are insensitive to task difficulty would be of enormous practical import. If police officers and courts had knowledge of the likely accuracy of an identification then, especially in cases where no other evidence was available, the influence of erroneous identifications could be reduced while maintaining the impact of accurate identifications, thereby, allowing the conviction of more guilty but fewer innocent individuals.

In addition, the obvious impact of difficulty on the CA relationship should serve as an important methodological warning to eyewitness identification researchers. Most studies in this domain tend to make use of a very small stimulus set, often only a single study video or live event and the related target-absent and target-present lineups. Consequently, the results of any given study, regardless of its rigor, could be influenced significantly by the difficulty of the stimuli chosen, and markedly different patterns could emerge with different stimuli. These findings serve as a useful reminder of the need to approach the investigation of eyewitness identification within a theoretical framework to avoid the necessity of replicating every result at all possible difficulty levels with every conceivable difficulty manipulation.

Finally, the calibration difference between positive and negative recognition decisions highlights an important issue for eyewitness identification investigators. Obviously, if the accuracy of negative recognition decisions cannot be reliably distinguished by confidence judgments, investigators must be careful to ignore the influence of a confidence judgment when considering the usefulness of a lineup rejection as evidence. It is important to note that this is not to say that a negative decision by a witness is not useful evidence for investigators. Indeed, Wells and Olsson (2002) dem-

onstrated the utility of lineup rejections for law enforcers. Rather, it suggests that the confidence with which the witness makes the rejection may not be a useful indicator of the accuracy of the witness's decision. We should note here, however, that one of the potential explanations for this positive–negative difference (i.e., the lack of a specific item in memory with which to compare stimuli judged as new), discussed in detail below, highlights a potentially important distinction between the recognition memory paradigm and the eyewitness identification task. In the latter, witnesses are required to decide not only whether they have seen a lineup member before but, more specifically, whether a lineup member was the perpetrator of the witnessed event. Therefore, regardless of the type of decision made by the witness (i.e., either a positive identification or a rejection of the lineup, i.e., a *not present* decision), the lineup member or members being considered can still be compared with a specific item in memory, the witness's memory of the appearance of the perpetrator. Consequently, confidence judgments for both types of decision could, in principle, be based on this type of specific comparison. These possibilities confirm that a useful avenue of future investigation is the source of this positive–negative calibration difference in face recognition memory.

In addition to these applied implications these data highlight a number of interesting theoretical issues. First, only one other study (Weber & Brewer, 2003) has reported a difference in calibration between positive and negative decisions. They too found that confidence in negative decisions, but not positive decisions, was relatively insensitive to accuracy. This finding highlights an issue yet to be addressed by calibration research. The majority of calibration studies have used relative judgment tasks, for example multiple-choice general knowledge questions or perceptual discrimination tasks, in which all responses are positive. Therefore, the identification of the cognitive mechanisms underlying the calibration difference between positive and negative recognition decisions and the extent to which it generalizes from face recognition to other tasks is an interesting area for future work.

Second, the finding that difficulty influences O/U for negative but not for positive recognition decisions has important ramifications for the debate over the source of the hard–easy effect. Some researchers (e.g., Gigerenzer et al., 1991) hold that the hard–easy effect is the result of the cognitive processes underlying decision making and the formation of confidence judgments. Others (e.g., Juslin et al., 2000) argue that the effect is, at least in part, due to artifacts such as scale-end effects, regression to the mean, and the linear dependence between difficulty and O/U. It is important to note that our findings are not consistent with the latter viewpoint as these artifacts should influence both positive and negative decisions equivalently. Our findings do not rule out the impact of these artifacts, as the pattern of O/U for positive decisions could be the result of the cancelation of the influence of the artifacts and the opposing influence of the cognitive mechanism responsible for confidence judgments about positive decisions. However, the finding does raise an interesting question about the nature of the hard–easy effect and the cognitive processes responsible for confidence judgments about positive and negative decisions.

In sum, this article has provided the first direct evidence that, when task difficulty is accounted for, little difference in calibration exists between absolute and relative judgments and also between more complex judgment tasks that differ in the extent to which they allow absolute and relative processing. Moreover, these experiments have provided further evidence for the existence of a hard–easy effect and the first demonstration of the equivalent impact of difficulty on CA calibration for absolute and relative recognition memory decisions. Finally, the identification of a calibration difference between positive and negative recognition decisions has highlighted an issue that is of importance not only to eyewitness identification researchers but also to the broader recognition memory and calibration domains.

References

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and Psychophysics*, *55*, 412–428.

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

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

Brewer, N., & Wells, G. L. (2004). *The confidence–accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity and target-absent base rates*. Manuscript in preparation.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Deffenbacher, K. A., Leu, J. R., & Brown, E. L. (1981). Memory for faces: Testing method, encoding strategy, and confidence. *American Journal of Psychology*, *94*, 13–26.

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

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.

Juslin, P., Winman, A., & Olsson, H. (2000). Naïve empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review*, *107*, 384–396.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370.

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior*, *24*, 685–697.

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science*, *9*, 215–218.

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, *70*, 556–564.

Martinez, A. M., & Benavente, R. (1998). *The AR Face Database* (CVC Tech. Rep. No. 24). Barcelona, Spain: Universitat Autónoma de Barcelona, Computer Vision Center.

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. (1999). Can self-reported encoding strategy and recognition skill be diagnostic of performance in eyewitness identifications? *Journal of Applied Psychology*, *84*, 42–49.

Olsson, N., Juslin, P., & Winman, A. (1998). Realism of confidence in

earwitness versus eyewitness identification. *Journal of Experimental Psychology: Applied, 4,* 101–118.

Read, J. D., Lindsay, D. S., & Nicholls, T. (1998). The relation between confidence and accuracy in eyewitness identification studies: Is the conclusion changing? In C. P. Thompson, D. J. Herman, J. D. Read, D. Bruce, D. G. Payne, & M. P. Toglia (Eds.), *Eyewitness memory: Theoretical and applied aspects* (pp. 107–130). Hillsdale, NJ: Erlbaum.

Sporer, S. L., Penrod, S. D., 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.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25,* 459–473.

University of Stirling. (2001). *Psychological Image Collection.* Retrieved March 2001, from /pics.psych.stir.ac.uk/

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. (2004). *Positive versus negative face recognition decisions: Confidence, accuracy, and response latency*. Manuscript in preparation.

Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). New York: Cambridge University Press.

Wells, G. L., & Olsson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied, 8,* 155–167.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22,* 603–647.

Wilson, J. L., Scott, J. H., & Power, K. G. (1987). Developmental differences in the span of visual memory for pattern. *British Journal of Developmental Psychology, 5,* 249–255.

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

Received February 9, 2004
Revision received March 31, 2004
Accepted April 7, 2004 ■