

In press: *Journal of Experimental Psychology: Applied*

Running Head: EYEWITNESS IDENTIFICATION

The confidence-accuracy relationship in eyewitness identification: Effects of lineup
instructions, foil similarity and target-absent base rates

Neil Brewer

Flinders University

Gary L. Wells

Iowa State University

Correspondence: Neil Brewer, School of Psychology, Flinders University, GPO Box 2100,

Adelaide, South Australia 5001

Telephone: 61-8-8201 2725

Fax: 61-8-8201 3877

Email: neil.brewer@flinders.edu.au

Abstract

Discriminating accurate from mistaken eyewitness identifications is a major issue facing criminal justice systems. This study examined whether eyewitness confidence assists such decisions under a variety of conditions using a confidence-accuracy (CA) calibration approach. Participants ($N = 1200$) viewed a simulated crime and attempted two separate identifications from 8-person target-present or absent-lineups. Confidence and accuracy were calibrated for choosers (but not nonchoosers) for both targets under all conditions. Lower overconfidence was associated with higher diagnosticity, lower target-absent base rates and shorter identification latencies. Although researchers agree that courtroom expressions of confidence are uninformative, our findings indicate that confidence assessments obtained immediately after a positive identification can provide a useful guide for investigators about the likely accuracy of an identification.

The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity and target-absent base rates

People who witness a crime are often asked by police to examine a lineup or photoarray to see if they can identify the offender. A witness's response at the identification test can have important ramifications. A positive identification is likely to shape subsequent police investigations and, ultimately, juror judgments. A failure to identify the suspect (i.e., a lineup rejection) may lead police to revise their theories about the likely perpetrator or, alternatively, to question the reliability of the witness and look for supportive evidence from other sources. As it is often the case that (a) the witness's identification is the key piece of evidence against a suspect, and (b) there is no way of knowing for certain if the witness's identification was accurate, the potential impact of an eyewitness identification can be a cause for concern. This concern is exacerbated by the now well-documented evidence from laboratory studies, which clearly indicates the fallibility of eyewitnesses (cf. Cutler & Penrod, 1995), and actual cases in which individuals convicted of serious crimes on the basis of eyewitness identifications have subsequently been exonerated by forensic DNA evidence (Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998).

Not surprisingly, therefore, researchers have invested considerable effort in trying to identify variables which might serve as independent markers of identification accuracy (labeled *assessment variables* by Sporer, 1993) and hence inform police, jurors and judges about the likely accuracy of an eyewitness identification. One variable that has attracted considerable attention from researchers is the eyewitness's confidence in his or her identification response. Apart from the fact that many people find it intuitively plausible that confidence and accuracy of various judgments should be at least reasonably closely related, the interest in the eyewitness confidence-accuracy (CA) relation can be traced to several

sources. Eyewitness identification confidence has been regarded in legal circles as an important indicator of accuracy. For example, the U.S. Supreme Court has endorsed confidence as one of the five criteria for assessing identification accuracy (Neil v. Biggers, 1972). Similarly, police, prosecuting and defence attorneys, and jury-eligible samples also share the view that confidence is an important indicator of likely eyewitness accuracy (Deffenbacher & Loftus, 1982; Noon & Hollin, 1987; Potter & Brewer, 1999). In addition, mock-juror studies have found that confidence has a major influence on mock-jurors' assessments of witness credibility and verdicts (Bradfield & Wells, 2000; Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988; Lindsay, Wells, & O'Connor, 1989; Lindsay, Wells, & Rumpel, 1981; Wells, Lindsay, & Ferguson, 1979).

Eyewitness researchers, however, have generally regarded eyewitness confidence as at best a relatively weak, and not practically useful, indicator of identification accuracy (cf. Kassin, Tubb, Hosch, & Memon, 2001). Most reviews and meta-analyses locate the average CA correlations as somewhere around zero to .3 (e.g., Bothwell, Deffenbacher, & Brigham, 1987; Cutler, Penrod, & Martens, 1987; Sporer, Penrod, Read, & Cutler, 1995; Wells & Murray, 1984), with stronger (though still modest) relations given optimal encoding conditions (Deffenbacher, 1980) or positive identifications (Sporer et al., 1995).

Various arguments have been advanced to explain the absence of a strong CA relation for eyewitness identification. For example, demonstrations of overconfidence (albeit only under specific conditions) across different domains such as general knowledge (Fischhoff, Slovic, & Lichtenstein, 1977; Koriat, Lichtenstein, & Fischhoff, 1980) and eyewitness recall (Bornstein & Zickafoose, 1999; Granhag, Strömwall, & Allwood, 2000) have led researchers to question whether common processes underlie the two types of judgments or whether variables that affect accuracy and confidence necessarily have equivalent effects on each. For

example, Busey, Tunnicliff, Loftus, and Loftus (2000) detected greater observer confidence for faces shown brighter at test than at encoding although accuracy was higher when brightness at test matched that at encoding. This was attributed to observers' confidence judgments being influenced by analytic (or metacognitive) cues when there was a mismatch between encoding and test stimuli. In a similar vein, superior exposure conditions at the time of the crime may increase the likelihood of an accurate identification but (to the extent that the witness relied upon such metacognitive cues when assessing confidence) may produce an even more marked increase in witness confidence. Consistent with this suggestion, Memon, Hope, and Bull (2003) found improved identification accuracy (i.e., hits and correct rejections vs. misses and false alarms) with longer exposure to the offender, but greater confidence for correct relative to incorrect identifications was no longer evident for target-present lineups.

Other factors have also been implicated in the dissociation of confidence and accuracy variables. One factor is the tendency for people to seek out only confirmatory information about their judgments (confirmatory bias), thereby producing overconfidence (Koriat et al., 1980), though note that the Koriat et al. findings have sometimes been difficult to replicate (Allwood & Granhag, 1996; Fischhoff & MacGregor, 1982). A second factor, suggested by Tversky and Koehler's (1994) research on support theory, is that judgments of uncertainty (quantitative or qualitative) cannot be made reliably because of the inevitable failure to take into account the influence of unavoidable possibilities or scenarios that should guide these judgments. A third and related factor is the difficulty that individuals have in accurately translating their internal or subjective states of confidence onto a numerical scale (cf. Gigerenzer, Hoffrage, & Kleinbölting, 1991; Windschitl & Wells, 1996). A fourth is the possibility that social factors, such as feedback from lineup administrators or other witnesses,

may affect confidence judgments for identification responses, independent of any effect on accuracy (Luus & Wells, 1994; Semmler, Brewer, & Wells, 2004; Wells & Bradfield, 1998, 1999). Finally, another factor that questions the likely stability of the CA relationship is the possible contribution of individual difference variables. Research outside the eyewitness identification domain points to possible influences of variables such as gender and personality on the realism of people's assessments of their own abilities (e.g., Furnham & Thomas, 2004), though the findings are not always supportive (Jonsson & Allwood, 2003).

Although factors such as those outlined above may have influenced eyewitness researchers to discount confidence as an indicator of identification accuracy, the broader human judgment literature does not support the view that people are always overconfident (or underconfident) in their judgments. For example, Ackerman, Beier, and Bowen (2002) reported considerable agreement between (a) participants' self-assessments of fluid and crystallized abilities and a diverse array of knowledge measures and (b) objective measures of those abilities and knowledge domains. Not only did participants recognize their areas of strength, but they also discriminated those from their areas of weakness. Furthermore, it has been consistently demonstrated that the presence (and degree) of over- or underconfidence is dependent on judgment difficulty. For both perceptual (Baranski & Petrusic, 1994; Petrusic & Baranski, 1997) and non-perceptual tasks (Gigerenzer et al., 1991; Lichtenstein & Fischhoff, 1977; Lichtenstein, Fischhoff, & Phillips, 1982), increased difficulty is associated with increased overconfidence. It has been pointed out that difficulty-related overconfidence (or underconfidence) may arise from a failure to recognize just how difficult (or easy) a particular task may be (Lichtenstein et al., 1982).

There are also theoretical grounds for expecting a meaningful, even if not perfect, CA relationship. A number of theories of human judgment and decision making point to robust

CA links by virtue of common underlying processes. Signal detection theory, for example, holds that strength of evidence determines both accuracy and confidence (Green & Swets, 1966; Macmillan & Creelman, 1991). Some theories of perceptual discrimination maintain that confidence is directly related to the strength of evidence favoring one decision alternative over another (Van Zandt, 2000; Vickers, 1979). Evidence strongly favoring a particular alternative will be a condition for high accuracy and confidence, with weak evidence associated with just the opposite. Finally, some models of recognition memory distinguish recollection based processes, characterized by high accuracy and confidence, and familiarity based processes, the accuracy and confidence of which vary with memory trace strength (Atkinson & Juola, 1974; Mandler, 1980; Wixted & Stretch, 2005; Yonelinas, 2002). To the extent that eyewitness identifications rely on such processes, their contributions to both accuracy and confidence judgments suggest a meaningful CA relationship.

The eyewitness identification literature also provides specific grounds for expecting more robust CA relationships than are generally acknowledged in this area. Several researchers (Brewer, Keast, & Rishworth, 2002; Brewer, Weber, & Semmler, 2005; Juslin, Olsson, & Winman, 1996; Weber & Brewer, 2003, 2004) have argued that a richer perspective on the CA relationship in the eyewitness identification domain is obtained if, instead of relying on the (typically used) point-biserial correlation to examine the CA relationship, a procedure called calibration is used. Although the point-biserial correlation is informative about the variance in identification accuracy explained by confidence across participants, its characteristics are such that meaningful relationships between identification confidence and accuracy are not necessarily detected. The point biserial correlation expresses the relation between a categorical scale of confidence (0%, 10%, ... 100%) and a binary identification test outcome (correct, incorrect). A number of researchers (Juslin et al., 1996; Lindsay,

Nilsen, & Read, 2000; Lindsay, Read, & Sharma, 1998) have argued that (a) the distribution of confidence scores in identification studies will generally be unimodal due to the invariant encoding and identification test conditions, and (b) the resultant restricted variance in confidence scores will produce CA correlations that underestimate the relationship likely to obtain in the forensic context where much variation in the encoding and test conditions is anticipated. Nor can the point-biserial correlation inform judgments of the likely reliability of judgments made with a particular level of confidence, or assist in determining whether a witness's confidence estimate is a realistic estimate of the likelihood of a correct identification rather than a marked under- or overestimate.

Recent research has shown that the alternative procedure, calibration, can reveal positive CA relations that are not suggested by the relatively weak point-biserial correlations, although investigation of calibration in the identification context requires very large sample sizes in order to generate stable data (Brewer et al., 2002; Juslin et al., 1996). Examinations of CA calibration in the eyewitness identification context have plotted identification decision confidence against identification accuracy, with the proportion of correct identifications recorded for each confidence category. A linear CA function, characterized by 100% accuracy for those witnesses who were 100% confident, 90% accuracy for witnesses who were 90% confident and so on, is considered to signal perfect calibration. In the forensic context such data can provide guidance on the likely reliability of identifications made with different levels of confidence (e.g., 100% vs. 70%), information that in a particular case is arguably more informative than a point-biserial correlation coefficient.

Using the calibration procedure, there are several possible ways of assessing the CA relation. Visual inspection of the CA calibration function indicates the degree of match between the obtained and the ideal function. Three statistics (see Baranski & Petrusic, 1994,

and Lichtenstein et al., 1982) are also informative. One is the calibration or C statistic which varies from 0 to 1, with zero indicating perfect calibration. The calculation of C involves: (1) assigning each confidence rating to class intervals (J); (2) finding the difference between the mean confidence level (c_j) and the proportion of correct responses (a_j) for each class interval (j); (3) multiplying the squared differences by the number of observations (n_j) in the interval; (4) summing the result of step (3) across class intervals and dividing by the total number of observations (n): i.e.,

$$C = \frac{1}{n} \sum_{j=1}^J n_j (c_j - a_j)^2$$

A second is an over/underconfidence statistic (O/U) which can vary from -1 to $+1$: under- and overconfidence are indicated by negative and positive scores, respectively. In the identification context, over/underconfidence is used to refer to whether participants' identification confidence estimates were, on average, greater than or less than their accuracy. The O/U statistic is calculated in the same way as the C statistic, except that the differences are not squared. (This is equivalent to mean confidence minus overall proportion correct. Thus, confidence intervals for the O/U statistic, for either choosers or nonchoosers, can be calculated by assigning a proportion correct of 0 or 1.0 to each participant's incorrect or correct identification response.)

Normal parametric analyses may be applied to these calibration statistics when multiple observations exist for each participant, thereby permitting the computation of C and O/U statistics for each participant (see, for example, Cutler & Penrod, 1989, and Weber & Brewer, 2003, with a face recognition paradigm). In the eyewitness identification paradigm where one (or occasionally several) data point per participant is the norm, these conventional analyses are not possible. A third measure, resolution (Baranski & Petrusic, 1994; Yaniv, Yates, &

Smith, 1991), indicates how well confidence judgments discriminate correct from incorrect decisions, and is indexed by the normalized resolution index (NRI) which ranges from 0 (no discrimination) to 1 (perfect discrimination), with associated effect sizes able to be calculated. The NRI is calculated as:

$$\left[\frac{1}{n} \sum_{j=1}^J n_j (a_j - a)^2 \right] / a(1-a),$$

where a_j denotes proportion of correct responses at confidence level j and a denotes overall mean accuracy.

There is now calibration evidence available from face recognition and eyewitness identification paradigms which suggests that, contrary to the conclusion generally drawn from CA correlation studies, confidence is informative about accuracy. Using an old-new face recognition task, Olsson, Juslin, and Winman (1998) reported a generally linear calibration curve ($C = .003$), despite a modest CA correlation (.36). Also with a face recognition paradigm, Weber and Brewer's (2003) calibration curves showed a similar trend for the 50-100% section of the confidence scale for both absolute and relative judgments, with within-subjects' C statistics ranging from .03 to .11. These patterns also occurred in the context of modest CA correlations ranging from .21 to .35. Weber and Brewer (2004) found similar calibration patterns for choosers' absolute and relative judgments using the conventional face recognition paradigm and a 'mini-lineup' version of that paradigm. Using an eyewitness identification paradigm, Juslin et al. (1996) also reported a clear positive CA relationship (only slight under- or overconfidence, indicated by O/U statistics of -.06 and .06, and the 95% confidence intervals including zero) for lineups based either on match-to-description or suspect similarity and for two different culprits, although none of the associated correlations surpassed .49. In contrast, Brewer et al. (2002) presented calibration

curves that showed virtually no relation between confidence and accuracy in a control condition (most likely because confidence judgments were made some five minutes after the identification response). When, however, witnesses were subjected, prior to producing their confidence estimates, to one of two manipulations (reflection or hypothesis disconfirmation) designed to improve confidence scaling, the CA curves indicated clear positive relationships. Again, however, CA correlations were modest in all conditions and did not differ significantly from each other.

Although it is encouraging that identification confidence and accuracy can be calibrated even though the point-biserial correlation is low, and there has been research into situational influences on the CA correlation (e.g., Bothwell, Deffenbacher, & Brigham, 1987; Deffenbacher, 1980), to date there has been limited examination of the influence of situational factors likely to affect identification responding on CA calibration, over- and underconfidence, and resolution. Here we explored how CA calibration, and the associated statistics, were affected by forensically relevant situational variables designed to produce variance in identification accuracy and/or confidence. Specifically, we investigated the effects of lineup instructions (biased vs. unbiased) and lineup foil similarity (high vs. low), and in a post hoc examination the proportion of target-absent lineups presented (.15 vs. .25 vs. .5). We examined the impact of these variables on CA relationships for witnesses who made a positive identification response (i.e., choosers) and for those who rejected the lineup (i.e., nonchoosers). In addition, we contrasted CA relationships for participants who made their identification decision rapidly versus slowly, a characteristic of responding that reliably discriminates accurate from inaccurate identifications (Brewer, Caon, Todd, & Weber, in press). Given the (previously outlined) social influences on eyewitness confidence assessments, confidence assessments were taken immediately after the identification response

and were provided without any possibility of influence from the lineup administrator. An immediate confidence assessment may also reduce the impact of any decay of, or interference with, confidence cues as a result of a delay between the identification test and the confidence assessment. Even at short delays, the information for confidence judgments that is available via any of the mechanisms outlined previously (i.e., recollection, familiarity, or balance of evidence) is likely to differ from that available at the time of identification. Yet, at the metacognitive or analytical level (cf. Busey et al., 2000), witnesses are unlikely to acknowledge the possibility of a deterioration in the confidence information over relatively short intervals and, in turn, make the necessary adjustments to their confidence estimates.

An important influence on identification performance is the decision criterion adopted by witnesses. A variety of factors is likely to influence decision criteria: for example, witnesses' expectations about the likelihood that the offender is in the lineup and the instructions given at lineup presentation. How factors that shape decision criterion adjustments affect the CA relationship is unknown. Here we used the presence or absence of bias in lineup instructions (i.e., whether the instructions fail to caution vs. caution the witness that the culprit may not be in the lineup) to manipulate criterion adjustments and, in turn, the pattern of identification responses.¹ Unbiased instructions are typically associated with lower rates of choosing (i.e., fewer positive or false identification responses) from target-absent lineups (Malpass & Devine, 1981; Steblay, 1997), reflecting the adoption of a stricter criterion for a positive identification. Consistent with this suggestion, unbiased instructions not only produce fewer false identifications (cf. biased instructions) from target-absent lineups, but the evidence also indicates lower rates of choosing (i.e., positive identifications) from target-present lineups (Clark, 2005). This is reflected in fewer correct identifications, foil identifications, or both. Whether such adjustments in identification response patterns are accompanied by

commensurate adjustments in confidence so that the CA relation is maintained was examined here.

Variations in the degree of similarity between the target and lineup foils (e.g., operationalized as the number of people in the lineup who are a plausible match for the offender or the degree of similarity between the target and each of the foils) are also likely to affect identification performance. For example, given the well-documented tendency of witnesses to make relative judgments and to pick the best match to the offender (Wells, 1993), the expectation is that correct identifications from target-present lineups should be more prevalent when the target is accompanied by low rather than high similarity foils. At the same time, an innocent foil who happens to match the offender's description is at much higher risk of being positively identified if there are few (or no) plausible alternatives in a target-absent lineup. Data consistent with these expectations have been reported by Lindsay and Wells (1980) and Tredoux (2002), although such manipulations have not always affected identification response patterns (Gonzalez, Ellsworth, & Pembroke, 1993).

It is also possible that foil similarity manipulations could produce other outcomes by influencing witnesses' metacognitions about the identification test process. For example, the presence of a large number of plausible foils, and the ensuing feeling of familiarity (cf. Chandler, 1994), might suggest to witnesses that the offender must be present in the lineup and increase the likelihood of positive identification responses from both target-present target-absent lineups. Depending on the precise pattern of identification responses across target-present and -absent lineups, one consequence may well be elevated confidence that is not matched by accuracy gains (i.e., overconfidence). Interestingly, Nosworthy and Lindsay (1990) noted that increasing the nominal size of the lineup increased the number of plausible foils which, in turn, resulted in a higher frequency of positive identifications (i.e., choosing).

The precise pattern of identification responses detected will most likely be dependent upon the particular combination of encoding conditions, lineup structure, encoding-test stimulus match, and so on. Whether the fluctuations in accuracy and/or confidence produced by variation in lineup foil similarity affect the CA relationship was examined in this study.

The third issue investigated was the extent to which confidence and accuracy were calibrated in those participants who rejected the lineup (i.e., nonchoosers) in either the target-present or target-absent conditions, thereby enabling us to determine whether confidence in a lineup rejection provides a pointer to whether the suspect is the culprit. Previous research has consistently reported higher CA correlations for choosers than for nonchoosers (Sporer, 1992; Sporer et al., 1995), but used only the point-biserial correlation. To date, the CA calibration procedure has not been used with an identification paradigm to explore this issue. The importance of answering this question is highlighted by recent studies using a face recognition paradigm (Weber & Brewer, 2003, 2004) which showed poorer calibration for negative responses (i.e., lineup rejections) than for positive identification responses.

The fourth major issue examined was the impact on the CA relationship of variations in the base rate of target-absent lineups. The real world target-absent base rate is unknown, but probably varies across police jurisdictions under the influence of a variety of factors. Although police and prosecuting attorneys might claim rates of zero, the DNA exoneration cases clearly show that this is not the case. Anecdotal estimates of .10 have previously been reported (Brewer et al., 2002) although others suggest rates perhaps as high as .50 are not out of the question (R. C. L. Lindsay, personal communication, April 4, 2005). However, no independent verification of such claims exists. Many eyewitness identification studies use target-absent base rates of .50, though again this rate may no more reflect actual practice than does a rate of zero. (Studies that present 50% target-absent lineups are generally doing so to

provide a robust assessment of the effects of their independent variables when the target is absent, not because of an assumption that the target-absent base rate is 50%.) While the target-absent base rate in the real world may be unknown, there are strong grounds for expecting it to affect the CA relationship. For example, it is expected that most participants who made a positive identification from a target-absent lineup would report a confidence level higher ($>$ zero) than their accuracy level (zero). If the proportion of such participants is reduced, we expect altered calibration patterns because of a reduction in overconfidence. Although the target-absent base rate is likely to be an important influence on CA calibration for eyewitness identifications, there is no empirical evidence on this issue. There is, however, relevant evidence from the decision making literature, with Ferrell and McGoe (1980) and Lichtenstein et al. (1982) showing that calibration curves were affected by varying the proportion of true statements in a set of items (with discriminability held constant). As the proportion of true statements rose (analogous to reducing the base rate of target-absent lineups), the degree of overconfidence declined. In the present study our very large sample with a base rate of .50 provided sufficient data points to conduct post hoc (i.e., non-experimental) contrasts of calibration at that base rate with that from randomly drawn subsamples that provided target-absent base rates of .25 and .15, respectively.

Finally, we examined how the CA relationship varied with characteristics of the identification response, focusing specifically on identification latency. Response latency in a recognition memory paradigm is considered to be a key indicator of the strength of a memory trace accuracy (see, for example, Kahana & Loftus, 1999; Murdock & Dufty, 1972) and, in turn, the likelihood that an old stimulus is recognized as having been seen previously (Atkinson & Juola, 1974; Gillund & Shiffrin, 1984). Consistent with this position, a robust finding in the eyewitness identification literature is that identification accuracy and latency

are correlated, with rapid identification responses more likely to be correct than slower responses (Brewer et al., in press; Brewer, Gordon, & Bond, 2000; Dunning & Perretta, 2002; Sporer, 1992, 1994; Weber, Brewer, Wells, Semmler, & Keast, 2004). It has also been reliably found across various domains that overconfidence is less marked in association with higher levels of accuracy, generally known as the *hard-easy effect* (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, Winman, & Olsson, 2000). Thus, not only would faster identification responses be more likely to be accurate, but we also predicted reduced overconfidence and, most likely improved CA calibration for faster identifications. Perhaps most important for this argument, however, for those participants for whom the memory trace is very strong, and hence likely to be accessed rapidly, the likelihood of (false) positive identifications from target-absent lineups should be lower. Given that any confidence value exceeding zero for such incorrect identification responses contributes to overconfidence, the likelihood of witness overconfidence should be less for rapid than for slower identifications as there is likely to be a lower proportion of target-absent misidentifications.

All of the above issues were examined for two separate eyewitness identifications that produced quite distinct identification response patterns. Any or all of a number of non-quantifiable factors (that characterize all identification encoding and test stimuli) may have contributed to the differences in response patterns: for example, which of the two stimuli would have been seen by participants as the central figure, different exposure durations for the two stimuli, different degrees of match between encoding and test stimuli, and differences in the precise degree of discriminability of target and foils. Regardless of the underlying reasons for the identification performance difference, the different patterns enabled us to obtain converging evidence on the CA relationship across various conditions. In sum, this experiment used a between-subjects design and the calibration procedure to examine the CA

relationship for two eyewitness identifications given biased versus unbiased lineup instructions, high versus low foil similarity lineups, and target-present versus target-absent lineups. Although we expected relatively weak point-biserial correlations across stimuli and experimental conditions, we hypothesized that (a) the calibration approach would reveal a meaningful positive CA relation, with confidence discriminating accurate from inaccurate identifications, (b) the CA relationship would be detected for participants making positive, but not negative, identification decisions, (c) the degree of overconfidence would increase with the difficulty of the identification task (indexed by the ratio of hits to false alarms, or diagnosticity), (d) reducing the target-absent base rate would reduce overconfidence and, to the extent that identification decisions were characterized by overconfidence, improve calibration, and (e) overconfidence would be less pronounced and (again to the extent that overconfidence was the norm for the particular stimulus materials) calibration better for relatively fast, compared with slower, identification decisions.

Method

Participants

Twelve hundred participants (478 male, 722 female) were recruited from undergraduate and community groups and paid for their participation. Participants' ages ranged from 16 to 60 years ($M = 23.9$, $SD = 8.3$).

Materials

Stimulus event. The simulated crime was a non-violent theft of a credit card at a restaurant. The video (duration = 140 s) showed the thief entering the restaurant and waiting in the background while a customer was leaving his credit card on a counter for the waiter to process. When the customer left, the thief asked the waiter about a reservation as the waiter moved the credit card off the counter. When the waiter turned away to check the reservation

book, the thief reached for the credit card but withdrew his hand when the waiter turned around. The waiter then turned away again and the thief took the card. The phone rang and, while the waiter was answering it, the thief left the restaurant and ran away. At different stages of the film, both the thief and the waiter were seen from different angles after editing shots from two cameras. Some part of the thief's face was available for 23 s; the corresponding value for the waiter was 72 s.

The lineups and confidence scale. The lineups for the thief and waiter each consisted of 8 color photos (4 cm × 5.75 cm), arranged in two rows of four and shown simultaneously on a 15-inch computer screen (resolution of 1024 × 768 pixels). Below each photo was a number (1-8), and at the lower center of the screen was a button marked "Not Present". The original photos were scanned with a resolution of 100 × 144 pixels. All photos involved a front view from chest up. Both lineup targets were dressed in different clothing to that worn in the video. The mismatch between encoding and test stimuli was probably greater for the thief as his hair was uncombed in the lineup photo and neatly combed in the video. The position of the thief/waiter in the target-present lineup (or the replacement in the target-absent lineup) varied across all 8 positions. An identification response was made by clicking with the mouse on one of the photos or on the "Not Present" button. The foils, and the thief's/waiter's replacement in the target-absent lineup, were selected from an array of photos on the basis of their sharing a similar physical description with the thief/waiter: specifically, for (a) the thief: male, 20-30 years, olive skin, part Asian background,² and medium length dark hair, and (b) the waiter: male, 25-30 years, short to medium length dark hair, brushed back. The targeted overall base rate of target-absent lineups was .50 (actual = .499). The confidence scale appeared immediately after the identification response was made, and consisted of a row of buttons labeled 0%, 10%, 20%, ... 100%. There were three anchors on the scale, located

below the 0%, 50% and 100% buttons. The anchors were 0%/50%/100% confident that my decision was correct. Participants selected their confidence level by clicking with the mouse on the appropriate button.

Experimental manipulations

A between-groups design was used, with participants randomly assigned to biased or unbiased lineup instructions, high or low foil similarity lineups, and a target-present or target-absent thief lineup. All participants viewed the thief lineup and provided the associated dependent measures for that lineup before viewing the waiter lineup. As eyewitness identification studies generally involve only a single identification and there are no data on the impact of asking witnesses to perform multiple identifications on any particular identification, we did not counterbalance thief and waiter lineup orders. We did not want to risk the possibility that conclusions based on the data for the thief might be compromised by an order effect, especially given the very large number of participants required for calibration. Rather, we saw the waiter data as providing a valuable second data point and a large sample replication, even though there was a possibility that behavior on this second identification test might be influenced by having done a previous identification test. To obtain both a target-present and a target-absent data point from each participant, the second (waiter) lineup reversed the target-presence status of the first (thief) lineup.

Lineup instructions. Participants received either unbiased or biased instructions prior to viewing the lineup. The unbiased instructions were: “Now we would like you to identify the thief/waiter. He may or may not be in the lineup below. Please indicate your choice by clicking the button underneath the corresponding face. If you think the thief/waiter is not in the lineup, click the ‘Not Present’ button.” The biased instructions did not note the possibility

that the thief/waiter may not be present in the lineup. They were: “Now we would like you to try to identify the thief/waiter. Please indicate your decision by clicking on a face.”

Foil similarity. For each of the targets, two lineups were constructed that differed with respect to the number of people in the lineup who were rated as close matches to the offender. For the thief lineup, the foil similarity manipulation was based on a combination of face similarity ratings and the distribution of lineup foil choices recorded in a previous study. The face similarity ratings involved 30 independent observers viewing a series of slides, each of which contained the lineup version of the target and one of the possible lineup foils. The observers rated the similarity of each foil to the target using a 7-point scale (1 = very low degree of similarity; 7 = very high degree of similarity). To construct the low similarity lineup, three foils from the high foil similarity lineup – who were rated most similar to the target and were also the most frequently chosen foils in a separate study (Brewer et al., 2002) – with mean similarity ratings of 4.4 (SD = 1.8), 4.0 (SD = 1.6) and 3.8 (SD = 1.6), were replaced with foils with mean similarity ratings of 3.6 (SD = 1.2), 3.3 (SD = 1.3) and 2.9 (SD = 1.4). For the waiter lineup, the procedure was slightly different because we had no existing data on the distribution of foil choices and fewer photographs of potential foils. The low foil similarity lineup was constructed by replacing two foils, with mean similarity ratings of 3.1 (SD = 1.4) and 2.7 (SD = 1.4), with two other foils, both with similarity ratings of 1.9 (SD = 1.0). (Insufficient match-description foil photographs were available for replacement of all lineup foils.) In the absence of an objective index of similarity, it is difficult to categorize the two versions of the lineups in terms of some absolute level of similarity; the ratings data suggest that they are most appropriately described as being of relatively high and low similarity.

Procedure

Participants attended the laboratory, ostensibly to participate in a forensic psychology study. They watched the videotape in small groups (2-4), then worked in individual cubicles on a pencil-and-paper filler task (puzzles) for 15 minutes before commencing the identification test on a computer. A series of screen prompts guided the participants through the identification test and confidence assessment for the thief, with these steps then repeated for the waiter. All responses involved a simple mouse click on a button on the screen.

Results

Identification performance

Table 1 shows frequency data for the different identification response categories for target-present and -absent thief lineups across the different experimental conditions. For target-present lineups, a 3 (identification response) \times 2 (instructional bias) \times 2 (foil similarity) hierarchical loglinear analysis revealed an Identification Response \times Instructional Bias interaction, $\chi^2(2, N = 601) = 25.32$, $p < .01$, $w = 0.20$, with biased instructions reducing lineup rejections and increasing correct and foil identification responses. (Cutoffs for w for small, medium and large effect sizes are .1, .3 and .5.)

The Identification Response \times Foil Similarity interaction was not significant, $\chi^2(2, N = 601) = 3.11$, ns, $w = 0.07$. A 2 (identification response) \times 2 (instructional bias) \times 2 (foil similarity) hierarchical loglinear analysis for target-absent thief data also revealed an Identification Response \times Instructional Bias interaction, $\chi^2(1, N = 599) = 12.58$, $p < .01$, $w = 0.14$, with biased instructions increasing false identifications (see Table 1). The other non-reported effects for target-present and -absent thief identifications were not significant.

The waiter lineup yielded quite different identification response patterns (see Table 2), perhaps reflecting differences between the thief and waiter stimulus materials in any, or all,

of the following: the encoding conditions (e.g., exposure duration, central vs. peripheral character in the video), the encoding stimulus-test stimulus match (e.g., similarity of hair style at encoding and test), and/or the discriminability of the lineup target (e.g., number of similar foils or average similarity of foils). The proportion of correct identifications from target-present lineups was much higher for the waiter than for the thief (61.3% vs. 36.9%), and the proportion of lineup rejections much lower (16.7% vs. 45.6%). Identical loglinear analyses conducted on the waiter target-present lineup data again showed a significant Identification Response \times Instructional Bias interaction, $\chi^2(2, N = 599) = 11.77, p < .01, w = 0.14$, with biased instructions reducing lineup rejections. The Identification Response \times Foil Similarity interaction was also significant, $\chi^2(2, N = 599) = 7.49, p < .05, w = 0.11$, with fewer incorrect and more correct identifications for low than for high foil similarity lineups. The three-way interaction was not significant, $\chi^2(2, N = 599) = 0.96, ns$. The higher choosing rate, which translated into superior accuracy for the waiter target-present lineups, also extended to the target-absent lineup, resulting in a much higher rate of false identifications than for the thief lineup (54.7% vs. 32.9%). The Identification Response \times Instructional Bias interaction, $\chi^2(1, N = 601) = 32.93, p < .01, w = 0.37$, confirmed the higher proportion of false identifications with biased than with unbiased instructions. An Identification Response \times Instructional Bias \times Foil Similarity interaction, $\chi^2(1, N = 601) = 8.15, p < .01$, reflected the greater impact of unbiased instructions on choosing when foil similarity was low (34.7% vs. 69.3% false identifications under unbiased and biased instructions, respectively) than when it was high (51.7% vs. 63.3%).

In sum, instructional bias significantly affected identification response patterns for both targets in both target-present and -absent lineups; foil similarity did likewise for one of the

two targets. We also examined diagnosticity ratios across targets and conditions.

Diagnosticity is a measure of how well the status of the identified person (target vs. non-target) can be predicted based on the identification decision of the witness. It refers to the ratio of the probability that the suspect is identified, given that the suspect is the offender, to the probability that the suspect is identified, given that the suspect is not the offender. For target-absent lineups, all identifications were treated as incorrect suspect identifications as, unlike the real-world situation, there is not a basis for designating a particular lineup member (even the target's replacement) as the suspect. When calculating diagnosticity ratios, we assumed that all lineup members are equally likely to be the designated (innocent) suspect and, hence, divided the false identification rates in the target-absent lineups by 8. At the descriptive level, as shown in Table 3, diagnosticity ratios for positive identification responses were larger under unbiased than biased instructions for both the thief and waiter lineups. Also, the diagnosticity ratio was larger for high than low foil similarity lineups for the thief, while the opposite pattern prevailed for the waiter. Overall, the diagnosticity ratios for the thief and waiter were almost identical.

Diagnosticity for not present responses (i.e., the ratio of the probability that the offender is considered to be not present, given not present, to the probability that the offender is considered to be not present, given that the offender is present) was higher for the waiter than the thief, but the effects of the manipulations were not completely consistent across the two targets (see Table 3).

Confidence

Table 4 shows the mean (and standard deviation) confidence ratings for the different identification response categories for target-present and -absent thief lineups across experimental conditions. For target-present thief lineups, a 3 (identification response) \times 2

(instructional bias) \times 2 (foil similarity) analysis of variance (ANOVA) on confidence ratings (see Table 5) produced a main effect for identification response. Correct identifications were made with more confidence than incorrect rejections which, in turn, were made more confidently than incorrect identifications. Neither instructional bias nor foil similarity had meaningful effects. An Instructional Bias \times Foil Similarity interaction reflected higher confidence for high than low similarity lineups only under unbiased instructions. For target-absent lineups, a 2 (identification response) \times 2 (instructional bias) \times 2 (foil similarity) ANOVA confirmed lower confidence for false identifications than for correct lineup rejections. None of the other main effects or interactions were significant.

The corresponding data sets for the waiter are shown in Table 6. The ANOVAs on confidence (see Table 7) revealed identification response patterns similar to those found for the thief for target-present and target-absent conditions. The instructional bias effect was also significant for both target-present and target-absent conditions. However, while biased instructions produced higher confidence for target-present lineups, the opposite pattern occurred for target-absent lineups. None of the other effects were significant.

The CA relation

Positive identifications of foils from target-present lineups were excluded from analyses of the CA relation (i.e., from correlation and calibration analyses) as it is known in advance that, in a single-suspect lineup, a foil is not the culprit. The patterns of point-biserial CA correlations were consistent with previous research. In all conditions and for both targets the correlations were, at best, modest (see Table 8). Examination of the correlations reveals that, for every cell for both targets, correlations for choosers (i.e., correct and incorrect identifications from target present lineups, and false identifications from target absent lineups) were higher than for nonchoosers (i.e., not present responses to target-present and -

absent lineups). For the thief, significant differences between choosers and nonchoosers were detected for the high functional size and biased instruction conditions, as well as overall ($z = 2.14 - 2.33$). For the waiter, the correlations differed significantly for both levels of instructional bias and foil similarity, and overall ($z = 2.14 - 6.60$).

Figure 1 shows the calibration curves for participants making a positive identification response (i.e., choosers) from the thief and waiter lineups under the instructional bias and foil similarity manipulations (each collapsed across the other variable): the probability of being correct (i.e., $[\text{correct identifications}^{\text{target-present}}] \div [\text{correct identifications}^{\text{target-present}} + \text{false identifications}^{\text{target-absent}}]$) plotted against confidence. Any lineup choice from target-absent lineups was again considered to be a false identification, despite the fact that this approach almost certainly leads to an over-representation of false positives with confidence higher (i.e., $> 0\%$) than accuracy.

Figure 2 shows the corresponding curves for nonchoosers (i.e., $[\text{correct rejections}^{\text{target-absent}}] \div [\text{correct rejections}^{\text{target-absent}} + \text{incorrect rejections}^{\text{target-present}}]$). To provide more stable estimates in each confidence category for both Figures 1 and 2, confidence categories (for both choosers and nonchoosers) were collapsed into 5 categories (0%-20%, 30-40%, 50%-60%, 70%-80%, 90%-100%), with proportion correct in each of the combined categories compared with the weighted mean confidence for that category. Table 8 summarizes the C, O/U and NRI statistics for each target in the various conditions.

For choosers, the key results to note are as follows. The calibration curves for both targets generally follow the slope of the identity line, with the more obvious departures being at the lower end of the confidence scale. Across conditions, the C statistics were generally in the .01 - .02 range, although overconfidence was associated with both targets: for example, the 90-100% confidence levels were associated with accuracy levels between 75% and 90%. Across

conditions, higher levels of overconfidence were associated with lower diagnosticity ratios.

For both targets, the *C* statistics were closer to zero under unbiased instructions than under biased instructions, and the *O/U* statistics were larger for the latter. For both the thief and the waiter, differences in the calibration curves for the two foil similarity conditions were also evident at the upper end of the confidence scale, but the direction of the difference was reversed for the two targets and followed the diagnosticity ratios.

The data patterns for nonchoosers did not neatly parallel the diagnosticity ratios. For the thief, the calibration curves clearly show overconfidence at the highest confidence level and underconfidence at the lowest. In contrast, for the waiter, probability of a correct decision remained about the same, or even declined, across confidence categories. The NRI statistics confirm that confidence discriminated accurate and inaccurate identifications more effectively for choosers than nonchoosers for both targets, with effect sizes moderate for both targets for choosers and small for nonchoosers. As NRI can be interpreted as eta-squared (see Baranski & Petrusic, 1994; Yaniv et al., 1991), and eta-squared is directly related to Cohen's f , cutoffs for small, moderate and large NRIs can be calculated using the .1, .25 and .4 cutoffs for f . The respective NRI values are .010, .059 and .138.

Table 9 shows the distributions of confidence ratings for each target for choosers and nonchoosers, and for each identification response category collapsed across conditions. Also shown are the diagnosticity ratios at each confidence level. The most striking features of these data are the substantial increases in diagnosticity for choosers (of both targets) at the 90-100% confidence levels. Also noteworthy is the finding that participants did make positive identifications accompanied by very low confidence estimates, a pattern that may well reflect the tendency to make a choice from a lineup regardless of the lineup instructions.

To examine how the CA relationship varied with target-absent base rate, we drew two post hoc) random samples from the full database (target-absent base rates were .501 and .499 for thief and waiter lineups, respectively) to produce one sample with a target-absent proportion of .25 and another with a .15 proportion. To provide stable estimates for the calibration curves and associated statistics after removing large numbers of target-absent cases, we conducted this examination across the entire sample, not for each condition. Calibration curves and associated statistics for these two base rates, and for the full sample, are shown in Figure 3 and Table 10. For choosers of both targets, several characteristics are noteworthy. First, underconfidence, rather than overconfidence, was apparent at the two lower base rates. Second, inspection of the calibration curve indicates that it did not follow the identity line as closely at the lower target-absent base rates. Third, as Figure 3 shows, the calibration curve departs more markedly from the ideal (i.e., flattens out) in the lower half of the confidence scale, a pattern reflected in the NRI statistics. In other words, for choosers the usefulness of confidence ratings in the lower half of the scale is very much dependent on the target-absent base rate. For nonchoosers, lower target-absent base rates were associated with increased overconfidence, poorer calibration, and weak resolution.

To examine variations in calibration with identification latency, participants were divided into three groups (quick, medium, slow) based on their identification latency (see Table 11). No effects of lineup position on identification latency were detected. CA calibration curves (Figure 4) and statistics (Table 12) were then derived for choosers and nonchoosers within each group. These analyses were again only done on the full sample. Collapsing across conditions produced variations in the number of cases from the instructional bias and foil similarity conditions in the respective latency groups, but sample sizes for each condition for choosers versus nonchoosers were too small to produce stable

estimates for each calibration curve and the associated statistics. In line with predictions, calibration curves for choosers for both the thief and the waiter lineups more closely approached the ideal, the C statistics were nearer zero, and overconfidence was less (with the 95% confidence intervals including zero) for the quick group than for the medium and slow groups. (Confidence-identification latency correlations for both thief and waiter choosers and nonchoosers were significant and ranged from $-.30$ to $-.40$).

As Table 12 indicates, not only were choosers' C and O/U statistics lower, and NRI statistics higher, for the quick groups than for the others, but the diagnosticity ratios were also higher for the quick groups. Examination of the distribution of correct and incorrect choices (from both target-present and target-absent lineups) revealed, for both targets, patterns consistent with our predictions about how the CA relationship would vary with identification latency (see Table 13). First, accuracy was between 22% and 35% higher for the quick group than for the other two groups. Second, the proportion of incorrect identifications from target-absent lineups was between 17% and 30% lower for the quick group than for the other two groups. Both of these trends should, as predicted, contribute to a reduction in overconfidence.

None of these patterns were replicated for the nonchoosers. For the thief, the quick group differed little from the medium group, with both calibration curves displaying a positive CA relation but substantial overconfidence at the upper end of the confidence scale. The slow group's curve indicates no systematic pattern. For the waiter, none of the three groups' curves indicate meaningful patterns of accuracy variations across the confidence range. Although diagnosticity ratios were higher for the quick group, the ratios were much lower than for choosers.

Discussion

The relationship between eyewitness identification confidence and accuracy was examined (a) for two separate sets of encoding and test stimuli that elicited different identification response patterns, and (b) under two independent manipulations that induced changes in identification accuracy for one or both targets. The point-biserial CA correlations, though discriminating choosers from nonchoosers, were modest across all conditions. Despite the modest CA correlations, plotting confidence against proportion correct for choosers clearly indicated a positive relationship between confidence and accuracy for both sets of stimulus materials under all experimental conditions. As shown by the point-biserial correlations, the resolution statistics indicated that confidence was useful in discriminating correct from incorrect positive identification responses. Further, diagnosticity for positive identifications made with 90-100% confidence was high. The same patterns were not evident for nonchoosers, providing the first clear evidence using an eyewitness identification (rather than a face recognition) paradigm that identification confidence and accuracy are not well calibrated for lineup rejection responses.

For choosers, overconfidence was evident across all conditions, especially in the upper half of the confidence scale. This pattern was more marked for biased than unbiased instructions, whereas the effect of foil similarity differed across the two targets. (Note that the degree of overconfidence was more marked when positive identifications of foils known not to be the culprit were retained in the calibration analyses.) Consistent with face recognition paradigm findings – and findings from domains such as general knowledge and psychophysical discrimination (Baranski & Petrusic, 1994; Lichtenstein & Fischhoff, 1977) – that have shown increased overconfidence as difficulty increases (Olsson, 2000; Weber & Brewer, 2004), overconfidence was more marked when the diagnosticity ratio was lower for

both sets of stimuli and within all conditions. Nevertheless, we note below that, while the variations in the CA relation with diagnosticity are consistent with literature in other areas, they are not always so important from an applied perspective.

The picture of the CA relationship provided by the calibration curves and associated statistics from the present study (using new stimulus materials and experimental manipulations) complements the eyewitness identification findings reported by Juslin et al. (1996) and evidence obtained using the calibration approach with a face recognition paradigm involving mini-lineups (Weber & Brewer, 2004). It contrasts, however, with the poor CA relationship reported by Brewer et al. (2002) in their control condition (though not with that detected in their experimental groups) using the calibration approach. There are two differences between the two studies. First, the target-absent base rate was much higher in the present study (.50 vs. .15). Given that a higher target-absent base rate will produce more (overconfident) false identifications, this feature cannot explain the different patterns. The other difference between the two studies seems likely to provide the key. In the present experiment participants provided their confidence assessments immediately after indicating their identification response. In Brewer et al. (2002), the two (well calibrated) experimental groups completed other activities for five minutes after making their identification response and prior to their confidence assessment and, accordingly, there was an equivalent five-minute delay for the control group. Participants were not able to discuss the experiment with anyone in that interval, thereby minimizing the likelihood of any social influences on their confidence judgments. Nevertheless, as we indicated earlier, it is possible that internal or mnemonic cues to confidence decayed or were degraded during the interval, thereby undermining the veridicality of the confidence judgment. Indeed, preliminary findings from our laboratory which compare calibration curves after a zero versus a five-minute delay

suggest better CA calibration, and less overconfidence, for immediate compared with delayed confidence assessments (Brewer, Weber, & Semmler, in press). Elsewhere, the effects of postidentification influences (e.g., postidentification feedback) that can significantly inflate confidence have been well documented (Wells & Bradfield, 1998, 1999). The contrast between the present findings and those reported in our previous research further highlight the importance of collecting confidence assessments without any intervening delay.

Two other new findings emerged. First, as our overall sample was exceptionally large, we were able to conduct a post hoc examination of the effects of selecting lower target-absent base rates. Lower target-absent base rates were, as expected, associated with reductions in overconfidence. The calibration procedure indicated underconfidence for both targets, although for the thief confidence only discriminated accuracy within the upper half of the confidence scale. Given that the actual (i.e., real world) target-absent base rate is not known, and that it is likely to vary depending on practices followed in different jurisdictions, the specificity of our conclusions is necessarily constrained. Nevertheless, the calibration curves and associated statistics for choosers presented here suggest an overall conclusion that is somewhat different to that which has typically characterized the literature in this area. Specifically, regardless of target-absent base rate, confidence estimates of 50% or higher assessed at the time of a positive identification provided a rough guide to the likelihood that an identification is accurate. Whether some degree of over- or underconfidence is likely to be the norm for identification responses cannot be determined without a broader database and objective data on target-absent base rates. The latter, of course, cannot be known. Nevertheless, it is possible that some crude estimates could be obtained by careful archival analyses of large samples of police investigations in which identification tests have been

conducted with a view to obtaining converging evidence on whether lineups were target-present or-absent.

Second, the variations in choosers' calibration patterns with identification latency provide a further indication that confidence estimates are indeed related to, rather than dissociated from, the quality of the witness's memory for the person observed. We hypothesized that witnesses with strong memories for the target face(s) would be likely to recognize the target rapidly and be characterized by higher identification accuracy. Further, they would be less likely to identify a non-target, with a resultant lower proportion of (confident) false identifications. The outcome should be, and was, more impressive calibration for rapid identifications.

As we indicated previously the eyewitness identification paradigm does not yield the multiple data points that permit the application of conventional statistical tests to the various statistics generated via the calibration approach. This, of course, raises questions about exactly how data obtained using this approach should be interpreted. If the C and O/U statistics are zero, it is clear that the CA relationship is a perfect one. But what kinds of departures from these values still denote a good or useful relationship? There is no single answer to this question. Rather, the answer will vary depending on the researcher's objectives: for example, a comparison of CA relationships across different conditions versus an interpretation of the meaning that might be attached to a particular witness's confidence judgment versus an examination of whether confidence assists in the discrimination of identification accuracy at the broader sample level. Nevertheless, conclusions regarding the likely realism of any witness's confidence estimate should be guided by careful examination of the calibration curve and associated statistics. A linear CA relationship following the

identity line, coupled with an O/U statistic with a narrow 95% confidence interval that spans the zero value, suggests realism in judgments.

Reliance on such data should, of course, be backed up by the knowledge that the sample size is sufficient to produce stable estimates. Lichtenstein et al. (1982) indicated that stable calibration data may require several hundred responses; Juslin et al. (1996) suggested more than 200 responses. Researchers can get a sense of the likely stability of their estimates if they consider the impact of adding a few data points to either the denominator or numerator when calculating accuracy at each confidence level. For example, 9/10 correct identifications in the 90-100% confidence range denotes a proportion correct of .90. An extra 3 cases in this confidence category could cause this value to vary between .69 and .92, thereby substantially altering the conclusions about the realism of confidence judgments.

So, from a practical perspective, what do our various findings allow us to say? First, let us emphasize that they should not be construed as offering any encouragement for the use of any expression of eyewitness identification confidence in the courtroom as an index of likely accuracy. Our confidence data were obtained at the time of the identification and without any possibility of interaction with the lineup administrator. Consequently, the confidence estimates were protected from the biasing effects of any of the typical social influences that can (a) operate between the time of making an identification and giving testimony in the courtroom, and (b) render any expression of confidence at that time uninterpretable. In sum, our data in no way challenge the previously stated view (e.g., Sporer et al., 1995) that identification confidence expressed in the courtroom (and not previously recorded at the time of the identification) should be ignored.

Our findings do, however, have important implications at the police investigative level. There is now converging evidence of a meaningful CA relation (for choosers) from our and

other laboratories (Juslin et al., 1996), using different stimulus materials, provided the confidence estimate is recorded at the time of the identification and is an independent assessment from the witness. These findings suggest that police ought to use witness confidence as one of the factors guiding their thinking about the possible identity of the offender. This is not to say that confident witnesses (even at the time of the identification) cannot be wrong; clearly, they can be and police need to be fully aware of this. But knowing that a highly confident identification is much more likely to be accurate than an unconfident one provides an important piece of information for the police: namely, that it is worthwhile checking out their hypothesis about this particular suspect very carefully. Conversely, knowing that an unconfident identification from a particular witness has, in many conditions, a low probability of being accurate should raise serious doubts about the offender's identity in the minds of the investigating police and suggest that they should cast their net more widely when evaluating suspects. In other words, we are making a clear distinction here between (a) the evidentiary value of confidence (expressed at the time of identification) for shaping investigators' probabilistic evaluations of their hypotheses about the offender's identity and (b) the probative value of confidence (expressed following some delay after the identification) for jurors.

As an aside, it is worth noting that, currently, lineup conduct practices vary across, and even within, police jurisdictions, including live lineups, photoarrays involving various different formats, and computerized presentations such as used in the present study. Although it is difficult to forecast the direction(s) which police procedures will take, we note that computer-presented lineups offer considerable advantages with respect to minimizing social influences on witness judgments and standardizing lineup administration.

Interestingly, our findings that the degree of over- and underconfidence varies with overall identification performance, or diagnosticity, is not in our view always of immense applied significance. There are two main reasons for saying this. First, for any individual identification in the real world, it is unlikely that, except under the most extreme conditions (e.g., the event took place at night and at a considerable distance from the witness), police investigators will be able to specify those conditions likely to affect diagnosticity (e.g., quality of encoding conditions or encoding-test stimulus match, or discriminability of target from foils) with sufficient precision to inform an appraisal of the possible under- or overconfidence of a positive identification, although some variables likely to affect diagnosticity can indeed be controlled by police investigators (e.g., effective size of the lineup, lineup instructions, lineup presentation mode). Second, our data reinforce what researchers have known for a long time: namely, an extremely confident witness can be “just plain wrong”, regardless of the encoding and identification test conditions that shape the patterns of identification responding. In other words, our findings merely reinforce the conclusion stated earlier that confidence can provide a strong pointer for police investigators but certainly cannot be taken as unequivocal evidence of identification accuracy.

Knowing that confidence is at its most useful as a guide to accuracy when the identification is made relatively rapidly potentially has some practical value, but again this value should not be exaggerated. While there is not any objective standard for what constitutes a fast identification under any particular set of circumstances (see, for example, Brewer et al., 2005; Weber et al., 2004), the different CA patterns for fast versus slow identifications certainly suggest that investigators should be very careful about attaching any importance to the witness’s confidence assessment when the witness has taken a long time to make their decision. Our data clearly indicate that confidence is a very dubious indicator of

accuracy when identifications are relatively slow (i.e., latency longer than 18 s and 24 s for the thief and waiter stimuli, respectively, and the particular testing conditions used here).

We also have provided the first decisive CA evidence using the calibration approach for those witnesses who rejected the lineup (i.e., nonchoosers). Although the findings may simply appear to confirm previous conclusions (e.g., Sporer, 1992; Sporer et al., 1995), those conclusions relied on the point-biserial correlation when examining the CA relation, an approach now clearly shown to be of limited value for evaluating a single confidence judgment (cf. Juslin et al., 1996), although it does inform our understanding of the variance in identification accuracy accounted for by confidence across a group of witnesses. Consistent with data for positive versus negative decisions obtained using a face recognition paradigm (Weber & Brewer, 2003, 2004), confidence is clearly not a useful indicator of the likely accuracy of a lineup rejection. This does not mean that a lineup rejection is not informative about whether or not the suspect is the culprit. Indeed, Wells and Olson (2002) have shown that lineup rejections provide important information about the likely involvement of the suspect, so clearly police ought to take very seriously a witness's judgment that a suspect in the lineup does not match their memory of the culprit. But, unfortunately, the witness's confidence in such a rejection appears to be unhelpful in assessing just how much weight such a rejection should receive.

More generally, our results underscore the fact that different methodological and statistical approaches can lead to different conclusions regarding the extent to which people can accurately self-report their abilities. As outlined earlier, Ackerman et al. (2002) demonstrated that self-reported abilities and knowledge were strongly correlated with objective measures of abilities and knowledge when they used an individual differences approach that included multi-item assessments of participants' relative standing on abilities

and knowledge. One of the characteristics of the eyewitness identification paradigm is that both the accuracy assessment and the confidence assessment are single items. Although using only one accuracy item and one confidence item has ecological validity (i.e., this is the case for actual crime witnesses), the reliability of the accuracy score and the confidence score cannot be assessed and reliability is likely to be severely limited by the single-item aspect of the method. Furthermore, individual differences in how participants use the confidence scale, their individual propensities toward underconfidence or overconfidence, and other matters related to individual differences contribute to error variance and cannot be removed with the single-item approach used in eyewitness identification research. Given the inability to account for individual differences and the questionable stability of a single-item assessment of accuracy and of confidence, the outcomes revealed here by the calibration approach are impressive.

To what extent individual differences actually do underpin the CA relation is, of course, an issue for future research. As Lichtenstein et al. (1982) argued, the influence of individual difference variables may well depend on task difficulty, with individual difference factors leading each person to be ‘best calibrated’ at a particular level of difficulty but under- or overconfident given alternative stimuli to judge. Although such information is likely to be of little practical value when it comes to evaluating a judgment made by a witness at a particular identification test, understanding the contribution of such factors is important for advancing theorizing about confidence judgment processes.

In summary, this study has highlighted meaningful CA relationships across stimulus materials and experimental conditions, relationships that could assist in guiding police investigations. For choosers, the calibration curves and associated statistics are at odds with the often expressed view that there is no meaningful (or useful) relationship between

eyewitness identification confidence and accuracy, though should not weaken conclusions about the minimal probative value of courtroom confidence. The high diagnosticity ratios for highly confident, positive identifications suggest that confidence, appropriately recorded and interpreted, may be a very useful indicator for police investigators. Future research should, however, continue to explore the boundaries of the CA relation using the calibration approach. For example, a more powerful version of the foil similarity manipulation used in the present study may produce more striking effects on the CA relationship than those detected here. Other manipulations may do likewise. For example, the interval between a crime and the identification test will be considerably longer than that used in this study: the average interval in the U.K. is a month (Pike, Brace, & Kynan, 2002). To what extent such variables differentially affect accuracy and confidence, perhaps by increasing the likelihood that witnesses will entertain analytic or metacognitive cues that are likely to influence confidence judgments, remains to be determined.

Research should also probe the impact of different approaches to the scaling of confidence. For example, at present, police are unlikely to collect witnesses' confidence estimates in a numerical form (although, as computerized lineup presentations become more prevalent, numerical estimates could be easily incorporated). It would be interesting to chart the CA relations when witnesses express confidence on purely verbal scales, including scales involving relatively few judgment categories (e.g., very sure, sure, unsure, very unsure). While such scales would not permit the calculation of the usual calibration statistics, they may reveal more impressive CA correspondences than those presented here.

Our results provide further strong evidence against the reliance on the point-biserial correlation for examining the CA relationship. Although use of that statistic has previously distinguished between CA relations for choosers and nonchoosers, it conceals meaningful CA

relations for positive identifications. Despite the possible temptation for researchers to continue with that approach to assessing the CA relationship, perhaps principally because of the large scale data collection required for the calibration approach, we believe it is often inappropriate to rely on this statistic. We advocate the following approach. If the primary focus of a piece of research is the CA relationship, and particularly the meaning that can be attached to particular levels of confidence (or to a single confidence level), researchers should collect sufficient data to provide stable estimates of the CA calibration function and the associated statistics. If the focus is the extent to which confidence explains variance in accuracy across participants – and how this might be affected by variables of interest – the point-biserial correlation (or the NRI) is informative. When, however, the CA relationship is a peripheral focus of a piece of research (as often appears to be the case in studies in which the CA point-biserial correlation is reported), the routine reporting of the point-biserial correlation will most likely make a limited contribution to the overall significance of the research.

From a practical perspective, a major challenge will be the communication of just what results such as ours mean for practitioners within the criminal justice system. Highlighting, and ensuring the maintenance of, distinctions between those confidence estimates that are likely to provide useful guidance versus those that will not, the relative versus absolute diagnostic value of a highly confident identification, and the use of confidence assessments obtained in the investigatory versus the courtroom phase represent significant challenges for researchers.

References

- Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really now about our abilities and our knowledge. Personality and Individual Differences, *33*, 587-605.
- Allwood, C. M., & Granhag, P. A. (1996). The effects of arguments on realism in confidence judgments. Acta Psychologica, *91*, 99-119.
- Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology: Vol. 1. Learning, memory & thinking (pp. 243-293). San Francisco, CA: Freeman.
- Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. Perception and Psychophysics, *55*, 412-428.
- Bornstein, B. H., & Zickafoose, D. J. (1999). "I know I know it, I know I saw it": The stability of the confidence-accuracy relationship across domains. Journal of Experimental Psychology: Applied, *5*, 76-88.
- Bothwell, R. K., Deffenbacher, K. A., Brigham, J. C. (1987). Correlations of eyewitness accuracy and confidence: Optimality hypothesis revisited. Journal of Applied Psychology, *72*, 691-695.
- Bradfield, A. L., & Wells, G. L. (2000). The perceived validity of eyewitness identification testimony: A test of the five Biggers criteria. Law and Human Behavior, *24*, 581-594.
- Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. Law and Human Behavior, *26*, 353-364.
- Brewer, N., Caon, A., Todd, C., & Weber, N. (in press). Eyewitness identification accuracy and response latency. Law and Human Behavior.

- Brewer, N., Gordon, M., & Bond, N. (2000). Effect of photoarray exposure duration on eyewitness identification accuracy and processing strategy. Psychology, Crime and Law, 6, 21-32.
- Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. Journal of Experimental Psychology: Applied, 8, 46-58.
- Brewer, N., Weber, N., & Semmler, C. (2005). Eyewitness identification. In N. Brewer & K. D. Williams (Eds.), Psychology and law: An empirical perspective (pp. 177-221). New York: Guilford.
- Brewer, N., Weber, N., & Semmler, C. (in press). A role for theory in eyewitness identification research. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. Toglia (Eds.), Handbook of eyewitness psychology: Volume 2: Memory for people. Mahwah, NJ: Lawrence Erlbaum.
- Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. Psychonomic Bulletin and Review, 7, 26-48.
- Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. Memory and Cognition, 22, 273-280.
- Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. Law and Human Behavior, 29, 395-424.
- Cutler, B. L., & Penrod, S. D. (1989). Moderators of the confidence-accuracy correlation in face recognition: The role of information processing and base-rates. Applied Cognitive Psychology, 3, 95-107.

- Cutler, B. L., & Penrod, S. D. (1995). Mistaken identification: The eyewitness, psychology, and the law. New York: Cambridge University Press.
- Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). The reliability of eyewitness identification. Law and Human Behavior, *11*, 233-258.
- Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Jury decision making in eyewitness identification cases. Law and Human Behavior, *12*, 41-56.
- Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about the relationship. Law and Human Behavior, *4*, 243-260.
- Deffenbacher, K. A., & Loftus, E. F. (1982). Do jurors share common understanding concerning eyewitness behavior? Law and Human Behavior, *6*, 15-30.
- Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10 - to - 12 second rule for distinguishing accurate from inaccurate positive identifications. Journal of Applied Psychology, *87*, 951-962.
- Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities Organizational Behavior and Human Performance, *26*, 32-53.
- Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. Forecasting, *1*, 155-172.
- Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance, *3*, 552-564.
- Furnham, A., & Thomas, C. (2004). Parents' gender and personality and estimates of their own and their children's intelligence. Personality and Individual Differences, *37*, 887-903.

- Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. Psychological Review, *98*, 506-528.
- Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. Psychological Review, *91*, 1-67.
- Gonzalez, R., Ellsworth, P. C., & Pembroke, M. (1993). Response biases in lineups and showups. Journal of Personality and Social Psychology, *64*, 525-537.
- Granhag, P. A., Strömwall, L. A., & Allwood, C. M. (2000). Effects of reiteration, hindsight bias, and memory on realism in eyewitness confidence. Applied Cognitive Psychology, *14*, 397-420.
- Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. New York: John Wiley.
- Jonsson, A. C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. Personality and Individual Differences, *34*, 559-574.
- 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.
- Kahana, M., & Loftus, G. (1999). Response time versus accuracy in human memory. In R. J. Sternberg (Ed.), The nature of cognition (pp. 323-384). Cambridge, MA: MIT Press.

- Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the “general acceptance” of eyewitness testimony research: A new survey of the experts. American Psychologist, 56, 405-416.
- Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, 107-118.
- Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? Organizational Behavior and Human Performance, 20, 159-183.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 306-334). Cambridge: Cambridge University Press.
- 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. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. Law and Human Behavior, 4, 303-314.
- Lindsay, R. C. L., Wells, G. L., & O'Connor, F. J. (1989). Mock-juror belief of accurate and inaccurate eyewitnesses. Law and Human Behavior, 13, 333- 339.
- Lindsay, R. C. L., Wells, G. L., & Rumpel, C. (1981). Can people detect eyewitness identification accuracy within and between situations? Journal of Applied Psychology, 66, 79-89.

- Luus, C. A. E., & Wells, G. L. (1994). The malleability of eyewitness confidence: Co-witness and perseverance effects. Journal of Applied Psychology, *79*, 714-723.
- Macmillan, N. A., & Creelman, C. D. (1991). Detection theory: A user's guide. New York: Cambridge University Press.
- Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. Journal of Applied Psychology, *66*, 482-489.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. Psychological Review, *87*, 252-271.
- Memon, A., Hope, L., & Bull, R. (2003). Exposure duration: Effects on eyewitness accuracy and confidence. British Journal of Psychology, *94*, 339-354.
- Murdock, B. B., Jr., & Dufty, P. O. (1972). Strength theory and recognition memory. Journal of Experimental Psychology, *94*, 284-290.
- Neil v. Biggers, 409 U.S. 188 (1972).
- Noon, E., & Hollin, C.R. (1987). Lay knowledge of eyewitness behaviour: A British survey. Applied Cognitive Psychology, *1*, 143-153.
- Nosworthy, G.J., & Lindsay, R.C.L. (1990). Does nominal lineup size matter? Journal of Applied Psychology, *75*, 358-361.
- Olsson, N. (2000). A comparison of correlation, calibration, and diagnosticity as measures of the confidence-accuracy relationship in witness identification. Journal of Applied Psychology, *85*, 504-511.
- Olsson, N., Juslin, P., & Winman, A. (1998). Realism of confidence in earwitness versus eyewitness identification. Journal of Experimental Psychology: Applied, *4*, 101-118.

- Petrusic, W. M., & Baranski, J. V. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *American Journal of Psychology*, *110*, 543-572.
- Pike, G., Brace, N., & Kynan, S. (2002). The visual identification of suspects: Procedures and practice (Briefing Note 2/02). London: Home Office.
- Potter, R., & Brewer, N. (1999). Perceptions of witness behaviour-accuracy relationships held by police, lawyers and jurors. *Psychiatry, Psychology and Law*, *6*, 97-103.
- Semmler, C., Brewer, N., & Wells, G. L. (2004). Effects of postidentification feedback on eyewitness identification and nonidentification confidence. *Journal of Applied Psychology*, *89*, 334-346.
- Sporer, S. L. (1992). Post-dicting eyewitness accuracy: Confidence, decision times and person descriptions of choosers and non-choosers. *European Journal of Social Psychology*, *22*, 157-180.
- Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision-times in simultaneous and sequential lineups. *Journal of Applied Psychology*, *78*, 22-33.
- Sporer, S. L. (1994). Decision times and eyewitness identification accuracy in simultaneous and sequential lineups. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 300-327). New York: Cambridge University Press.
- 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.
- Stebly, N. M. (1997). Social influence in eyewitness recall: A meta-analytic review of lineup instruction effects. *Law & Human Behavior*, *21*, 283-297.

- Tredoux, C. (2002). A direct measure of facial similarity and its relation to human similarity perceptions. Journal of Experimental Psychology: Applied, 8, 180-193.
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. Psychological Review, 101, 547-567.
- Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. Journal of Experimental Psychology: Learning, Memory, & Cognition, 26, 582-600.
- Vickers, D. (1979). Decision processes in visual perception. New York: Academic Press.
- 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). Confidence-accuracy calibration in absolute and relative face recognition judgments. Journal of Experimental Psychology: Applied, 10, 156-172.
- Weber, N., Brewer, N., Wells, G. L., Semmler, C., & Keast, A. (2004). Eyewitness identification accuracy and response latency: The unruly 10-12 second rule. Journal of Experimental Psychology: Applied, 10, 139-147.
- Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist*, 48, 553-571.
- Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. Journal of Applied Psychology, 83, 360-376.
- Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitnesses' recollections: Can the postidentification-feedback effect be moderated? Psychological Science, 10, 138-144.
- Wells, G. L., Lindsay, R. C. L., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. Journal of Applied Psychology, 64, 440-448.

- Wells G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), Eyewitness testimony: Psychological perspectives (pp. 155-170). New York: Cambridge University Press.
- Wells, G. L., & Olson, 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.
- Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. Journal of Experimental Psychology: Applied, 2, 343-364.
- Wixted, J. T., & Stretch, V. (2005). In defense of the signal-detection interpretation of remember/know judgments. Psychonomic Bulletin and Review, 11, 616-641.
- Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. Psychological Bulletin, 110, 611-617.
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. Journal of Memory & Language, 46, 441-517.

Footnotes

¹We recognize that biased instructions are not recommended lineup practice. They are used here as this manipulation is a reliable way of adjusting decision criteria.

²Comparisons of Asian background (generally around 5-10% of the undergraduate component of our samples) and non-Asian participants have revealed no significant differences in identification performance for this target (Brewer et al., 2002).

Author Notes

Neil Brewer, School of Psychology, and Gary L. Wells, Department of Psychology.

This research was supported by grants 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.

We gratefully acknowledge Sarah Hollitt for her Herculean data collection efforts and Nathan Weber for cranking analyses and providing many useful comments.

Correspondence concerning this article should be addressed to Neil Brewer, School of Psychology, Flinders University, GPO Box 2100, Adelaide, South Australia, Australia.

Electronic mail may be sent to neil.brewer@flinders.edu.au

Table 1

Frequency and Percentage of Identification Responses in Each Condition for Target-Present and Target-Absent Thief Lineups

| | | Identification Response: Target-Present Lineup | | | | | | |
|-----------------------|----------------------|--|------|-----------|------|-------------|------|-------|
| | | Correct | | Incorrect | | Not Present | | Total |
| Instructional Bias | | N | % | N | % | N | % | N |
| | High Foil Similarity | | | | | | | |
| Unbiased | | 56 | 37.1 | 18 | 11.9 | 77 | 51.0 | 151 |
| Biased | | 64 | 42.7 | 36 | 24.0 | 50 | 33.3 | 150 |
| Overall | | 120 | 39.9 | 54 | 17.9 | 127 | 42.2 | 301 |
| Low Foil Similarity | | | | | | | | |
| Unbiased | | 45 | 30.0 | 17 | 11.3 | 88 | 58.7 | 150 |
| Biased | | 57 | 38.0 | 34 | 22.7 | 59 | 39.3 | 150 |
| Overall | | 102 | 34.0 | 51 | 17.0 | 147 | 49.0 | 300 |

Table 1 (cont.)

| Identification Response: Target-Absent Lineup | | | | | |
|---|-------------------|------|----------------------|------|-------|
| Instructional Bias | Correct Rejection | | False Identification | | Total |
| | N | % | N | % | N |
| High Foil Similarity | | | | | |
| Unbiased | 107 | 71.8 | 42 | 28.2 | 149 |
| Biased | 94 | 62.7 | 56 | 37.3 | 150 |
| Overall | 201 | 67.2 | 98 | 32.8 | 299 |
| Low Foil Similarity | | | | | |
| Unbiased | 114 | 76.0 | 36 | 24.0 | 150 |
| Biased | 87 | 58.0 | 63 | 42.0 | 150 |
| Overall | 201 | 67.0 | 99 | 33.0 | 300 |

Table 2

Frequency and Percentage of Identification Responses in Each Condition for Target-Present and Target-Absent Waiter Lineups

| | | Identification Response: Target-Present Lineup | | | | | | |
|---------------|--|--|------|-----------|------|-------------|------|-------|
| | | Correct | | Incorrect | | Not Present | | Total |
| Instructional | | N | % | N | % | N | % | N |
| Bias | | N | % | N | % | N | % | N |
| | | High Foil Similarity | | | | | | |
| Unbiased | | 86 | 57.7 | 31 | 20.8 | 32 | 21.5 | 149 |
| Biased | | 83 | 55.3 | 48 | 32.0 | 19 | 12.7 | 150 |
| Overall | | 169 | 56.5 | 79 | 26.4 | 51 | 17.1 | 299 |
| | | Low Foil Similarity | | | | | | |
| Unbiased | | 94 | 62.7 | 24 | 16.0 | 32 | 21.3 | 150 |
| Biased | | 104 | 69.3 | 29 | 19.3 | 17 | 11.3 | 150 |
| Overall | | 198 | 66.0 | 53 | 17.7 | 49 | 16.3 | 300 |

Table 2 (cont.)

| Identification Response: Target-Absent Lineup | | | | | |
|---|-------------------|------|----------------------|------|-------|
| Instructional Bias | Correct Rejection | | False Identification | | Total |
| | N | % | N | % | N |
| High Foil Similarity | | | | | |
| Unbiased | 73 | 48.3 | 78 | 51.7 | 151 |
| Biased | 55 | 36.7 | 95 | 63.3 | 150 |
| Overall | 128 | 42.5 | 173 | 57.5 | 301 |
| Low Foil Similarity | | | | | |
| Unbiased | 98 | 65.3 | 52 | 34.7 | 150 |
| Biased | 46 | 30.7 | 104 | 69.3 | 150 |
| Overall | 144 | 48.0 | 156 | 52.0 | 300 |

Table 3

Diagnosticity Ratios in Each Condition for Both Targets for Positive Identifications and Not Present Responses

| Target - Condition | Identification Response | |
|----------------------|-------------------------|-------------|
| | Positive Identification | Not Present |
| Thief | | |
| Overall | 8.99 | 1.47 |
| Unbiased | 10.29 | 1.35 |
| Biased | 8.13 | 1.66 |
| High Foil Similarity | 9.72 | 1.59 |
| Low Foil Similarity | 8.23 | 1.37 |
| Waiter | | |
| Overall | 8.95 | 2.71 |
| Unbiased | 11.15 | 2.65 |
| Biased | 7.52 | 2.81 |
| High Foil Similarity | 7.87 | 2.49 |
| Low Foil Similarity | 10.15 | 2.94 |

Table 4

Means and Standard deviations for Confidence of Different Identification Responses in
Each Condition for Target-Present and Target-Absent Thief Lineups

| | | Identification Response: Target-Present Lineup | | | | | | | |
|----------------------|--|--|-----------|-------------|-----------|-------------|-----------|-------------|-----------|
| | | Correct | | Incorrect | | Not Present | | Overall | |
| Instructional | | <u>Mean</u> | <u>SD</u> | <u>Mean</u> | <u>SD</u> | <u>Mean</u> | <u>SD</u> | <u>Mean</u> | <u>SD</u> |
| Bias | | | | | | | | | |
| High Foil Similarity | | | | | | | | | |
| Unbiased | | 71.3 | 20.4 | 63.3 | 28.5 | 60.7 | 25.0 | 64.9 | 24.2 |
| Biased | | 70.2 | 23.3 | 47.8 | 18.5 | 63.4 | 23.9 | 62.5 | 24.0 |
| Overall | | 70.7 | 21.9 | 53.0 | 23.3 | 61.7 | 24.5 | 63.7 | 24.1 |
| Low Foil Similarity | | | | | | | | | |
| Unbiased | | 63.8 | 19.5 | 48.8 | 24.2 | 55.9 | 24.3 | 57.5 | 23.3 |
| Biased | | 72.1 | 22.7 | 53.2 | 23.8 | 58.1 | 22.6 | 62.3 | 24.1 |
| Overall | | 68.4 | 21.7 | 51.8 | 23.8 | 56.8 | 23.6 | 59.9 | 23.8 |
| Overall | | | | | | | | | |
| Unbiased | | 67.9 | 20.2 | 56.3 | 27.1 | 58.1 | 24.7 | 62.0 | 21.2 |
| Biased | | 71.1 | 23.0 | 50.4 | 21.3 | 60.6 | 23.2 | 60.5 | 23.5 |
| Overall | | 69.6 | 21.8 | 52.4 | 23.4 | 59.1 | 24.1 | 61.1 | 22.6 |

Table 4 (cont.)

| Identification Response: Target-Absent Lineup | | | | | | |
|---|----------|-------------------|-----------|----------------------|-----------|-----------------------|
| | | Correct Rejection | | False Identification | | Total |
| Instructional | Bias | <u>Mean</u> | <u>SD</u> | <u>Mean</u> | <u>SD</u> | <u>Mean</u> <u>SD</u> |
| High Foil Similarity | | | | | | |
| | Unbiased | 69.6 | 20.7 | 59.1 | 14.1 | 66.6 19.6 |
| | Biased | 68.8 | 21.3 | 57.3 | 54.0 | 64.5 21.5 |
| | Overall | 69.3 | 21.0 | 58.1 | 17.7 | 65.6 20.1 |
| Low Foil Similarity | | | | | | |
| | Unbiased | 65.7 | 21.9 | 54.2 | 20.2 | 62.9 22.0 |
| | Biased | 68.3 | 22.3 | 54.0 | 22.0 | 62.3 23.2 |
| | Overall | 66.8 | 22.0 | 54.0 | 21.2 | 62.6 22.6 |
| Overall | | | | | | |
| | Unbiased | 67.6 | 21.3 | 56.8 | 17.2 | 63.5 23.3 |
| | Biased | 68.6 | 21.8 | 55.5 | 21.1 | 65.6 22.6 |
| | Overall | 68.0 | 21.5 | 56.0 | 19.6 | 64.4 23.0 |

Table 5

ANOVA on Confidence Summary Statistics for the Thief Lineup

| Target presence | Source | df | F | f |
|------------------------|-----------------------------|-----------------------------|----------|---------|
| Present | Identification response (R) | 2 | 19.14** | 0.25 |
| | Instructional bias (I) | 1 | 0.01 | 0.00 |
| | Foil Similarity (F) | 1 | 3.71 | 0.08 |
| | R × I | 2 | 1.39 | 0.07 |
| | R × F | 2 | 0.15 | 0.02 |
| | I × F | 1 | 5.12* | 0.09 |
| | R × I × F | 2 | 1.86 | 0.08 |
| | (Error) | (589) | (531.53) | |
| | Absent | Identification response (R) | 1 | 41.74** |
| Instructional bias (I) | | 1 | 0.00 | 0.00 |
| Foil Similarity (F) | | 1 | 2.94 | 0.07 |
| R × I | | | 0.25 | 0.02 |
| R × F | | | 0.26 | 0.02 |
| I × F | | | 0.44 | 0.03 |
| R × I × F | | | 0.06 | 0.01 |
| (Error) | | (591) | (438.30) | |

* $p < .05$ ** $p < .01$

Table 6

Means and Standard deviations for Confidence of Different Identification Responses in Each Condition for Target-Present and Target-Absent Waiter Lineups

| | | Identification Response: Target-Present Lineup | | | | | | | |
|----------------------|--|--|-----------|-------------|-----------|-------------|-----------|-------------|-----------|
| Instructional | | Correct | | Incorrect | | Not Present | | Overall | |
| Bias | | <u>Mean</u> | <u>SD</u> | <u>Mean</u> | <u>SD</u> | <u>Mean</u> | <u>SD</u> | <u>Mean</u> | <u>SD</u> |
| High Foil Similarity | | | | | | | | | |
| Unbiased | | 66.3 | 22.2 | 54.2 | 21.6 | 62.2 | 20.1 | 62.9 | 22.0 |
| Biased | | 70.2 | 20.2 | 53.8 | 20.5 | 67.4 | 26.6 | 64.6 | 22.3 |
| Overall | | 68.2 | 21.3 | 53.9 | 20.8 | 64.1 | 22.6 | 63.8 | 22.2 |
| Low Foil Similarity | | | | | | | | | |
| Unbiased | | 66.7 | 21.8 | 48.8 | 22.5 | 54.4 | 26.5 | 61.2 | 23.4 |
| Biased | | 71.4 | 20.1 | 54.1 | 23.5 | 69.4 | 21.4 | 67.8 | 21.8 |
| Overall | | 69.1 | 21.0 | 51.7 | 23.0 | 59.6 | 25.7 | 64.5 | 23.1 |
| Overall | | | | | | | | | |
| Unbiased | | 66.5 | 21.9 | 51.8 | 22.0 | 58.3 | 23.7 | 60.2 | 22.2 |
| Biased | | 70.9 | 20.1 | 53.9 | 21.5 | 68.3 | 24.0 | 59.6 | 22.7 |
| Overall | | 68.7 | 21.1 | 53.0 | 21.6 | 61.9 | 24.1 | 59.8 | 22.5 |

Table 6 (cont.)

| Identification Response: Target-Absent Lineup | | | | | | |
|---|-------------------|-----------|----------------------|-----------|-------------|-----------|
| Instructional Bias | Correct Rejection | | False Identification | | Total | |
| | <u>Mean</u> | <u>SD</u> | <u>Mean</u> | <u>SD</u> | <u>Mean</u> | <u>SD</u> |
| High Foil Similarity | | | | | | |
| Unbiased | 64.0 | 24.9 | 57.1 | 20.6 | 60.4 | 23.0 |
| Biased | 56.7 | 23.0 | 50.1 | 22.3 | 52.5 | 22.7 |
| Overall | 60.9 | 24.3 | 53.2 | 21.8 | 56.5 | 23.1 |
| Low Foil Similarity | | | | | | |
| Unbiased | 61.5 | 23.4 | 51.7 | 19.1 | 58.1 | 22.4 |
| Biased | 58.5 | 24.6 | 52.2 | 20.1 | 54.1 | 21.7 |
| Overall | 60.6 | 23.7 | 52.1 | 19.7 | 56.1 | 22.1 |
| Overall | | | | | | |
| Unbiased | 62.6 | 24.0 | 54.9 | 20.1 | 61.4 | 24.0 |
| Biased | 57.5 | 23.6 | 51.2 | 21.1 | 60.4 | 24.1 |
| Overall | 60.7 | 24.0 | 52.7 | 20.8 | 61.0 | 24.0 |

Table 7

ANOVA on Confidence Summary Statistics for Waiter Lineup

| Target presence | Source | df | F | f |
|------------------------|-----------------------------|-----------------------------|----------|---------|
| Present | Identification response (R) | 2 | 25.07** | 0.29 |
| | Instructional bias (I) | 1 | 6.98** | 0.11 |
| | Foil Similarity (F) | 1 | 0.53 | 0.03 |
| | R × I | 2 | 0.89 | 0.06 |
| | R × F | 2 | 0.43 | 0.04 |
| | I × F | 1 | 1.64 | 0.05 |
| | R × I × F | 2 | 0.48 | 0.04 |
| | (Error) | (587) | (470.42) | |
| | Absent | Identification response (R) | 1 | 15.29** |
| Instructional bias (I) | | 1 | 4.90** | 0.09 |
| Foil Similarity (F) | | 1 | 0.27 | 0.02 |
| R × I | | 1 | 0.26 | 0.02 |
| R × F | | 1 | 0.11 | 0.01 |
| I × F | | 1 | 2.36 | 0.06 |
| R × I × F | | 1 | 0.18 | 0.02 |
| (Error) | | (593) | (494.23) | |

** $p < .01$

Table 8

Calibration (C), Over/Underconfidence (O/U) and 95% Confidence Intervals (CI) for O/U, Normalized Resolution Index (NRI) and Point-Biserial Correlation (r) Statistics for Choosers and Nonchoosers from Thief and Waiter Lineups for Each Condition

| Choosing | Condition | Thief | | | | |
|-------------|--------------------|----------|------------|---------------|------------|----------|
| | | <u>C</u> | <u>O/U</u> | <u>95% CI</u> | <u>NRI</u> | <u>r</u> |
| Choosers | Instructional Bias | | | | | |
| | Unbiased | .016 | .066 | -.004 - .137 | .108 | .25** |
| | Biased | .029 | .130 | .069 - .190 | .124 | .36** |
| | Foil Similarity | | | | | |
| | High | .024 | .100 | .036 - .164 | .137 | .33** |
| | Low | .019 | .106 | .039 - .173 | .086 | .30** |
| | Overall | .020 | .103 | .057 - .149 | .107 | .32** |
| Nonchoosers | Instructional Bias | | | | | |
| | Unbiased | .021 | .063 | .013 - .113 | .045 | .20** |
| | Biased | .031 | .031 | -.026 - .089 | .070 | .17** |
| | Foil Similarity | | | | | |
| | High | .029 | .051 | -.004 - .105 | .035 | .16** |
| | Low | .023 | .048 | -.004 - .101 | .078 | .21** |
| | Overall | .021 | .049 | .012 - .087 | .043 | .19** |

Table 8 (cont.)

| Choosing | Condition | Waiter | | | | |
|-------------|--------------------|----------|------------|---------------|------------|----------|
| | | <u>C</u> | <u>O/U</u> | <u>95% CI</u> | <u>NRI</u> | <u>r</u> |
| Choosers | Instructional Bias | | | | | |
| | Unbiased | .014 | .036 | -.019 - .090 | .092 | .26** |
| | Biased | .017 | .123 | .078 - .168 | .178 | .43** |
| | Foil Similarity | | | | | |
| | High | .020 | .112 | .062 - .163 | .115 | .33** |
| | Low | .006 | .057 | .009 - .105 | .155 | .39** |
| | Overall | .012 | .084 | .049 - .119 | .132 | .36** |
| Nonchoosers | Instructional Bias | | | | | |
| | Unbiased | .051 | -.114 | -.176 - -.051 | .013 | .08 |
| | Biased | .127 | -.134 | -.225 - -.042 | .056 | -.20* |
| | Foil Similarity | | | | | |
| | High | .082 | -.097 | -.175 - -.020 | .017 | -.06 |
| | Low | .072 | -.143 | -.213 - -.073 | .011 | .02 |
| | Overall | .074 | -.121 | -.173 - -.069 | .002 | -.02 |

* $p < .05$ ** $p < .01$

Table 9

Frequency of Responses and Diagnosticity Ratios for the Different Confidence Categories for Choosers and Nonchoosers for Thief and Waiter Targets

| Target – Response | Confidence Level (%) | | | | | Overall |
|--------------------------|----------------------|-------|-------|-------|--------|---------|
| | 0-20 | 30-40 | 50-60 | 70-80 | 90-100 | |
| Thief – Choosers | | | | | | |
| Correct identification | 9 | 21 | 50 | 88 | 54 | 222 |
| Foil identification | 12 | 23 | 36 | 24 | 10 | 105 |
| False identification | 13 | 40 | 73 | 61 | 10 | 197 |
| Overall | 34 | 84 | 159 | 173 | 74 | 524 |
| Diagnosticity Ratio | 3.52 | 3.29 | 5.86 | 13.63 | 38.31 | 8.99 |
| Waiter – Choosers | | | | | | |
| Correct identification | 8 | 48 | 96 | 117 | 98 | 367 |
| Foil identification | 12 | 34 | 44 | 31 | 11 | 132 |
| False identification | 34 | 73 | 131 | 74 | 17 | 329 |
| Overall | 54 | 155 | 271 | 222 | 126 | 828 |
| Diagnosticity Ratio | 3.64 | 6.53 | 7.22 | 10.93 | 20.39 | 8.95 |

Table 9 (cont.)

| Target – Response | Confidence Level (%) | | | | | Overall |
|-----------------------------|----------------------|-------|-------|-------|--------|---------|
| | 0-20 | 30-40 | 50-60 | 70-80 | 90-100 | |
| Thief – Nonchoosers | | | | | | |
| Correct rejection | 15 | 32 | 110 | 161 | 84 | 402 |
| Incorrect rejection | 23 | 48 | 85 | 76 | 42 | 274 |
| Overall | 38 | 80 | 195 | 237 | 126 | 676 |
| Diagnosticity Ratio | 1.02 | 0.85 | 1.21 | 1.79 | 2.26 | 1.47 |
| Waiter – Nonchoosers | | | | | | |
| Correct rejection | 26 | 45 | 76 | 79 | 46 | 272 |
| Incorrect rejection | 11 | 13 | 28 | 29 | 19 | 100 |
| Overall | 37 | 58 | 104 | 108 | 65 | 372 |
| Diagnosticity Ratio | 1.22 | 2.79 | 2.20 | 3.15 | 4.92 | 2.71 |

Table 10

Calibration (C), Over/Underconfidence (O/U) and 95% Confidence Intervals (CI) for O/U, and Normalized Resolution Index (NRI) Statistics for Choosers and Nonchoosers Given Different Base Rates of Target-Absent Lineups

| Choosing | Target-Absent Base Rate | Thief | | | |
|-------------|-------------------------|----------|------------|---------------|------------|
| | | <u>C</u> | <u>O/U</u> | <u>95% CI</u> | <u>NRI</u> |
| Choosers | .50 | .020 | .103 | .057 - .149 | .107 |
| | .25 | .029 | -.123 | -.170 - -.075 | .099 |
| | .15 | .056 | -.177 | -.223 - -.131 | .054 |
| Nonchoosers | .50 | .021 | .049 | .012 - .087 | .043 |
| | .25 | .102 | .286 | .239 - .333 | .047 |
| | .15 | .202 | .399 | .351 - .447 | .023 |
| Waiter | | | | | |
| Choosers | .50 | .012 | .084 | .049 - .119 | .132 |
| | .25 | .020 | -.116 | -.153 - -.080 | .107 |
| | .15 | .055 | -.205 | -.237 - -.172 | .099 |
| Nonchoosers | .50 | .074 | -.121 | -.173 - -.069 | .002 |
| | .25 | .079 | .154 | .074 - .234 | .015 |
| | .15 | .153 | .281 | .193 - .370 | .014 |

Table 11

Number of Responses, Latency Range and Mean and Standard Deviation Latencies (in seconds) for the Three Identification Latency Groups for Choosers and Nonchoosers for Thief and Waiter Lineups

| | | Thief | | | |
|-------------|------------------------|----------|----------------------|-------------|-----------|
| Choosing | Identification Latency | <u>N</u> | Latency <u>Range</u> | <u>Mean</u> | <u>SD</u> |
| | | | | Latency | Latency |
| Choosers | Quick | 174 | 3.5-14.4 | 10.2 | 2.7 |
| | Medium | 174 | 14.5-23.7 | 19.0 | 2.6 |
| | Slow | 176 | 24.2-127.2 | 39.1 | 16.4 |
| Nonchoosers | Quick | 226 | 4.1-17.3 | 12.7 | 3.2 |
| | Medium | 224 | 17.4-27.2 | 21.8 | 2.9 |
| | Slow | 226 | 27.4-198.4 | 45.1 | 21.4 |
| | | Waiter | | | |
| Choosers | Quick | 273 | 0.4-10.0 | 7.0 | 1.8 |
| | Medium | 277 | 10.0-18.5 | 13.7 | 2.5 |
| | Slow | 278 | 18.7-104.1 | 31.3 | 13.4 |
| Nonchoosers | Quick | 122 | 4.1-13.5 | 9.5 | 2.3 |
| | Medium | 124 | 13.7-21.4 | 17.2 | 2.3 |
| | Slow | 126 | 21.7-147.1 | 35.0 | 18.5 |

Table 12

Calibration (C), Over/Underconfidence (O/U) and 95% Confidence Intervals (CI) for O/U, Normalized Resolution Index (NRI) Statistics, and Diagnosticity Ratios for Choosers and Nonchoosers For Different Identification Latencies

| Choosing | Identification Latency | Thief | | | | | Diagnosticity Ratio |
|-------------|------------------------|----------|------------|---------------|------------|-------|---------------------|
| | | <u>C</u> | <u>O/U</u> | <u>95% CI</u> | <u>NRI</u> | | |
| Choosers | Quick | .016 | .015 | -.057 - .087 | .108 | 17.45 | |
| | Medium | .031 | .153 | .071 - .235 | .080 | 7.77 | |
| | Slow | .038 | .146 | .062 - .230 | .032 | 5.39 | |
| Nonchoosers | Quick | .020 | .099 | .036 - .162 | .075 | 2.02 | |
| | Medium | .019 | .044 | -.022 - .109 | .057 | 1.42 | |
| | Slow | .037 | .005 | -.063 - .074 | .039 | 1.12 | |
| Waiter | | | | | | | |
| Choosers | Quick | .003 | .013 | -.040 - .067 | .118 | 13.83 | |
| | Medium | .026 | .118 | .053 - .182 | .033 | 8.36 | |
| | Slow | .029 | .124 | .061 - .187 | .055 | 6.41 | |
| Nonchoosers | Quick | .065 | .019 | -.073 - .112 | .025 | 4.20 | |
| | Medium | .048 | -.145 | -.226 - -.064 | .032 | 2.46 | |
| | Slow | .134 | -.237 | -.329 - -.145 | .066 | 1.97 | |

Table 13

Percentage of Correct (for Target-Present) and Incorrect (for Target-Present and -Absent)

Lineup Choices For Different Identification Latencies

| Target | Identification Latency | Lineup Choice | | |
|--------|------------------------|---------------|-------------------------------|------------------------------|
| | | Correct | Incorrect (Target-Present) | Incorrect (Target-Absent) |
| Thief | Quick | 60.0 | 16.6 | 23.4 |
| | Medium | 37.1 | 22.3 | 40.6 |
| | Slow | 29.9 | 21.3 | 48.9 |
| Waiter | Quick | 64.1 | 13.8 | 22.1 |
| | Medium | 39.9 | 15.6 | 44.6 |
| | Slow | 29.0 | 18.5 | 52.5 |

Figure Captions

Figure 1. Calibration curves for choosers from the thief and waiter lineups in the instructional bias and foil similarity conditions. Lighter line denotes perfect calibration. Error bars denote standard error of a proportion.

Figure 2. Calibration curves for nonchoosers from the thief and waiter lineups in the instructional bias and foil similarity conditions.

Figure 3. Calibration curves for choosers (upper panel) and nonchoosers (lower panel) from the thief and waiter lineups for different target-absent lineup base rates.

Figure 4. Calibration curves for choosers (upper panel) and nonchoosers (lower panel) from the thief and waiter lineups for different identification latencies.

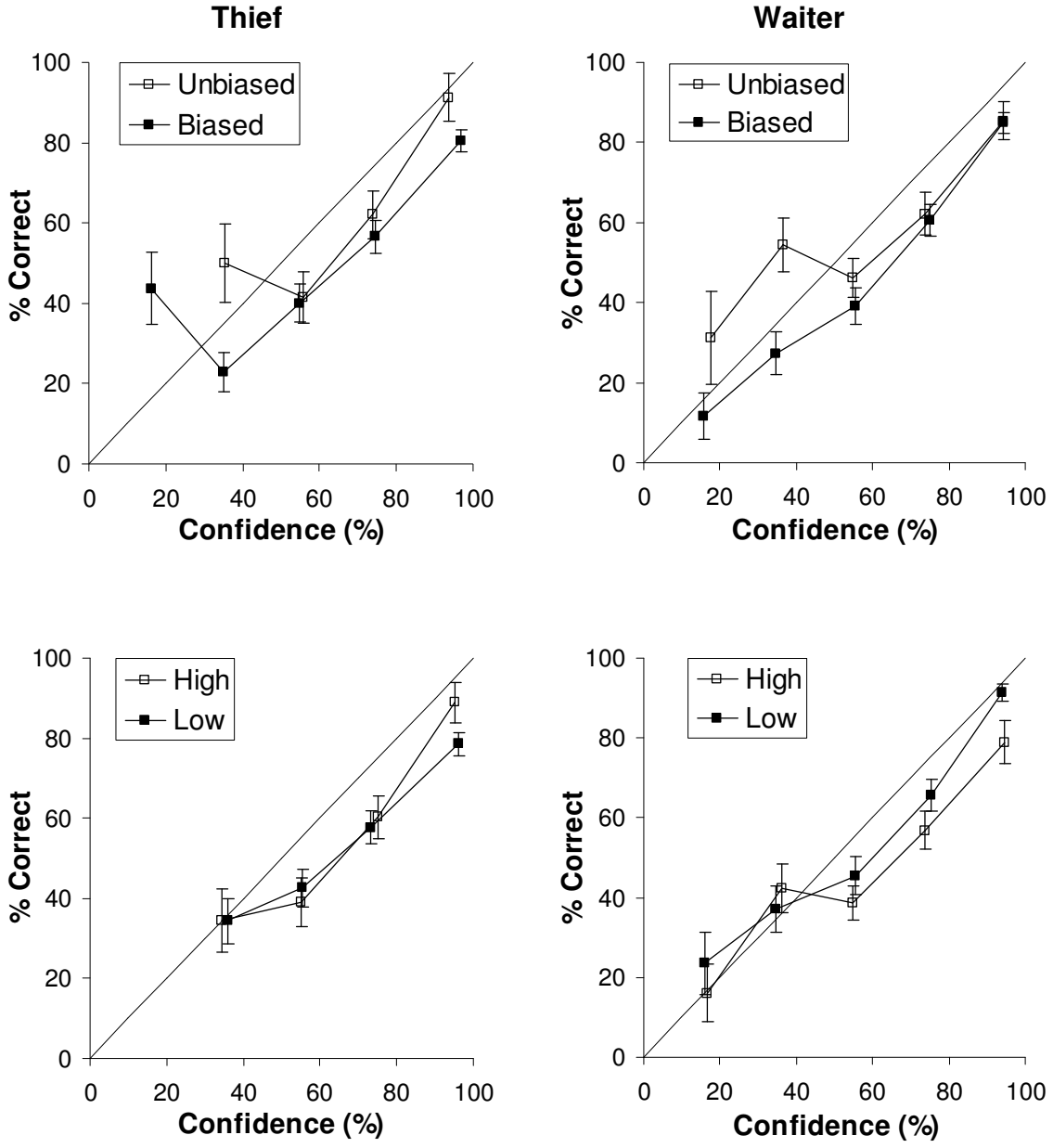


Figure 1

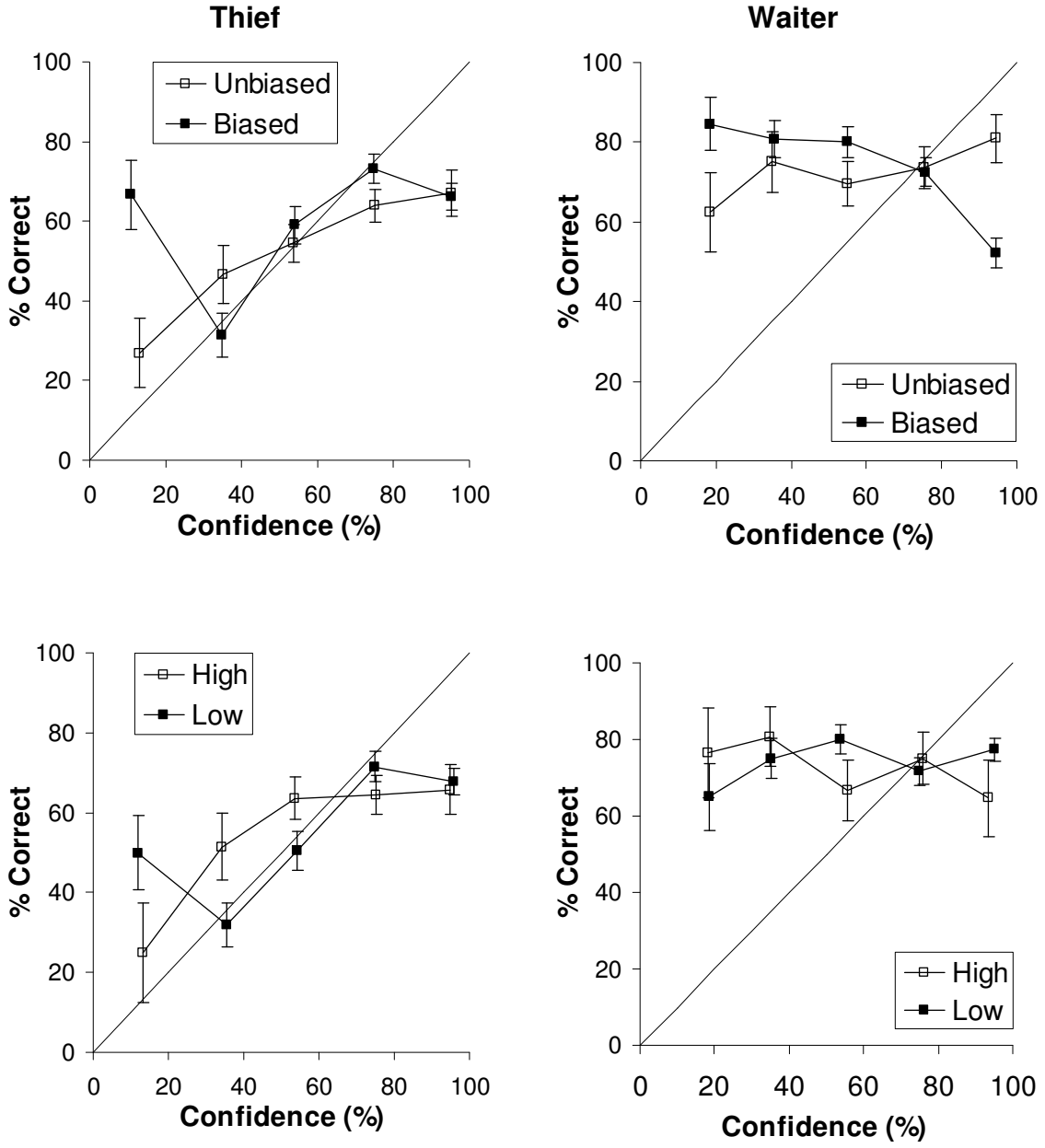


Figure 2

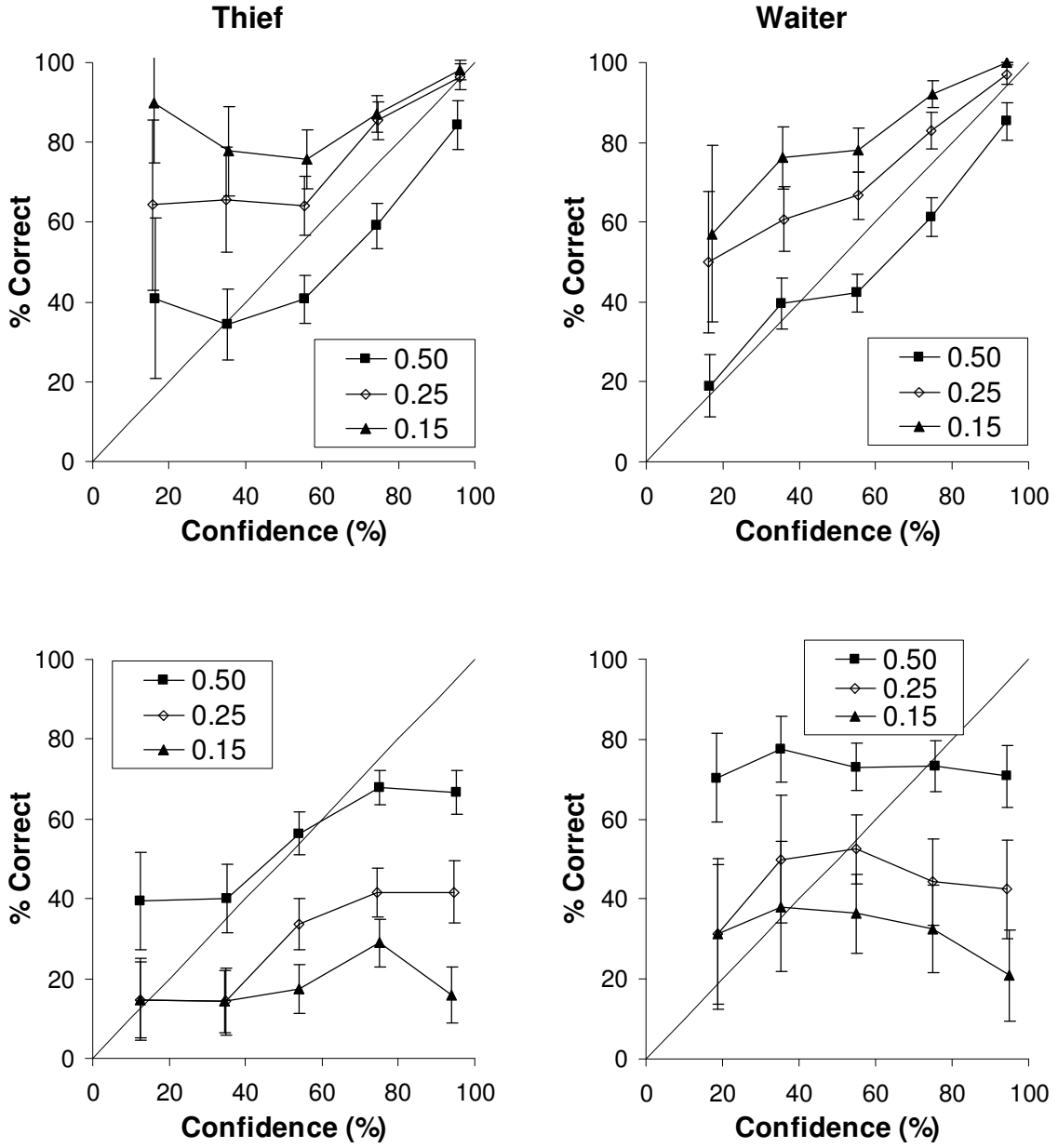


Figure 3

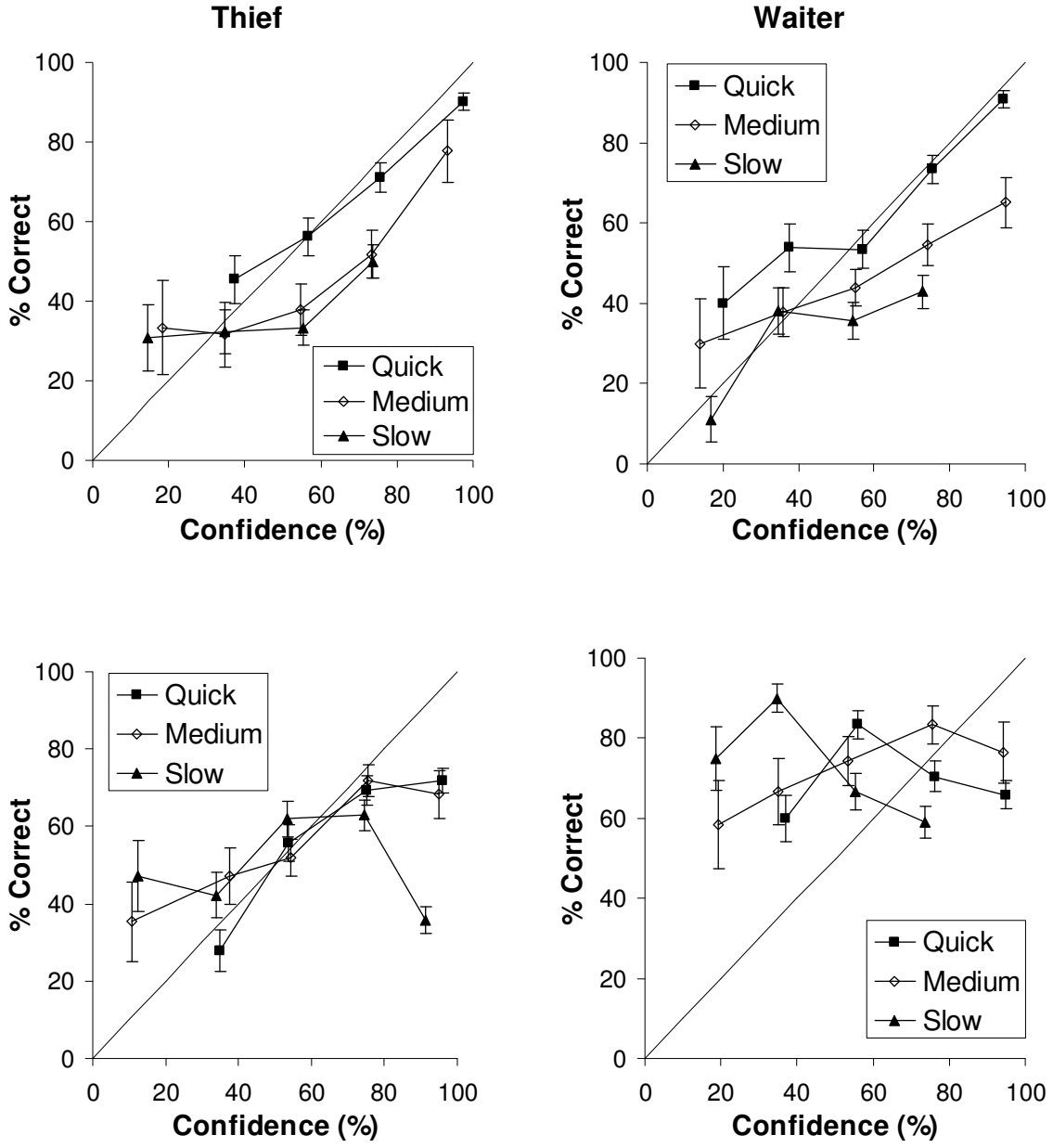


Figure 4