The Effect of Retention Interval on the Confidence–Accuracy Relationship for Eyewitness Identification

James Sauer · Neil Brewer · Tick Zweck · Nathan Weber

Published online: 22 July 2009
© American Psychology-Law Society/Division 41 of the American Psychological Association 2009

Abstract Recent research using a calibration approach indicates that eyewitness confidence assessments obtained immediately after a positive identification decision provide a useful guide as to the likely accuracy of the identification. This study extended research on the boundary conditions of the confidence–accuracy (CA) relationship by varying the retention interval between encoding and identification test. Participants (N = 1,063) viewed one of five different targets in a community setting and attempted an identification from an 8-person target-present or -absent lineup either immediately or several weeks later. Compared to the immediate condition, the delay condition produced greater overconfidence and lower diagnosticity. However, for choosers at both retention intervals there was a meaningful CA relationship and diagnosticity was much stronger at high than low confidence levels.

Keywords Eyewitness identification · Confidence–accuracy · Retention interval · Calibration

Criminal justice systems often use eyewitness identification evidence when assessing the likely guilt of a suspect or defendant. Yet, the likelihood of eyewitness identification error is well documented by laboratory- and field-based research demonstrating that, when presented with a lineup and asked to make an identification decision, witnesses sometimes (a) misidentify innocent lineup members as the culprit or (b) fail to identify the culprit when s/he is present in the lineup (Cutler & Penrod, 1995; Innocence Project, 2009; Pike, Brace, & Kynan, 2002; Wells et al., 1998). Such identification errors divert investigative attention from the actual culprit and are likely to undermine the effectiveness of the criminal justice system. Their impact has motivated a substantial amount of research aimed at identifying markers capable of discriminating accurate from inaccurate identification decisions.

Eyewitness confidence is one possible marker of identification accuracy that has been used by forensic decision makers. Not only has confidence been endorsed by the U.S. Supreme Court as one of the criteria to be considered when assessing the likely accuracy of identification evidence (Neil v. Biggers, 1972) but there is also a substantial literature demonstrating that eyewitness confidence influences assessments of likely identification accuracy made by police officers, lawyers, jurors, and jury-eligible samples (e.g., Bradfield & Wells, 2000; Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988; Deffenbacher & Loftus, 1982; Lindsay, Wells, & Rumpel, 1981).

Moreover, there are sound theoretical grounds for predicting a meaningful confidence–accuracy (CA) relationship for eyewitness identification decisions, which are a form of recognition memory decision. A number of theories of decision making and confidence processing—such as signal detection theory (Egan, 1958; Green & Swets, 1966; Macmillan & Creelman, 1991) and accumulator models of decision making and perceptual discrimination (Van Zandt, 2000; Vickers, 1979)—suggest a shared evidential basis for response and response confidence in recognition memory tasks. Both classes of theory hold that confidence stems from the same evidence that drives the decision-making
process and, consequently, conditions facilitating accurate responding (e.g., long exposure durations, focused attention, short retention intervals) should also produce high confidence. Conversely, conditions that hinder accurate responding should also lead to decreased confidence.

Although there have been repeated demonstrations of weak or, at best, modest, CA correlations (e.g., Bothwell, Deffenbacher, & Brigham, 1987; Sporer, Penrod, Read, & Cutler, 1995; empirical support for the diagnostic utility of eyewitness identification confidence—under certain conditions—has grown (e.g., Brewer & Wells, 2006; Juslin, Olsson, & Winman, 1996; Lindsay, Nilsen, & Read, 2000; Lindsay, Read, & Sharma, 1998; Sauer, Brewer, & Wells, 2008; Sauerland & Sporer, 2009; Weber & Brewer, 2004). Continued research interest in the CA relationship has been stimulated by two lines of enquiry suggesting that the early correlational work underestimated the CA relationship. First, Lindsay et al. (2000, 1998) argued that the homogeneity of encoding and testing conditions (e.g., exposure duration, witnesses’ attention to the target stimulus, retention interval, etc.) evident in most correlational investigations of the CA relationship for eyewitness identification tasks restricts variation in the quality of participants’ memories for the target. Thus, variations in accuracy and confidence are constrained, and the CA relationship underestimated. Lindsay et al. demonstrated substantial CA correlations across participants making a positive identification when witnessing conditions were varied to produce changes in the quality of the witness’ memory for the target.

Second, Juslin et al. (1996) argued that the point-biserial correlation provides only a limited perspective on the CA relationship, whereas an alternative approach—calibration—provides (a) a more detailed representation of the CA relationship and (b) more forensically useful information. The calibration approach compares the objective and subjective probabilities of a response being correct, determining the proportion of correct responses at each confidence level (typically measured on 0–100% scale). Perfect calibration is obtained when, for example, 100% of all responses made with 100% confidence are accurate, 90% of all responses made with 90% confidence are accurate, etc. This information is typically plotted on a graph, with the resulting calibration function compared to the ideal function, to assess the CA relationship. In addition to visual inspection of the curve, the calibration approach incorporates a number of statistical tools for assessing the CA relation. First, the calibration (C) statistic indexes the degree of correspondence between the subjective assessment (i.e., confidence) and the objective probability (i.e., accuracy) of correct recognition, and varies from 0 (perfect calibration) to 1. To calculate the C statistic, the difference between proportion correct and confidence level is computed, and squared, for each confidence level. These values, each multiplied by the number of judgments at the respective confidence level, are then summed and divided by the total number of judgments in the sample. Second, the computation of an over/under-confidence (O/U) statistic indicates the extent to which participants are, generally, more or less confident than they are accurate. The O/U statistic is calculated by subtracting the mean accuracy from the mean confidence of the sample. The O/U statistic can range from −1 to 1, with negative and positive scores indicating underconfidence and overconfidence, respectively. Finally, resolution (like the CA correlation) assesses the extent to which confidence discriminates correct from incorrect decisions. The Normalized Resolution Index (NRI) ranges from 0 (no discrimination) to 1 (perfect discrimination). The forensic utility of the calibration approach, when compared to correlation, lies in its indication of probable accuracy for each level of confidence. As Juslin et al. (1996) note, the knowledge that the CA correlation is, for example, .28 does not help assess the accuracy of an individual identification made with 80% confidence. On the other hand, knowing that 80% (or 70, or 90%) of identifications made with 80% confidence are correct provides a guide for assessing the likely reliability of an individual identification decision.

Studies using the calibration approach have not only provided detailed information on the CA relationship but, in so doing, have also demonstrated robust CA relationships when participants positively identify a lineup member as the culprit (e.g., Brewer & Wells, 2006; Juslin et al., 1996; Sauerland & Sporer, 2009), provided confidence is assessed immediately after the identification is made (Bradfield, Wells, & Olson, 2002; Brewer, Weber, & Semmler, 2007). The reason for the poor CA relations typically observed for non-choosers remains unclear. However, it is well understood why delaying the assessment of confidence is harmful to the CA relation. As outlined above, the relationship between memory quality, confidence, and accuracy is fundamental to the CA relationship. However, confidence can be shaped not only by memory quality but also by various social, environmental, and meta-cognitive influences (see Wells, 1993). As the influence of these non-memorial factors increases, the degree to which confidence reflects the evidential basis it shares with accuracy decreases and, in turn, the CA relation weakens.

Research testing the boundary conditions for CA calibration is under way. The difference in CA relations for choosers and non-choosers, and the deleterious effects of delaying assessments of confidence on the CA relationship, are well documented. Brewer and Wells (2006) examined the effects on CA calibration of varying instructional bias, foil similarity, and target-absent base rates, while Weber and Brewer (2003) tested the effect of varying the
confidence scale on CA calibration in basic face recognition tasks. The present study further probes the boundary conditions for CA calibration by investigating the effects of varying the retention interval between encoding and the identification test on the CA relationship.

Retrieval interval is a variable of particular interest for three main reasons. First, witnesses to actual crimes commonly experience delays ranging from hours to months between viewing an event and being asked to make an identification decision. For example, Pike et al. (2002) report UK survey data revealing a median delay of over 10 weeks between police requesting and administering a lineup, although they noted that more conservative estimates put the average delay at just over a month. Regardless, it seems safe to assume that the average retention interval for the witness (i.e., between viewing the crime and viewing the lineup) is longer. In contrast, retention intervals employed to date in laboratory-based investigations of CA calibration for eyewitness (e.g., 15 min in Brewer & Wells, 2006) and face recognition memory tasks (e.g., 3–10 min in Weber & Brewer, 2003, 2004, 2006) are considerably shorter and less varied in range. Juslin et al.’s (1996) CA calibration study provides an exception by employing 1 h and 1 week retention intervals, and their findings are addressed below. The emphasis placed on confidence by decision makers in the forensic setting makes understanding the effect of lengthened delays on the efficacy of confidence in discriminating accurate from inaccurate identification decisions a matter of forensic significance.

Second, theories of recognition and recall memory function suggest that, in general, the quantity, quality, and/or accessibility of information stored in memory decreases over time. This claim is supported by a large body of research literature demonstrating that, across a variety of memory task paradigms, increases in retention interval generally produce decreases in recognition and recall memory performance (e.g., Deffenbacher, Bornstein, McGorty, & Penrod, 2008; Ebbinghaus, 1964; Schacter, 1999). Thus, variations in retention interval should produce variations in accuracy. While memory strength is the proposed basis for both confidence and accuracy (e.g., Egan, 1958; Green & Swets, 1966; Macmillan & Creelman, 1991) and, hence, variations in memory strength should affect both components of the CA relationship, it is unclear whether the effects on confidence and accuracy will be equivalent. Previous research demonstrates that changes in accuracy are not always accompanied by equivalent changes in confidence (e.g., Weber & Brewer, 2004) and, further, that various manipulations can influence confidence, and the CA association, independent of effects on accuracy (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000). Investigations of the CA relation for eyewitness recall memory suggest that repeated questioning produces confidence inflation (Shaw, 1996; Shaw & McClure, 1996). For recognition memory, providing post-identification feedback, encouraging witnesses to reflect on whether their encoding conditions were likely to facilitate or hinder identification accuracy, and having witnesses consider their behavior during the identification process all produce variations in the CA relation, without affecting accuracy (e.g., Bradfield et al., 2002; Brewer, Keast, & Rishworth, 2002; Kassin, 1985; Kassin, Rigby, & Castillo, 1991). In sum, it is clear that despite the strong theoretical links between confidence, accuracy, and memory strength, non-memorial factors can lead to CA dissociation. Thus, while the effect of increased retention interval on memory strength (and accuracy) is predictable, the effect of increased retention interval on CA calibration is not.

Third, while numerous studies have investigated the effect of varied retention interval on recognition and recall memory accuracy (see Deffenbacher et al., 2008 for a review), studies probing the effect of varied retention interval on the CA relationship are scarce. Lindsay et al. (1998) varied retention interval, but it was manipulated in conjunction with a number of other variables in an effort to exert a compounded effect on memory quality. Further, they assessed the CA relation using correlation and, thus, their findings do not allow specific predictions regarding CA calibration (see also Lindsay et al., 1981). As mentioned above, Juslin et al. varied retention interval and found no difference in CA calibration for identifications made after retention intervals of 1 h and 1 week. However, Juslin et al.’s investigation is limited in two important ways. First, their manipulation of retention interval exerted a negligible effect on accuracy (correct identification rates were .69 and .64 for the 1 h and 1 week conditions, respectively). Thus, there is no evidence that participants’ memories were challenged by the additional delay, and these findings are unable to speak to the effect of delay-induced memory degradation on CA calibration. Juslin et al. presented an additional CA calibration curve, based on a different dataset from that described in the article, which (a) combined data from a 1 week and 3 month retention interval condition and (b) suggested a meaningful CA relation in the upper half of the confidence scale. However, for three reasons, this curve is not informative regarding the effect of retention interval. First, the experimental methodology and data underpinning this curve remain (to our knowledge) unpublished. Second, the absence of any accuracy data precludes an assessment of any decline in memory associated with the increased retention interval. Third, derivation of a calibration curve given such a small sample required collapsing data across retention interval conditions, and no indication was given of the relevant contribution of data from each retention interval.
interval. Thus, we have no way of knowing to what extent this curve reflects the influence of either the shorter or longer of the two retention intervals. Simply put, Juslin et al.’s initial manipulation of retention interval was not strong enough to affect memory quality, and the introduction of the additional data did not overcome this limitation.

Second, after presenting a lineup but prior to making an identification decision, Juslin et al. (1996) had participants rate their confidence that any lineup member was presented at encoding. Brewer et al. (2002) found that having participants consider encoding conditions prior to rating confidence improved CA calibration. In a similar way Juslin et al.’s initial rating task may have aided calibration. For example, if a participant rates the likelihood that a lineup member was present at encoding as high, (s)he is likely to pick and do so with high confidence. Alternatively, if (s)he rates this likelihood as low but still chooses, confidence (and accuracy) is likely to be low. This pre-decision rating task may have improved CA calibration. Further, other research suggests that encouraging witnesses to consider confidence prior to making an identification can alter the decision making process and decision accuracy to consider confidence prior to making an identification can alter the decision making process and decision accuracy (e.g., Fleet, Brigham, & Bothwell, 1987).

In addition to these two major limitations, two idiosyncrasies in Juslin et al.’s (1996) methodology may have affected the CA relation observed. First, Juslin et al. used a target-absent base rate of .25, rather than the .50 base rate typical of eyewitness CA calibration research (and used in this research). While there is no reason to assume a .50 target-absent base rate in the applied setting (with the typically used .50 target-absent base rate perhaps representing a considerable overestimation), differences in the target-absent base rate affect CA calibration (Brewer & Wells, 2006). Second, the researchers provided instructions on calibration and interpretation of the confidence scale. Prior to eliciting confidence estimates, Juslin et al. informed participants that a positive identification accompanied by a confidence estimate of 0% amounted to a contradiction. While this logic may be sound, positive identifications are sometimes made with very low (even 0%) confidence, and this instruction may have influenced participants’ confidence estimates and, consequently, the CA relationship observed. Taken together, these differences are sufficient to raise doubts about the generalizability of Juslin et al.’s findings. Specifically, given that accuracy was barely affected by the manipulation, and that the rating task and lower target-absent base rate may have enhanced calibration and reduced underconfidence (cf. Brewer et al., 2002; Brewer & Wells, 2006), Juslin et al.’s (1996) study does not represent an adequate test of the effect of increased retention interval on CA calibration. CA calibration in Juslin et al.’s shorter retention interval condition was already strong. Thus, any over-estimation of the CA relation resulting from Juslin et al.’s methodology would most likely also manifest in the longer retention interval, increasing the likelihood of obtaining similar CA relations across conditions.

CA calibration research in the eyewitness identification area is in its infancy. The paucity of research in this area is understandable given the large number of participants required to generate stable estimates of CA calibration. Indeed, most of what is currently understood in this area relies on laboratory research using a limited range of stimulus materials. Only one study has previously examined CA calibration using a field study methodology (Sauerland & Sporer, 2009). The present research advances understanding of the CA relationship in three main ways. First, we used the CA calibration approach to examine the effect of retention interval on the CA relation, contrasting the CA relation for a virtually immediate identification test with that for one conducted between 3 and 7 weeks after the encoding event (and producing lower identification accuracy). Second, we used five different sets of encoding and test stimuli and, third, we tested the robustness of the CA relation in a field setting that provided varied and more realistic encoding conditions (cf. Lindsay et al., 1998).

METHOD

Design

A 2 (retention interval: immediate test versus delayed test) \(\times\) 2 (target-presence: present versus absent), between-subjects design was used to test the effect of varied retention interval on the confidence–accuracy relationship using multiple target stimuli in a field setting.

Participants

A total of 1,063 (548 female) participants provided data for this research. Participant ages ranged from 15 to 85 (\(M = 29.21, SD = 14.33\)). A functional grasp of the English language was the only prerequisite for participation.

Materials

Photographs of the target were cropped to present the individual, from the shoulders up, against a plain white/gray background, and were approximately 55 mm \(\times\) 55 mm in size. Non-target (i.e., foil) photographs were selected from our laboratory’s large database using a match-description strategy, with foil selection requiring agreement between the researchers and the experimenter from each pair that the foils matched the target’s
description. In sum, five different sets of target-present and target-absent lineups were constructed. For each target, identical foils were used for target-present and -absent lineups. Target-absent lineups were created by replacing the target with another foil photograph. However, as discussed in the “Results” section, because the designation of individual foils as target-replacements was arbitrary, the target-replacement is not analogous to an innocent suspect.

Procedure

Ten female, third-year honors psychology students collected data as part of a work experience course-component. The 10 students split into pairs with one acting as the researcher and the other as the target. Targets were of either Caucasian or Mediterranean appearance. Data were collected at various locations ranging from on-campus to city streets to parkland areas. While the target remained out of sight, the researcher approached members of the public (individually) and asked if they would like to participate in a psychology experiment. If the individual agreed, the researcher signaled to her partner who moved into the participant’s view, and remained in view for 10 s. Targets were viewed at a pre-measured distance of 10 m, and participants were instructed to attend to the target for the full 10 s.

After encoding, participants were allocated to either an immediate or delayed testing condition. Data were obtained from 691 participants in the immediate condition and from 372 participants in the delayed condition (i.e., only about 55% of participants approached in the delayed condition responded). Participants in the immediate condition were asked to perform an identification task. The researcher read the following instructions to the participant: “I’m now going to ask you to try and pick the person you just saw out of a group of photographs on this sheet…” The researcher then presented the participant with a laminated piece of A4 paper displaying eight, clearly numbered, color photographs organized into two rows of four faces. The instructions continued: “The person may or may not be present in the lineup. If you think the person is not present, please say ‘not present’. Please indicate the number of the person who is the person you have just viewed”. The researcher then recorded the participant’s response, asked the participant to indicate their confidence in the accuracy of their response on an 11-point scale (0–100%), and collected some demographic information.

Participants in the delayed condition provided an email address and were contacted approximately 18–21 days after encoding, and provided with a link to an online data collection system. Actual retention intervals ranged from 20 to 50 days ($M = 23$, $SD = 5$). When entered into the system, participant email addresses were matched to the relevant researcher/target pair to ensure that each participant viewed the correct lineup for their target stimulus. Participants accessed the online system and were presented with instructions generally identical to those reported above. However, rather than indicating responses verbally, participants in the delayed condition made identification decisions by either (a) clicking the photo of the lineup member they believed to be the target, or (b) clicking a button labeled “Not Present” at the bottom of the screen. Similarly, participants entered their confidence estimate by clicking one of eleven on-screen buttons representing the levels of confidence indicated above. Participants in the delayed condition were asked for the same demographic information as those in the immediate condition. Target-presentation was counterbalanced in both the immediate and delayed conditions to achieve an equal number of target-present and -absent trials.

RESULTS

Retention Interval and Accuracy

Chi-square analyses performed on response accuracy for the delayed and immediate conditions found predictable effects of retention interval for both choosers, $\chi^2(1, N = 614) = 11.59, p < .001$, and non-choosers, $\chi^2(1, N = 449) = 13.85, p < .001$. In both cases, accuracy was greater in the immediate condition (62 and 82% for choosers and non-choosers, respectively) than in the delayed condition (47 and 66% for choosers and non-choosers, respectively). Thus, the effect of increased retention interval on identification accuracy was consistent with the expected reduction in memory quality. As found by Juslin et al. (1996) and Sauerland & Sporer (2009), accuracy rates for non-choosers were significantly higher than for choosers in both the immediate, $\chi^2(1, N = 691) = 32.24, p < .001$, and delayed conditions, $\chi^2(1, N = 372) = 13.24, p < .001$. The present non-chooser accuracy and diagnosticity data (see below) lend support to previous research demonstrating that lineup rejections can inform assessments of the likely guilt of a suspect (e.g., Clark, Howell, & Davey, 2008; Wells & Olson, 2002).

Retention Interval and the CA Relation

To enhance the stability of the plotted CA calibration functions, confidence data were collapsed from the 11 initial confidence categories (i.e., 0–100%) to five (i.e., 0–20%, 20–40%, 40–60%, 60–80%, 80–100%) (see Brewer & Wells, 2006; Juslin et al., 1996). Moreover, because foils are known in advance to be innocent, we excluded target-
present, foil identifications from our calibration analyses (see Brewer & Wells, 2006). However, as there was no actual police suspect in the target-absent lineups, all false identifications of foils from target-absent lineups were included in calibration analyses, a practice that necessarily inflates the degree of overconfidence. Table 1 presents the distributions of confidence ratings for choosers and non-choosers, in the immediate and delayed conditions, according to identification response.

Given the well-documented differences in the CA relation for choosers and non-choosers, we present CA calibration analyses separately for these two groups (see Tables 1 and 2, and Fig. 1). In both retention interval conditions, meaningful CA relationships for choosers are apparent. Visual inspection of choosers’ CA calibration functions (Fig. 1) shows increasing accuracy as confidence increases for both retention intervals. Moreover, in the upper section of the confidence scale, the immediate and delayed condition curves are almost identical. While reliance on visual inspection may appear to lack rigor, the standard error bars for each confidence interval permit an estimation of the stability of the results obtained. Overlapping standard error bars (evident for the two highest confidence intervals of the chooser curves) denote non-reliable differences between groups.

Table 1 presents the diagnosticity ratios for each confidence category. Diagnosticity ratios indicate the likely reliability of an identification decision, in this case, according to the level of confidence expressed. Chooser diagnosticity ratios compare the likelihood that a guilty suspect will be identified to the likelihood that an innocent suspect will be identified. The procedure for separating suspect from foil identifications from target-absent lineups is complex. In contrast to the forensic setting, the laboratory setting provides no basis for designating any particular member of a target-absent lineup as the suspect (cf. Brewer & Wells, 2006). Accordingly, we calculated target-absent suspect identification rates by dividing the total number of target-absent false identifications by the number of lineup members (i.e., eight). Non-chooser diagnosticity ratios compare the probability that the witness responds not-present, given the target is not-present, to the probability that the witness responds not-present, given the target is present. Both retention interval conditions show increased diagnosticity at each successive confidence interval. Thus, when a suspect is identified, an increase in witness confidence is accompanied by an increase in the probability that the identified suspect is guilty.

There are, however, some differences apparent between the two retention interval conditions for choosers. A modified jackknife procedure (Koriat, Lichtenstein, & Fischhoff, 1980; Mosteller & Tukey, 1968) was performed on the C, O/U, and NRI statistics obtained for each retention interval condition. The jackknife procedure involves repeated calculation of each of the three statistics above, with each calculation omitting data from a different, individual participant. As many calculations are run as there are participants. This permits the calculation of mean and standard error data (Table 2) for the statistics obtained which, in turn, allows an assessment of differences in the relevant measures between groups. While these jackknife mean and standard error data cannot be subjected to inferential testing, they are intended to allow researchers to

### Table 1 Diagnosticity ratios and number of responses (according to response type) for each confidence interval, for choosers and non-choosers in the immediate and delayed testing conditions

<table>
<thead>
<tr>
<th>Condition &amp; response</th>
<th>Confidence level (%)</th>
<th>Immediate—choosers</th>
<th>Delayed—choosers</th>
<th>Immediate—non-choosers</th>
<th>Delayed—non-choosers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0–20</td>
<td>30–40</td>
<td>50–60</td>
<td>70–80</td>
<td>90–100</td>
</tr>
<tr>
<td>Correct identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foil identification</td>
<td>5</td>
<td>12</td>
<td>40</td>
<td>103</td>
<td>90</td>
</tr>
<tr>
<td>False identification</td>
<td>1</td>
<td>6</td>
<td>18</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Overall</td>
<td>15</td>
<td>31</td>
<td>89</td>
<td>158</td>
<td>112</td>
</tr>
<tr>
<td>Diagnosticity ratio</td>
<td>6.68</td>
<td>8.87</td>
<td>11.08</td>
<td>18.74</td>
<td>37.79</td>
</tr>
<tr>
<td>SE&lt;sub&gt;Diagnosticity&lt;/sub&gt;</td>
<td>2.71</td>
<td>2.40</td>
<td>1.91</td>
<td>2.56</td>
<td>8.61</td>
</tr>
<tr>
<td>Incorrect rejection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct rejection</td>
<td>6</td>
<td>11</td>
<td>42</td>
<td>91</td>
<td>84</td>
</tr>
<tr>
<td>Overall</td>
<td>10</td>
<td>13</td>
<td>52</td>
<td>113</td>
<td>98</td>
</tr>
<tr>
<td>Diagnosticity ratio</td>
<td>1.01</td>
<td>4.68</td>
<td>3.91</td>
<td>4.23</td>
<td>6.35</td>
</tr>
<tr>
<td>SE&lt;sub&gt;Diagnosticity&lt;/sub&gt;</td>
<td>0.57</td>
<td>5.91</td>
<td>1.31</td>
<td>0.89</td>
<td>1.71</td>
</tr>
</tbody>
</table>

1 Including only target-replacement identifications as false identifications from target-absent lineups resulted in only 13 and 6% (in the immediate and delayed conditions, respectively) of all target-absent misidentifications being available for calibration analyses. Split over the five confidence intervals, these data is insufficient to provide stable estimates of calibration.
draw inferences in conditions where data violate assumptions of conventional inferential testing techniques (Sheskin, 2004). Because the jackknife means replicated the original values in every case, only the original values are reported.

Inspection of the calibration functions, together with the O/U statistics (Table 2), suggests greater overconfidence for the delayed compared to the immediate condition. However, two aspects of the calibration information justify qualification of this general observation. First, for the two highest confidence categories, the standard error bars for the two functions overlap suggesting no meaningful difference in over/underconfidence. The applied value of this similarity at the higher confidence intervals is addressed in the “Discussion”. Second, the overall difference in overconfidence between conditions is, in fact, exaggerated by underconfidence in the lower half of the calibration curve in the immediate condition. This produces differences between conditions in three measures of the CA relation: the visual appearance of calibration function, the O/U statistic, and the NRI statistic. First, the calibration curve for the immediate condition flattens out in the lower half of the confidence scale, rather than following the ideal function. Further, the curve for the immediate condition shifts from overconfidence in the higher confidence intervals to underconfidence in the lower confidence intervals, a pattern not observed in the delayed condition. In addition to producing a visual flattening of the curve, this transition from overconfidence to underconfidence has important effects on two of the statistical measures of the CA relationship. It drives the immediate condition O/U statistic toward its mid-point (i.e., 0). Consequently, although the immediate condition curve exhibits noticeable underconfidence and overconfidence at the relevant extremes of the confidence scale, this is not reflected in the O/U statistic for that condition, thereby exaggerating the apparent difference in overconfidence between conditions. Finally, as evidenced by the NRI statistics (Table 2), it reduces the overall level of discrimination provided by confidence in the immediate condition. This discrepancy between conditions at the lower confidence extremes is addressed further in the “Discussion”.

The second difference between the CA relations for choosers in the delayed and immediate conditions is evident in the diagnosticity ratios reported for each confidence interval (Table 1). Consistent with the reported overall drop in identification accuracy associated with the delayed condition, the degree of diagnosticity at each confidence interval is greater in the immediate than delayed condition. Further, although no difference in overconfidence is apparent between conditions in the higher confidence brackets, the difference in diagnosticity persists. Nonetheless, as outlined above, the finding of increased diagnosticity with increased confidence is consistent (for choosers) across conditions.

In sum, the CA relations observed for choosers in the two retention interval conditions differ in terms of the

![Fig. 1](image-url)
general overconfidence and discriminability, due primarily to the trend toward underconfidence at low confidence levels in the immediate condition. However, in the upper half of the confidence scale, the conditions produce highly similar calibration functions. For non-choosers, both retention interval conditions produced the typically weak CA relations observed in previous CA calibration research. Further, any variations in diagnosticity between confidence levels were small and unsystematic in both conditions. While this absence of resolution might normally be taken as an indication that a confident rejection should not be given any special status, this needs to be considered in the context of accuracy rates for rejections usually being high. Thus, from an applied perspective, provided the conditions are such that non-chooser accuracy is high (e.g., unbiased lineup instructions, good encoding conditions), it is important to note that a highly confident rejection is as good a guide to (in)accuracy as a confident ID. Importantly also, an unconfident rejection is also likely to be as accurate as a confident ID.

The CA correlation patterns are generally in line with previous research (e.g., Lindsay et al., 1998; Sporer et al., 1995). CA correlations of moderate strength were found for choosers in both the immediate ($r (405) = 0.32, p < .001$) and delayed conditions ($r (209) = 0.41, p < .001$). While these values lie toward the high end of typically reported CA correlations, the relationships are still only moderate in size. Consistent with previous research, correlations for non-choosers were weak and non-significant in both the immediate ($r (286) = .09, ns$) and delayed conditions ($r (163) = .06, ns$).

**DISCUSSION**

While the dominant perspective in eyewitness identification research has been that the CA relationship is, at best, a weak one, recent research—underpinned by theoretically motivated changes in design and analysis techniques—has demonstrated meaningful CA relationships when certain pre-conditions are met. The present study extends this research, providing an important test of the boundary conditions of the CA relation. Variation in retention interval is (a) theoretically linked to variation in memory quality (and, thus, confidence and accuracy), (b) typical in the forensic setting, and (c) atypical in psychological investigations of the CA relation. Further, the emphasis placed on confidence when assessing the reliability of identification evidence in the forensic setting makes the effect of varied retention interval on the CA relationship an issue of applied and theoretical relevance.

The most striking feature of our examination of the effect of retention interval on the CA relationship is the consistency of the findings across retention interval conditions. Consistent with previous calibration research in the eyewitness and face recognition paradigms (e.g., Brewer & Wells, 2006; Juslin et al., 1996; Sauerland & Sporer, 2009; Weber & Brewer, 2003, 2004, 2006), confidence and accuracy were meaningfully related for choosers in both the immediate and delayed conditions, particularly in the upper half of the confidence scale. Further, both conditions show systematic increases in diagnosticity with increased witness confidence. Compared to the immediate condition, the delayed condition demonstrated an increase in general overconfidence and a decrease in the absolute levels of diagnosticity. However, such differences are equally likely to occur when retention interval is held constant but target stimuli or instructional bias are varied (e.g., Brewer & Wells, 2006). Of primary importance is the finding that the fundamental nature of the CA relationship, as evidenced by the shape of the calibration functions and the systematic relationship between confidence and diagnosticity, did not vary meaningfully between conditions. As Bruck and Poole (2002) note, albeit in a different context, when assessing consistency across conditions, patterns of findings are often more informative than individual numbers.

While our conclusions may be similar to those of Juslin et al. (1996) in that CA calibration was still evident when the retention interval was extended, our findings add significantly to our understanding of the effect of retention interval on the CA relation. Whereas there was no evidence that Juslin et al.’s retention interval manipulation affected memory strength, our manipulation clearly affected recognition memory performance and yet evidence of CA calibration persisted. Moreover, CA calibration was evident at the longer retention interval in our study, despite the absence of several methodological features contained in Juslin et al.’s research that may have buttressed the CA calibration detected at their longer retention interval. This suggests that these idiosyncrasies were not sufficient to affect the CA association. Additionally, by providing data from a field setting using multiple sets of encoding and test materials, our study provides an important pointer to the likely generality of the above conclusions.

The improved diagnosticity in both retention interval conditions evident at the upper confidence levels has significant forensic implications. Highly confident identifications, when compared to those made with low confidence, are likely to have a greater impact on police investigations and jury decision making. For example, in the absence of other compelling evidence, police are more likely to proceed with a case given a highly confident identification than given an identification made with low confidence. Further, compared to an identification made with low confidence, an identification made with high confidence is likely to be more persuasive in the
courtroom, and thus exert a more pronounced effect on juror assessments of likely guilt. Thus, it is reassuring that the identification decisions likely to exert the greatest influence in criminal justice systems are those for which (a) diagnosticity is greatest and (b) there was no significant variation in CA relationship associated with increased retention interval. We emphasize here, of course, that we are talking only about relationships detected when confidence was measured and recorded immediately after the identification, and not when opportunities for influencing confidence judgments had occurred.

A potentially interesting difference between the CA relations obtained in the two conditions presents in the lower half of the confidence scale for the chooser curves. As previously outlined, while the immediate condition curve exhibited underconfidence in the lower confidence levels, the delayed condition curve maintained its resemblance to the ideal function (i.e., low confidence ratings were accompanied by equivalently poor identification performance). As noted earlier, confidence judgments may be shaped not only by memory strength but also by various non-memorial factors. It may be the case that, because the immediate condition provided virtually no time for the memory trace to degrade, very low confidence estimates in this condition reflected the influence of misleading metacognitive inferences. In contrast, the delayed condition allowed for significantly greater degradation in memory trace, and, consequently, a greater drop in identification accuracy than did the immediate condition. In the delayed condition, very low confidence was perhaps more likely to reflect poor memory quality and, consequently, predict very poor performance. Thus, in this condition, confidence and accuracy corresponded more closely at the lower confidence levels, and the overall level of confidence-based discrimination increased (as evidenced by the NRI statistics). The improved resolution associated with the longer retention interval in the present study supports claims made by Lindsay et al. (2000, 1998) that the CA relation (and, in particular, resolution) is likely to be most evident in conditions that produce greatest variability in witnesses’ memory strength. However, given the low number of data points for these confidence categories, any conclusions must be tentative. Moreover, from an applied perspective, the data clearly show that low confidence identifications are associated with low accuracy (regardless of the existence of over- or underconfidence).

We should note three features of this study that might possibly have influenced the pattern of results obtained. First, despite email reminders to participants in the delay condition, there was still significant attrition. If it turns out that those conscientious enough to respond were also more conscientious—and, importantly, effective—when determining confidence judgments, then it is conceivable that the strength of the CA relation is overestimated in our delay condition. However, we know of no evidence that could sustain an argument either way on this issue. Second, our retention interval manipulation was confounded with method of responding. Participants in the immediate test condition provided their responses during face-to-face interactions with the researcher, while delayed condition participants responded via computer. As previously noted, social influence can undermine the confidence–accuracy relationship. However, given (a) the similarity of CA relationships evident between conditions in this experiment and (b) the similarity in CA relationships between the immediate condition in this experiment and previous work using similar (i.e., relatively short though not immediate) retention intervals and non-face-to-face responding (Brewer & Wells, 2006; Weber & Brewer, 2003, 2004, 2006), there is no reason to believe that method of responding exerted a significant effect on the results obtained. Third, for ethical reasons the encoded event in our field study did not involve a crime. Whether this might influence the CA relationship is also not known, though there is no obvious reason why this variable should interact with retention interval. What we do know, of course, is that the most reliable determinant of variations in the degree of over/underconfidence is task difficulty (see Brewer, 2006; Weber & Brewer, 2004), with our various stimuli providing tasks of sufficient difficulty to produce over- rather than under-confidence and, predictably, greater overconfidence in the delay condition.

In sum, this research asked: Does an increase in retention interval undermine the meaningful CA relationships reported in recent research? These results suggest not, at least not for retention intervals in the range used here. For choosers in both the delayed and immediate conditions, increased confidence was associated with increased probable accuracy. While this finding is encouraging, one important caveat is required. Although retention interval did not affect the CA relationship observed, many factors capable of distorting the CA relation over time in the forensic setting (e.g., confirming feedback/interaction with co-witnesses, repeated post-event questioning) were not addressed in our approach. It would be premature to suggest that, in the forensic setting, confidence-based discrimination of accuracy will not ever vary with increased retention interval. Simply increasing retention does not, by itself, seem to dampen the CA relation, but increased retention intervals may be associated with increased exposure to other factors likely to affect the relationship between confidence and accuracy. Moreover, it should be noted that retention intervals long enough to reduce identification accuracy to chance levels (i.e., likely much longer than in this study) would constrain variation in accuracy, reducing the extent to which confidence can discriminate accurate from inaccurate identification decisions.
Acknowledgments This research was supported by grant DP0556876 from the Australian Research Council and a Flinders Research Grant. We are grateful to Monica Beshara, Megan Cant, Danielle Chant, Kelly Ferber, Suzana Freegard, Caitlin Hithcock, Michaela O’Keefe, Lucy Pillay, Carla Raphael, Nancy Whitaker, and Anneke Woods for their assistance with data collection.

REFERENCES


