For A. Smith, M. Toglia, & J. Lampinen (2018). *Methods, measures, and theories in eyewitness identification tasks*. Taylor & Francis

Measuring the Relationship Between Eyewitness Identification Confidence and Accuracy

Neil Brewer and Carmen A. Lucas

Flinders University

James D. Sauer and Matthew A. Palmer

University of Tasmania

Supported by funding from Australian Research Council grant DP150101905 to N. Brewer et al.

Corresponding author: Neil Brewer
College of Education, Psychology and Social Work
Flinders University
GPO Box 2100, Adelaide
South Australia 5001
Email: neil.brewer@flinders.edu.au

Measuring the Relationship Between Eyewitness Identification Confidence and Accuracy

The relationship between confidence and accuracy for eyewitness identification decisions has been of interest for psychology and legal researchers, police and legal practitioners for many decades. The reason for this interest is easy to understand. For many crimes, a witness's positive identification of a suspect plays a crucial role in subsequent police investigations and legal proceedings. Yet, witnesses' eyewitness identification decisions are sometimes incorrect. The errors vary in nature, including identification of one of the known-innocent lineup fillers in a culprit-present or -absent lineup, rejection of a culprit-present lineup, and a mistaken identification of an innocent suspect from a culprit-absent lineup. Such errors are not isolated events, with meta-analyses of lab studies suggesting error rates as high as 50% (Steblay, Dysart, Fulero, & Lindsay, 2003; Steblay, Dysart, & Wells, 2011). Archival data, field studies, and US Innocence Project cases reinforce the propensity for witness error (see, for example, Horry, Halford, Brewer, Milne, & Bull, 2014; Innocence Project, 2018; Pike, Brace, & Kynan, 2002). These different types of errors may have adverse consequences, such as false convictions of innocent suspects, offenders evading apprehension, both of the aforementioned, and wasted investigative resources. Given the extent of these problems, researchers have been particularly interested in whether an easily obtainable measure such as the witness's confidence in their decision can provide reliable information about the accuracy of that decision.

Should identification confidence prove to be a reliable indicator of identification decision accuracy, it would simplify the investigative process by highlighting whether a suspect should be investigated more closely or alternative suspects considered. It could also assist judges and jurors as they weigh up the likely guilt or innocence of a defendant who has been positively identified (or not identified). In this chapter our primary focus is on how (a) the relationship between identification confidence and accuracy has been assessed, and (b) the format of this

assessment has shaped the nature of the relationship detected and, in turn, the sometimes strikingly different recommendations about the interpretation of identification confidence statements obtained from witnesses. First, however, we briefly highlight some of the contrasting perspectives advanced about the nature of the confidence-accuracy relationship.

Is There Consensus about the Confidence-Accuracy Relationship?

The enduring interest in the confidence-accuracy relationship has not led to a broad consensus about the nature and strength of the relationship. Perhaps unsurprisingly, the view of the non-research community appears to have aligned with what intuition would suggest: namely, that we might expect identifications made with high confidence to be accurate and those made hesitantly or unconfidently to be inaccurate. This perspective has been reflected in judicial judgments (e.g., Neil v. Biggers, 1972) and captured in surveys of police, lawyers and jurors (e.g., Deffenbacher & Loftus, 1982; Potter & Brewer, 1999).

This optimistic view of the confidence-accuracy relationship has not always been shared by eyewitness memory researchers—although, if one follows the timeline of research in this area over the last four decades or so, the impression gained is one of steadily growing optimism among eyewitness memory researchers about the strength of the confidence-accuracy relationship. Skeptical views about the usefulness of confidence for diagnosing accuracy can be found as far back as Münsterberg (1908) and were reinforced by a series of reviews published in the 1980s that emphasized the very weak correlations detected between identification confidence and accuracy (e.g., Bothwell, Deffenbacher, & Brigham, 1987; Cutler & Penrod, 1989; Wells & Murray, 1983). A much less pessimistic perspective on the strength of the relationship between the two variables was provided by a subsequent meta-analysis that highlighted how confidence may provide a more reliable indication of likely accuracy after allowing for moderator variables such as whether the witness chose from or rejected the lineup (Sporer, Penrod, Read, & Cutler,

1995). A more recent series of studies using a calibration approach—essentially a plot of accuracy at different levels of confidence—showed that, provided confidence is assessed immediately following the identification decision, a positive linear association between confidence and accuracy for positive identification decisions is evident under a variety of conditions (e.g., Brewer & Weber, 2008; Brewer & Wells, 2006; Palmer, Brewer, Weber, & Nagesh, 2013; Sauer, Brewer, Zweck, & Weber, 2010), although often accompanied by some degree of overconfidence depending on encoding and test conditions. And, even more recently, some researchers have endorsed an extremely optimistic perspective on the diagnostic value of identification confidence by concluding that, provided identification testing conditions are not procedurally biased in some way (i.e., conditions are what those researchers labeled as "pristine"), high confidence denotes high accuracy (Wixted & Wells, 2017).

Should we assume from this historical trend that there is now general agreement among eyewitness memory researchers that identification confidence is an extremely reliable indicator of accuracy? Although some readers of very recent reviews might arrive at that conclusion, we—despite being long-time advocates for the potential informativeness of identification confidence—take a somewhat more conservative position for reasons that will become apparent later in this chapter. We set the scene for that discussion by considering two issues: (i) Are there reasonable theoretical grounds for anticipating a strong and reliable (i.e., invariant) confidence-accuracy relationship? (ii) How can the discrepant conclusions that are so obvious in the literature be explained?

Should a Strong and Invariant Confidence-Accuracy Relationship be Expected?

As has been argued in a number of papers emanating from our laboratory, there are indeed sound theoretical and empirical grounds for expecting meaningful confidence-accuracy relationships (e.g., Brewer, 2006; Brewer & Weber, 2008; Brewer & Wells, 2006; Palmer et al.,

2013; Sauer et al., 2010). In various domains of human decision making, researchers have argued that common processes are likely to underlie confidence and accuracy. For example, perceptual discrimination theories have linked judgmental confidence to the strength of the information (commonly referred to by cognitive psychologists as "evidence strength")—a determinant of accuracy—favoring one decision over another (e.g.,Vickers, 1979). Of course, evidence strength is also a widely used term in the legal system, but here we are referring to "evidence" that is recruited by cognitive or neural mechanisms in support of a basic perceptual or memorial judgment. Evidence strength is considered a determinant of both confidence and accuracy of psychological judgments by signal detection theory (Green & Swets, 1966). And various models of recognition memory highlight the relations between memory strength, accuracy and confidence (e.g., Atkinson & Juola, 1974; Van Zandt, 2000; Yonelinas, 2002). Given the apparent parallels between those decision making domains and the eyewitness identification task, it is unsurprising that many psychologists would anticipate identification decision confidence and accuracy being strongly related.

Should we assume that this relationship is likely to be strong or invariant under all conditions? How individual researchers answer this question may depend on the extent to which they view the identification task as simply a recognition or discrimination task versus a task on which performance may also be shaped by social influence, metacognitive factors and individual difference variables. Those researchers who are more likely to acknowledge the influence of variables in the latter category are probably also more likely than others to expect that there will be conditions under which identification confidence and accuracy may become dissociated. It is probably fair to say that all eyewitness memory researchers acknowledge how post-identification social influence from lineup administrators or co-witnesses may profoundly inflate confidence judgments while not affecting accuracy (Steblay, Wells, & Douglass, 2014), thereby undermining

the confidence-accuracy relationship. It is this pattern of findings, of course, that underpins the recommendation that only a confidence judgment made immediately after the identification decision is likely to be a reliable diagnostic indicator (e.g., Brewer, Keast, & Rishworth, 2002; Brewer & Wells, 2006). Various other procedural factors—sometimes referred to as non-pristine test conditions (e.g., more than one suspect per lineup, a suspect who stands out, a failure to caution that the culprit may not be present, not using double blind testing and not obtaining the confidence statement at the time of testing)—that undermine the validity of confidence judgments have been identified (Wixted & Wells, 2017).

There are other factors that cannot be put down to non-pristine test conditions that might undermine the validity of a confidence assessment in individual cases and, at this stage, we simply do not know just how pervasive their influence might be. These factors have also been widely canvassed in previous work from our laboratory. They include, for example, overconfidence that may reflect the influence of misleading metacognitive cues, the tendency to seek only confirmatory information to support assessments of the likely accuracy of a judgment, and individual difference or dispositional factors (see, for example, discussions in Brewer, 2006; Brewer & Weber, 2008; Brewer & Wells, 2006; Palmer et al., 2013; Sauer et al., 2010). The point we wish to emphasize is that we currently know little about how prevalent the influence of such factors is likely to be in any individual case or about the likely extent of any impact. We return to what this means for the use of confidence judgments to assess guilt in individual cases later in this chapter.

Further grounds for speculating that the relationship might not always be strong are provided by findings in the broader psychological literature that the confidence-accuracy relation is affected by task difficulty, with overconfidence typically more marked in association with greater task difficulty (e.g., Gigerenzer, Hoffrage, & Kleinbölting, 1991). This pattern has also

been neatly demonstrated in studies of absolute and relative face recognition judgments using an old-new face recognition paradigm and a mini-lineup paradigm requiring recognition of previously studied photographs of faces from among groups of stimuli (Weber & Brewer, 2003, 2004). At present there are few studies in the literature that have probed the confidence-accuracy relation under what might be called challenging encoding and retention conditions, despite the fact that such conditions may not be uncommon in real life cases. We return to a consideration of the possible effects of such conditions later in the chapter.

In sum, we believe there are perfectly reasonable grounds for expecting a range of possible relations between identification confidence and accuracy as laboratory studies expand, rather than an invariant pattern—but only time will tell. We turn now to what the currently available data say about the relationship and how the message varies depending on the measurement approach taken.

Measuring the Confidence-Accuracy Relationship

In this section we review the main approaches used to assess the relationship between identification confidence and accuracy and summarize the major findings from each approach. As will become apparent, there is a close connection between researchers' measurement approaches and their conclusions about the nature of the relationship.

Point-Biserial Correlation

For many years the confidence-accuracy relationship was examined almost exclusively using a point-biserial correlation. To calculate a point-biserial correlation, identification accuracy is coded as 1 or 0 (accurate vs. inaccurate) and confidence can be measured on one of a variety of continuous scales (e.g., *1–7*; *1–10*; *0–100*). The correlation coefficient is interpreted in the same way as a standard bivariate correlation, with rough guidelines of .1, .3, and .5 suggested for interpreting small, medium, and large effects (Cohen, 1988).

As already noted, the eyewitness identification literature is replete with examples of weak confidence-accuracy correlations for identification decisions, with the consistency of those findings underpinning the once widely held conclusion that identification confidence is an uninformative indicator of likely accuracy. However, the interpretations based on much of the correlational data have been challenged by three different lines of argument and empirical findings. First, a meta-analysis that made a distinction between the performance of lineup choosers and non-choosers reported a weighted $r = .37$, 95% CI [.20, .55] for choosers whereas for non-choosers it was much smaller, $r = .12$, 95% CI [.07, .17] (Sporer et al., 1995), thereby indicating a meaningful relationship for positive identification decisions. Second, the common practice of employing invariant encoding and test stimuli in individual studies of the confidence-accuracy relation is likely to restrict the distribution of confidence scores in a way that is unlikely to be seen in everyday forensic contexts, and is likely to constrain the correlation (Juslin, Olsson, & Winman, 1996; Lindsay, Nilsen, & Read, 2000; Lindsay, Read, & Sharma, 1998). Indeed, when encoding conditions were allowed to vary, correlations extending to the .5–.6 range have been detected (Lindsay et al., 2000; Lindsay et al., 1998).

Third, and most important, the point-biserial coefficient provides only limited information about the confidence-accuracy relationship. Juslin et al. (1996) provided a cogent demonstration that a robust correlation between identification confidence and accuracy demands a level of discrimination from any sample of witnesses that is almost certainly unachievable. For example, a coefficient of 1.0 requires correct identifications all to be made with the same level of confidence, incorrect identifications all to be made with the same level of confidence, and the latter confidence level to be lower than the former. One consequence of this characteristic of the point-biserial coefficient is that (as will become apparent in the next section) it is quite possible that meaningful confidence-accuracy relationships may exist despite relatively low confidence-

accuracy correlations. Another limitation of this statistic is that the information it provides is not readily amenable to interpretation in the forensic setting because the coefficient does not speak directly to the likely accuracy of an identification response made with any particular level of confidence.

Confidence-Accuracy Calibration

Constructing and interpreting calibration curves. When Juslin et al. (1996) highlighted the limitations of the point-biserial correlation for understanding the identification confidence-accuracy relation, they advocated an alternative approach—referred to as confidence-accuracy calibration—that previously has been widely used in various other judgment and decision making domains (e.g., Cooke, 1906; Lichtenstein, Fischhoff, & Phillips, 1982). For identification decisions, the calibration approach first involves plotting the percentage of accurate decisions at each level of confidence: that is, charting the proportion of accurate decisions made with 100% confidence, 90% confidence, 80% confidence, etc.[1] If identification confidence and accuracy are perfectly calibrated, all decisions made with 100% confidence will be correct, 90% of decisions made with 90% confidence will be correct, and so on. The resulting graphical plot is usually the first port of call in assessing the relationship. Simply inspecting where the obtained calibration curve sits in relation to the ideal function indicating perfect calibration often provides a very useful guide as to the nature of the relationship. In the panels of Figure 1 we provide schematic representations of the various types of relationships described below. If accuracy increases systematically with increases in confidence it indicates that the two variables are (to some degree) calibrated. An obtained curve that very closely follows the ideal curve suggests that not only are the two variables calibrated but also that there is neither over- nor under-confidence. If points on the obtained curve fall below the ideal (e.g., accuracy for responses made with 90% confidence is 70%), over-confidence is indicated; if points are above the ideal (e.g., accuracy for

responses made with 70% confidence is 90%), under-confidence is suggested. If the obtained curve is flattish rather than sloped as per the ideal function (i.e., confidence does not provide strong discrimination between accurate and inaccurate responses), then resolution is relatively poor (and vice versa).

Figure 1 here

There are a number of important points to bear in mind when examining calibration curves for eyewitness identification decisions. First, given that the typical eyewitness identification study gathers one data point per participant, an obvious concern is the number of data points (and hence participants) required to generate what are likely to be relatively stable curves. Juslin et al. (1996) recommended around 200 data points per condition and our experience across many studies suggests this really should be an absolute minimum target. Note also that as the confidence-accuracy relationship is quite different for lineup choosers and nonchoosers—and hence separate calibration curves for each are mandatory—a minimum of 300–400 data points might be targeted per condition depending on the choosing rate associated with the particular encoding and test conditions. To generate more stable curves, data are often collapsed from 11 confidence categories (i.e., 0% to 100%) to five categories (i.e., 0%–20%, 30%–40%, 50%–60%, 70%–80%, 90%–100%) to increase the number of data points per category and "smooth" the curves (see, for example, Brewer et al., 2002; Brewer & Wells, 2006). When collapsing across confidence categories containing different numbers of observations, the weighted average confidence should be used in all plots and analyses as the confidence value for each new category.

Second, interpretation of calibration curves via visual inspection is facilitated by two additional pieces of information. To assess the likely stability of the calibration function it is informative to indicate the number of observations for each point on the curve, either in a

separate table or on the curve itself. This information allows the reader to assess how vulnerable any point on the curve might be to distortion if a few additional data points went in a different direction. Additionally, providing standard error bars at each confidence plot point assists in interpretation of the stability of the patterns, with overlapping standard error bars pointing to non-reliable differences between groups.

Third, following Brewer and Wells (2006), it has been common in calibration studies to remove filler identifications made from target-present lineups because, in single-suspect lineups, filler picks are already known to be picks of innocent lineup members.[2] In the absence of a designated innocent suspect in target-absent conditions, our calibration studies have typically adopted a "conservative" approach by counting any target-absent lineup pick as an error. This, of course, means that the degree of overconfidence observed in calibration curves will increase as the base rate of target-absent arrays increases, as illustrated by Brewer and Wells (2006). Our rationale for adopting this approach was to guard against possible overestimation of the ability of identification confidence to diagnose accuracy.

To demonstrate how such overestimation might occur we highlight one mechanism via which the diagnostic value of confidence may come to be overestimated. In an actual case, police should select their suspect based on descriptive information provided by the witness (likely combined with various other pieces of information they believe points to a particular suspect). Lineup fillers should be selected who also match the witness's description of the suspect (unless, of course, the suspect deviates from that description) and bear a reasonable degree of similarity to the suspect (who may or may not be the culprit). The lab situation is obviously radically different to an actual case in that there is no ongoing investigation culminating in the pinpointing of a suspect. Rather, the researchers know exactly who the culprit is. Consequently, lineup fillers in target-absent conditions of experimental studies are usually

selected based on similarity to the culprit. In real cases, however, the police conduct an investigation and pinpoint a suspect. This suspect may or may not be the culprit—but it is this suspect (in conjunction with the witness's description) who provides the focus for the construction of the lineup. Thus, if the suspect is not the culprit, the lineup fillers will be selected based on their similarity to the innocent suspect rather than to the culprit.

Clark and Tunnicliff (2001) have highlighted how this traditional experimental approach to target-absent lineup construction may underestimate the likelihood of mistaken identifications of innocent suspects which, in turn, can lead to an overestimation of the ability of confidence to diagnose accuracy. They found that (a) the false identification rate from target-absent arrays was considerably lower when fillers were selected to match the target rather than the innocent suspect (approximately 5% vs. 25%),[3] and (b) the conditional probability that an innocent suspect was identified when someone was picked from the array did not exceed chance when fillers were matched to the target, but reliably exceeded chance when fillers were matched to the suspect. Assuming at least some of the identifications of innocent suspects—had fillers been matched to the suspect instead of the target—might have been made with high confidence, accuracy at high confidence levels would be overestimated by matching fillers to the target. To our knowledge, this issue has not been allowed for in any of the prognostications about the reliability of high confidence identifications and yet it clearly demands further research.

Finally, an increasingly common practice in eyewitness identification research is to use multiple sets of encoding and test materials. There are two main reasons for following this practice. First, in the real world, encoding and test conditions clearly vary from case to case and often do so substantially. Second, it is reassuring if researchers can show that experimental findings are not constrained to some particular set of stimulus materials. One practice that calibration researchers who adopt this approach may find useful is to check that the resultant

calibration patterns are not driven by particular materials. Yet, because calibration studies require such large numbers of participants, conducting separate analyses for each set of materials might not be practical. One alternative approach that may still provide sufficient power is to repeat the analyses, each time removing data obtained via one set of stimulus materials, to ensure that the key patterns replicate across the remaining materials (cf. Palmer et al., 2013, p. 63). Another approach is to control for stimulus set within a mixed-effects statistical analysis (Baayen, Davidson, & Bates, 2008).

Descriptive statistics for a calibration approach. Visual inspection of calibration curves can be complemented by an examination of descriptive statistics that index calibration, over-/under-confidence and resolution (see, for example, Baranski & Petrusic, 1994; Lichtenstein et al., 1982). The calibration statistic (*C*) may range from 0 to 1, with 0 indicating perfect calibration. The over-/under-confidence statistic (*O/U*) may range from -1 to +1, with negative and positive values indicating under- and over-confidence, respectively. The method for computing these statistics is shown in Brewer and Wells (2006, p. 13). Resolution can be indexed by the adjusted normalized resolution index (*ANRI*) which can be interpreted in terms of proportion of variance explained (see Weber, Woodard, & Williamson, 2013, p. 155; Yaniv, Yates, & Smith, 1991). Values may range from 0 to 1, with the former indicating zero discrimination and the latter perfect discrimination.

To guide interpretation of these statistics, inferential 95% confidence intervals can permit testing for significant differences between experimental conditions (see Palmer et al., 2013; Sauer et al., 2010). On occasions, researchers have used experimental paradigms in which participants complete multiple trials rather than the typical single-trial identification test. One example is the use of an old-new face recognition paradigm in which participants study a series of faces and are subsequently asked to decide if each of a series of test faces was previously seen or not seen.

When such a paradigm is used, descriptive statistics such as those described above can be computed for each participant and the data analyzed using conventional parametric analyses (e.g., Weber & Brewer, 2003, 2004).

In some studies diagnosticity ratios (e.g., Brewer & Wells, 2006), or natural log diagnosticity ratios interpreted as log odds ratios (e.g., Palmer et al., 2013), have also been computed for lineup choosers and nonchoosers at each of the different confidence levels. A diagnosticity ratio of zero indicates that investigators should not adjust their prior estimates of the odds of suspect guilt on the basis of a witness's identification decision made with a certain level of confidence (i.e., a decision made with this level of confidence is not informative about the likely guilt of the suspect). Progressively higher positive diagnosticity ratios indicate that investigators should make progressive upward adjustments to their prior estimates of the odds of suspect guilt (or innocence, in the case of lineup rejection decisions) on the basis of the witness's identification decision.

Findings from calibration studies. Despite the very large sample sizes required to generate stable calibration curves using an eyewitness identification paradigm, a number of studies have now been reported in the literature. Here we provide a brief overview of the calibration findings indicating a robust confidence-accuracy relationship. Then we note findings indicating witness over-confidence despite indications of reasonable calibration. Finally, we summarize findings that clearly demonstrate poor calibration. Note that we only consider studies that have used an eyewitness identification paradigm: that is, studies in which a mock-witness viewed a filmed or live event and subsequently makes an identification decision from a lineup (photo-array or live). Not considered here are studies which used an old-new recognition paradigm or other studies in which a photo-array may have been used but the encoding stimulus was simply an identical photograph (or series of photographs) of the face(s) that appeared at test.

A positive linear association between confidence and accuracy has been demonstrated under a variety of memory encoding conditions: different stimulus exposure durations (e.g., Palmer et al., 2013), divided and full attention (Palmer et al., 2013), central versus peripheral target (Juslin et al., 1996), weapon present versus absent (Carlson, Dias, Weatherford, & Carlson, 2017), and cross- versus same-race identifications (Dobson & Dobolyi, 2016). Likewise, similar relationships have been observed under encoding-test retention intervals that vary in length (e.g., Palmer et al., 2013; Sauer et al., 2010). A number of aspects of lineup composition and presentation have also been associated with positive linear confidence-accuracy relations: high versus low similarity fillers (Brewer & Wells, 2006), target-absent base rate variations (Brewer & Wells, 2006), different levels of sequential lineup backloading (i.e., leading witnesses to believe they will see more lineup images than will actually be presented; Horry, Palmer, & Brewer, 2012), unbiased versus biased lineup instructions (Brewer & Wells, 2006), and numerical versus verbal confidence scales (Dobson & Dobolyi, 2016; Weber, Brewer, & Margitich, 2008).

The existence of positive linear confidence-accuracy relationships has not guaranteed, however, that calibration functions are not characterized by some degree of over-confidence (i.e., confidence levels are not matched by the corresponding accuracy levels). Data from many calibration studies indicate some degree of over-confidence (e.g., Brewer & Wells, 2006; Colloff, Wade, & Strange, 2016), with such a pattern particularly marked for long (cf. short) latency identifications (Brewer & Wells, 2006; Dodson & Dobolyi, 2016), cross-race identifications (Dodson & Dobolyi, 2016), and when the target-absent base rate is high rather than low (Brewer & Wells, 2006). We will return to this discussion of over-confidence later in the chapter.

A number of studies have shown that confidence and accuracy may sometimes be poorly calibrated. There is overwhelming support for the conclusion that confidence and accuracy are poorly calibrated for lineup non-choosers (e.g., Brewer & Wells, 2006; Palmer et al., 2013; Sauer

et al., 2010; Sauerland & Sporer, 2009) and for children (Keast, Brewer, & Wells, 2007). Delaying the witness's confidence judgment by as little as five minutes has also been associated with poor calibration. However, it appears that the relationship can be restored after such a delay if the witness is required to think deeply about the encoding and test conditions and how these conditions might be related to the likely accuracy of the identification decision (Brewer et al., 2002).

Confidence-Accuracy Characteristic (CAC) Analysis

Another approach to measuring the confidence-accuracy relationship was recently proposed by Mickes (2015). Labelled confidence-accuracy characteristic (CAC) analysis, this approach is—as Mickes (2015) noted—a close relative of the calibration approach, and has now been applied to a number of data sets. Most of the outcomes have been summarized in reviews by Wixted and his colleagues (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted & Wells, 2017). Mickes (2015) argued that a critical issue for judges and jurors who must evaluate identification evidence is the reliability of an eyewitness who has identified the suspect with a particular level of confidence. Assuming the witness was presented with a single-suspect lineup, any pick of a lineup filler is known to be a pick of an innocent person. Thus, she argued that what is of interest to triers of fact is whether a pick of the suspect is a pick of the culprit and, hence, the likelihood that the suspect is guilty when picked with a certain level of confidence. Consequently, unlike calibration analyses which, in most studies, exclude filler picks from target-present lineups but not from target-absent lineups, studies using the CAC approach have focused exclusively on picks of the culprit (target-present lineups) and the innocent suspect (target-absent lineups). Moreover, a primary focus for most researchers adopting the CAC approach has been on accuracy levels when identifications are made with very high confidence.

How then is the number of innocent suspect identifications from target-absent lineups determined, given that there is typically no actual innocent suspect as there may be in a police lineup? One approach is to designate one of the fillers as the innocent suspect. The other approach – and the one adopted for most of the CAC data reported – is to divide the number of filler identifications by the number of lineup members.

Constructing and interpreting CAC curves. Like calibration, the CAC approach provides a plot of accuracy at different confidence levels. Each point on a CAC curve represents suspect identification accuracy at a particular confidence level. The only responses of interest are identifications of the culprit from target-present lineups and identifications of the innocent suspect from target-absent lineups. At each confidence level the number of suspect identifications from target-present lineups is divided by the number of suspect identifications from target-present lineups + the number of innocent suspect identifications from target-absent lineups, with the result multiplied by 100. Thus, if 30 guilty suspect and 2 innocent suspect identifications were made with 100% confidence, accuracy at the 100% confidence level would be 93.75%. This process would be repeated at the 90% confidence level, and so on. Unlike calibration, it does not matter what confidence scale is used. It could be *100, 90, 80 … 0%*; *7, 6, 5 … 1*; *absolutely certain, maybe, unsure*; etc. Regardless of the scale, the data might be collapsed, for example, into two bins with the highest confidence level responses classified as high and all other responses classified as low. Or they might be collapsed into three bins such as high (*90-100%*), medium (*70-80%*), and low (*0-60%*). The classification system used has not mattered because the primary focus has been on the accuracy of suspect identifications made with very high confidence.

Findings from CAC studies. Wixted and Wells (2017) provide an overview of CAC analyses conducted on a number of data sets, most of which use a traditional eyewitness

identification paradigm. Many of these datasets were originally collected for studies using a calibration approach, with around half of them emanating from our laboratory. Thus, like the calibration data, the data encompass a range of experimental conditions including different stimulus exposure durations, divided and full attention, central versus peripheral targets, weapon present versus absent, cross- versus same-race identifications, different retention intervals, high versus low similarity lineup fillers, different levels of sequential lineup backloading, and unbiased versus biased lineup instructions. The data patterns from the studies reported are remarkably consistent:

> Analyses of suspect-ID accuracy show that for a wide range of base rates, high confidence implies high accuracy (with no sign that witnesses are overconfident) and low confidence implies much lower accuracy. This is true of both lab studies and police department field studies, so long as pristine testing conditions are used. … Under pristine testing conditions, a high-confidence suspect ID appears to be highly probative of guilt. Ignoring that fact—as the legal system is increasingly inclined to do—only serves to inappropriately exonerate the guilty. At the same time, ignoring low confidence at the time of an initial ID inappropriately imperils the innocent. The take-home message is that initial eyewitness confidence obtained from a pristine eyewitness-identification procedure serves both of the fundamental goals of the criminal justice system: to clear the innocent and to convict the guilty. (Wixted & Wells, 2017, pp. 49–50)

As the above quotations indicate, Wixted and Wells (2017) imposed the qualifier that the lineup conditions should be pristine: that is, a single suspect who should not stand out, cautioning the witness that the culprit may not be present, double blind testing and obtaining the confidence statement at the time of testing. Elsewhere in their manuscript, they reinforced a point, first noted by Brewer and Wells (2006), that high confidence identifications may be less reliable if the base

rate of target-present lineups were low. Nevertheless, the conclusions about the diagnostic value of high confidence identifications based on CAC analyses are clearly somewhat stronger than the conclusions that emanated from the earlier calibration studies—with the take-away message (as signaled in the above quotation) that a high confidence identification is a basis for conviction. We question, however, whether such a strong conclusion is justified based on our current knowledge.

Recently, Sauer, Palmer, and Brewer (2018) conducted CAC analyses on several data sets from published studies for which CAC analyses have not previously been reported. Their study was neither a meta-analysis nor a comprehensive review. The studies were chosen simply because the data provide striking exceptions to the conclusions of Wixted and Wells (2017), despite being collected using stimulus materials constructed to ensure that all lineup fillers provided strong matches to description and in most cases accompanied by demonstrations that their "fair" lineup conditions were characterized by high functional size. In other words, Sauer et al. argued that the lineups used in those studies could easily be the sort of lineups that police might construct in real cases. Yet, Sauer et al. showed how the reliability of very high confidence identifications broke down, sometimes very badly (e.g., between 4 and 8 out of 10 highly confident identifications in error), under conditions where an innocent suspect proved to be a particularly plausible choice for a witness or the identification task proved to be difficult. Other researchers who endorse the very strong conclusions sometimes made on the basis of CAC data will almost certainly argue that, despite the systematic lineup construction procedures followed in these studies (especially if we were to compare them with what is likely to happen in actual cases), the lineups were obviously not fair or perhaps the task (due to some combination of encoding and/or test conditions) was inordinately difficult. We deal with these issues in the next section.

Limitations of the Confidence-Accuracy Calibration and CAC data

What should be the takeaway message from the studies using calibration and CAC approaches? In 2006, Brewer and Wells concluded:

> 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. However, 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. (p. 25)

This conclusion is now buttressed by a substantial literature capturing an array of experimental conditions that indicates that identification confidence can provide a useful guide to investigators regarding the merits of following up a particular suspect if the confidence judgment was related to a positive identification decision, was obtained immediately after the identification, and was obtained from an adult witness. As already noted, studies using CAC analyses have drawn somewhat stronger conclusions about the likely accuracy of high confidence identifications (e.g., Wixted & Wells, 2017). Unless we have misinterpreted the tenor of those conclusions—and the quotation in the preceding section suggests we have not—we suspect that many readers would interpret those findings as indicating that an identification of a suspect made with very high confidence is tantamount to saying the suspect is guilty provided the testing conditions were pristine. We do not subscribe to such a strong conclusion for five main reasons, the last of which we view as the real "kicker".

First, in the section on constructing and interpreting calibration curves, we noted Clark and Tunnicliff's (2001) research showing how the practice of matching target-absent lineup fillers to the target rather than the innocent suspect—a practice that has been almost ubiquitous in

identification studies—deflates the rate of false identifications and very likely underestimates the prevalence of high confidence mistaken identifications.

Second, we believe that the available data fall short in terms of identifying possible boundary conditions for drawing the very strong conclusions. The number and nature of calibration and CAC studies conducted is limited. Although calibration and/or CAC analyses have been examined under a variety of forensically relevant conditions (e.g., long vs. short exposure duration, full vs. divided attention, short vs. long viewing distance, short vs. long retention interval, no weapon vs. weapon focus), the impact of most of the experimental manipulations on identification accuracy or discriminability has generally been relatively modest in magnitude (see Sauer et al., 2018) so it is difficult to know what the effects on, for example, accuracy of high confidence identifications are when discriminability is severely affected. There are convincing demonstrations in various domains of decision making that the confidence-accuracy relationship is undermined by increases in task difficulty because confidence judgments do not keep step with performance changes. For example, recognition memory studies involving old-new face recognition have shown a clear deterioration in the relationship as task difficulty (operationalized by encoding and test difficulty manipulations such as exposure time, decision type, other race stimuli) increased (e.g., Nguyen, Pezdek, & Wixted, 2016; Weber & Brewer, 2004). Identification research shows that over-confidence is much more marked for positive identifications decisions made by witnesses who take a long time to make a decision than for those whose identifications are made rapidly (Brewer & Wells, 2006). The determinants of slow decisions may be varied but it is very likely that, at least in part, they reflect difficulties in matching the lineup stimuli to memory.

We believe that we have a lot to learn about how variables that provide significant challenges for effective encoding and/or retention will affect the relationship and, especially, how

the accuracy of high confidence identifications is affected. Consider just a few examples. Relationships have been examined for retention intervals of two to three weeks (Palmer et al., 2013; Sauer et al., 2010). But what will happen, for example, when retention intervals extend to several months or a year or two as can happen in actual cases? Perhaps, witnesses will recognize deficiencies in their memory and either reject the lineup or, if they feel pressured to make a positive decision, respond with low confidence. Using CAC analysis, Wixted, Read, and Lindsay (2016) clearly favored this view when, based on remarkably high accuracy at high confidence levels for retention intervals of 3-9 months, they drew the extremely strong conclusion: "Thus, based on the available evidence, it seems reasonable to suppose that the findings reported here may be generally applicable" (p. 199). They may well prove to be right, but we would certainly prefer further data on the issue, while emphasizing several features of their study: their conclusions were based on data obtained from puny sample sizes using a single encoding stimulus and lineup under what appear to be very favorable encoding conditions.

In a similar vein to Wixted et al. (2016), again using a CAC approach, Semmler, Dunn, Mickes, & Wixted (2018 online first) —based on a re-analysis of identification data reported by Lindsay, Semmler, Weber, Brewer, & Lindsay (2008) for different viewing distances— concluded that "The evidence suggests that though suboptimal estimator variables decrease discriminability (i.e., the ability to distinguish innocent from guilty suspects), they do not decrease the reliability of identifications made with high confidence" (p. 1). Yet, even in their own study they could only speculate as to why this conclusion only held for one of the three judgment conditions used.

Or what will happen when witnesses are confronted with all the attentional demands associated with the hurly-burly of some real-life scenarios? Will witnesses respond much like those witnesses in Palmer et al.'s (2013) attentional manipulation conditions—where divided

attention involved making one of two different keyboard presses in response to intermittent high or low tones while viewing a crime video—and either reject the lineup or respond with low confidence when they become aware of possible memory deficiencies? Again, who knows? In sum, to date studies have not pushed the limits on implementing very challenging encoding and retention conditions.

Third, as we intimated earlier in the chapter, at present we know little about the influence of a number of factors that could conceivably give rise to high confidence judgments in individual cases. For example, we know little about overconfidence reflecting the influence of misleading metacognitive cues, the tendency to seek only confirmatory information to support assessments of the likely accuracy of a decision, or individual difference or dispositional factors (see, for example, discussions in Brewer, 2006; Brewer & Weber, 2008; Brewer & Wells, 2006; Palmer et al., 2013; Sauer et al., 2010). Large sample studies such as those already reviewed likely capture considerable variance in each of these areas, so the aggregated data from such studies suggest that such factors are probably not very important. But this may not apply at the level of an individual case.

Fourth, in the preceding section on CAC analyses we highlighted the findings reported by Sauer et al. (2018) that show how the relationship can break down badly when, despite conscientious attempts at lineup construction resulting in high functional size arrays, the innocent suspect appears particularly plausible to a witness or the task is very difficult for some unknown reason. We really have little idea about how frequently such conditions might prevail in real life individual cases. Nor do we know anything about possible interactions between such conditions and other variables such as very long retention intervals, the perceived consequences associated with making a decision in relation to a heinous crime, and so on. For example, might an innocent suspect in a heinous rape and murder seem more plausible six months after the crime than they

might have a few days post-crime? The answer to such questions involves considerable guesswork. Consequently, we remind readers of a comment from Wixted and Wells (2017) when discussing the findings of Colloff et al. (2016), a comment that (because of the sheer size of their manuscript) we fear many readers might gloss over:

> … high-confidence accuracy in the fair condition was noticeably lower than the 95% correct levels of accuracy typically observed in the other studies reviewed here. Because there is no obvious reason for the observed difference, this result serves as a reminder that the determinants of high-confidence accuracy are not fully understood and that more research is needed to identify the conditions under which high-confidence accuracy can be compromised even when fair lineups are used. (p. 38)

Finally, and most important, we find ourselves at odds with some of the interpretations that have emerged from the body of work using CAC analyses. In the most comprehensive review of the work, Wixted and Wells (2017) make the point that high confidence will only mean high accuracy under pristine conditions, one of which is that "The suspect should not stand out in the lineup" (p. 20). But they also acknowledge that the criteria for creating fair lineups are poorly specified: "But there is a need to articulate more precisely what the criteria should be for making lineups fair. What tools can be developed for officers who are tasked with creating a lineup to make their job easier and more objective?" (p. 54). And, in the same paper, precisely what constitutes an unfair or biased lineup looks a little bit like a moving target. Indeed, the underlying premise appears to be that, even if all lineup members strongly match description, and the associated functional size index is high, the lineup will be deemed unfair if an innocent suspect in the lineup is picked more often than fillers: "In any given study, it might be the case that, by chance, the replacement filler was chosen more often than the other fillers. If so, a

conclusion derived from that study alone would apply more to unfair lineups than to fair lineups." (Wixted & Wells, 2017, p. 25).

This approach seems eminently reasonable to us from the perspective of understanding the response patterns in individual experiments. But it is unhelpful from the perspective of deciding whether a confidence judgment *in an individual case* speaks convincingly about the suspect's guilt or innocence. We make three observations about this issue. First, we acknowledge that witness descriptions will sometimes be limited and thus allow for great variation between lineup members meeting that description. But it seems that what constitutes unfair or biased—leaving aside obvious cases of low functional size lineups that can be detected by a knowledgeable researcher's visual inspection of the lineup—is a post hoc judgment based on the fact that the pattern of target-absent lineup selections indicates that a particular innocent suspect appears especially plausible to witnesses for unknown reasons. Second, how then could we ever know in any individual case that a police lineup of *apparent* high functional size (i.e., a judgment based on some expert's visual comparison of the lineup with the witness's description) may have been unfair or fair—that is, that an innocent suspect may or may not have appeared particularly plausible to a witness? Is it reasonable to expect that police can somehow reconstruct the encoding conditions at the crime and conduct an experiment to see if a suspect is picked more than other lineup members before finalizing their lineup? At present, eyewitness researchers are not even able to operationalize how similar lineup fillers should be to the suspect or provide an objective index of similarity. Third, based on the authors' descriptions of the procedures employed in lineup construction, the procedures used in the studies singled out for illustrative purposes by Sauer et al. (2018) strike us as being thoughtful and meticulous, despite the fact that some lineup member ended up appearing particularly plausible. It is easy to criticize the adequacy of those lineups, but could we ever really imagine that police would follow such

elaborate procedures? And, in turn, could we imagine, even if they produced lineups of high functional size, that it would be guaranteed that an innocent suspect was not more plausible than any other lineup member? Consequently, we believe it is extremely important that researchers do not push to one side those datasets or conditions that are characterized by high functional size but do not conform to the broader confidence-accuracy pattern. We believe this is extremely important because (a) although the lineups may be unfair according to the Wixted and Wells (2017) criteria, they may be exactly the sort of lineups that the best intentioned and trained police may be at risk of producing, and (b) the data sets on which we currently rely are likely derived from a very limited sampling of the encoding and test conditions likely to prevail in real crimes and lineups.

Conclusion

The overall body of research obtained from calibration and CAC approaches reinforces our belief that the (previously quoted) conclusion based on one of the earliest calibration studies conducted in our lab is at present the most appropriate one:

> 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. However, 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. (Brewer & Wells, 2006, p. 25)

In other words, we believe that given existing data it is fine to advise police that, if a highly confident identification has been made under "pristine" conditions, they should look very hard for other evidence to make the case against their suspect. And if an identification of their suspect is made with very low confidence, the police might want to think again.

We are worried, however, by the much stronger tenor of the conclusions in reporting on the CAC studies. What specifically concerns us is that the impression that we believe is created—namely, that psychological science can almost guarantee that a highly confident ID in an individual case implies guilt. Let us be clear. It would be wonderful if this conclusion eventually proves to be correct —it would certainly simplify the job of police and the courts and it would look neat from a theoretical perspective. But we are worried that there are big issues that need to be resolved before advocating such a strong conclusion: for example, what might make an innocent suspect plausible despite the lineup being carefully constructed, what effect does matching target-absent fillers to the target rather than to the suspect have on high confidence false identifications, and have the boundary conditions of the confidence-accuracy relationship been adequately probed, are all, in our view, important questions to answer. In sum, the tenor of some of the recent conclusions is such that we worry that non-scientists may misread what the science tells us. And we are worried about the prospect that a court's judgment about an individual suspect's guilt may end up riding on assertions that are supported by some strands of psychological research that might prove to be much less generalizable than they appear.

References

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (pp. 243–290). San Francisco, CA: Freeman.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. doi: 10.1016/j.jml.2007.12.005

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgements. *Perception & Psychophysics, 55*, 412–428. doi: 10.3758/BF03205299

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. doi: 10.1037/0021-9010.72.4.691

Brewer, N. (2006). Uses and abuses of eyewitness identification confidence. *Legal and Criminological Psychology, 11*, 3–23. doi: 10.1348/135532505X79672

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied*, *8*, 44–56. doi: 10.1037/1076-898X.8.1.44

Brewer, N., & Weber, N. (2008). Eyewitness confidence and latency: Indices of memory processes not just markers of accuracy. *Applied Cognitive Psychology, 22*, 827–840. doi: 10.1002/acp.1486

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, functional size and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*, 11–30. doi: 10.1037/1076-898X.12.1.11

Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2017). An investigation of the weapon focus effect and the confidence–accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition, 6*, 82–92. doi: 10.1016/j.jarmac.2016.04.001

Clark, S. E., & Tunnicliff, J. (2001). Selecting lineup foils in eyewitness identification experiments: Experimental control and real-world simulation. *Law and Human Behavior, 25*, 199–216. doi: 10.1023/A:1010753809988

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

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science*, *27*, 1227–1239. doi: 10.1177/0956797616655789

Cooke, E. (1906). Forecasts and verifications in Western Australia. *Monthly Weather Review*, *34*, 23–24. doi: 10.1175/1520-0493(1906)34<23:FAVIWA>2.0.CO;2

Cutler, B. L., & Penrod, S. (1989). Moderators of the confidence-accuracy correlation in face recognition. *Applied Cognitive Psychology, 3*, 95–107. doi: 10.1002/acp.2350030202

Deffenbacher, K. A., & Loftus, E. F. (1982). Do jurors share a common understanding concerning eyewitness behavior? *Law and Human Behavior, 6*, 15–30. doi: 10.1007/BF01049310

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology, 30*, 113–125. doi: 10.1002/acp.3178

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528. doi: 10.1037/0033-295X.98.4.506

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Horry, R., Halford, P., Brewer, N., Milne, R., & Bull, R. (2014). Archival analyses of eyewitness identification test outcomes: What can they tell us about eyewitness memory? *Law and Human Behavior*, *38*, 94–108. doi: 10.1037/lhb0000060

Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied*, *18*, 346–360. doi: 10.1037/a0029779

Innocence Project. (2018). Retrieved from http://www.innocenceproject.org

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. doi: 10.1037/0278-7393.22.5.1304

Keast, A., Brewer, N., & Wells, G. L. (2007). Children's metacognitive judgments in an eyewitness identification task. *Journal of Experimental Child Psychology*, *97*, 286–314. doi: 10.1016/j.jecp.2007.01.007

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, UK: 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. doi: 10.1023/A:1005504320565

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. doi: 10.1111/1467-9280.00041

Lindsay, R. C. L., Semmler, C., Weber, N., Brewer, N., & Lindsay, M. R. (2008). How variations in distance affect eyewitness reports and identification accuracy. *Law and Human Behavior, 32*, 526-535. doi 10.1007/s10979-008-9128-x

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*, 93–102. doi: 10.1016/j.jarmac.2015.01.003

Münsterberg, H. (1908). *On the witness stand.* Garden City, NY: Doubleday.

Neil v. Biggers, 409 U.S. 188 (1972).

Nguyen, T. B., Pezdek, K., & Wixted, J. T. (2017). Evidence for a confidence-accuracy relationship in memory for same- and cross-race faces. *The Quarterly Journal of Experimental Psychology*, *70*, 2518–2534. doi: 10.1080/17470218.2016.1246578

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention

interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55–71. doi: 10.1037/a0031602

Pike, G., Brace, N., & Kynan, S. (2002). The visual identification of suspects: Procedures and practice. London, UK: Policing and Reducing Crime Unit, Home Office Research, Development and Statistics Directorate. doi:10.1037/e665432007-001

Potter, R., & Brewer, N. (1999). Perceptions of witness behaviour accuracy relationships held by police, lawyers and jurors. *Psychiatry, Psychology and Law, 6*, 97–103. doi: 10.1080/13218719909524952

Sauer, J. D., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337–347. doi: 10.1007/s10979-009-9192-x

Sauer, J. D., Palmer, M. A., & Brewer (2018). What can eyewitness confidence really tell us about identification accuracy? Manuscript under revision.

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*, 46–62. doi: 10.1037/a0014560

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018, online). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied*. doi: 10.1037/xap0000157

Sporer, S. L., Penrod, S. D., Read, 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. doi: 10.1037/0033-2909.118.3.315

Steblay, N. K., Dysart, J. E., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *27*, 523–540. doi: 10.1023/A:1025438223608

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law*, *17*, 99–139. doi: 10.1037/a0021650

Steblay, N. K., Wells, G. L., & Douglass, A. B. (2014). The eyewitness post identification feedback effect 15 years later: Theoretical and policy implications. *Psychology, Public Policy, and Law, 20*, 1–18. doi: 10.1037/law0000001

Sučić, I., Tokić, D., & Ivešić, M. (2015). Field study of response accuracy and decision confidence with regard to lineup composition and lineup presentation. *Psychology, Crime & Law*, *21*, 798–819. doi: 10.1080/1068316X.2015.1054383

Van Zandt, T. (2000). Roc curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 582–600. doi: 10.1037/0278-7393.26.3.582

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. doi: 10.1037/0021-9010.88.3.490

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgements. *Journal of Experimental Psychology: Applied, 10*, 156–172. doi: 10.1037/1076-898X.10.3.156

Weber, N., Brewer, N., & Margitich, S. (2008). The confidence-accuracy relation in eyewitness identification: Effects of verbal versus numeric confidence scales. In K. H. Kiefer (Ed.), *Applied psychology research trends* (pp. 103–118). Hauppauge, NY: Nova Science Publishers.

Wells, G. L., & Murray, D. M. (1983). What can psychology say about the *Neil v. Biggers* criteria for judging eyewitness accuracy? *Journal of Applied Psychology, 68*, 347–362. doi: 10.1037/0021-9010.68.3.347

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L., III. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*, 515–526. doi: 10.1037/a0039510

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence-accuracy relationship. *Journal of Applied Research in Memory and Cognition, 5*, 192–203. doi: 10.1016/j.jarmac.2016.04.006

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10–65. doi: 10.1177/1529100616686966

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*, 611–617. DOI 10.1037/0033-2909.110.3.611

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517. doi: 10.1006/jmla.2002.2864

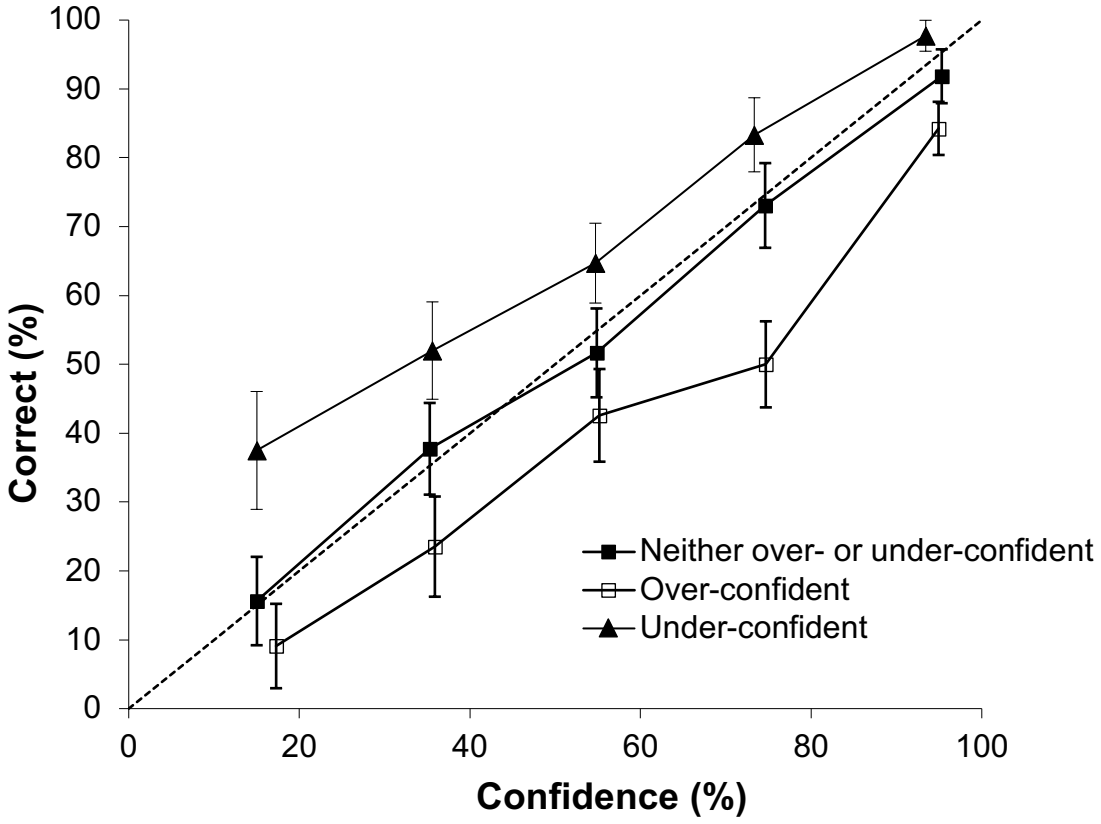
Footnotes

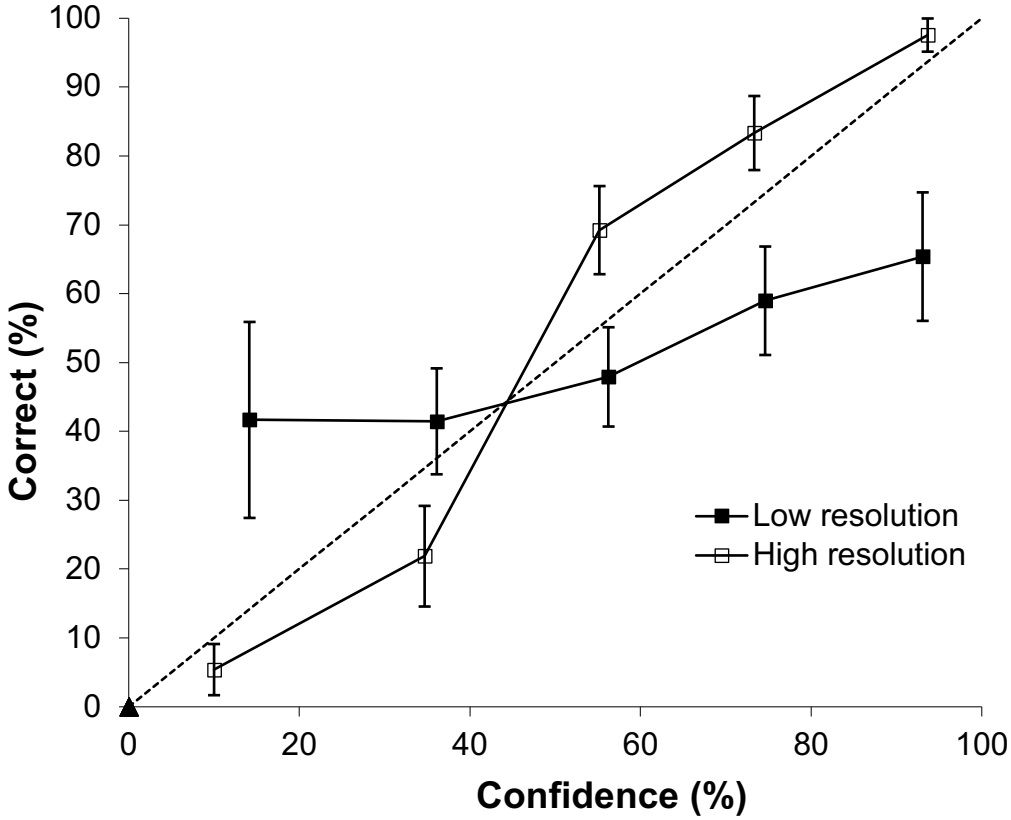
[1]In many previous applications, accuracy (proportion correct) has been plotted against subjective probability of being correct, which may or may not be the same as confidence.

[2]It is not always clear whether filler selections were included in calibration analyses (e.g., Sučić, Tokić, & Ivešić, 2015).

[3]At first glance, findings from Oriet and Fitzgerald (2018) suggest the opposite: namely, lower rates of false identifications from suspect-matched arrays. Note, however, that although this pattern was obvious in their Experiment 2 when the innocent suspect was seldom chosen, in Experiment 3 this pattern was far less apparent when the fillers were of very low similarity and was completely reversed when the fillers were of high similarity.

Figure 1. Example of over- versus under-confidence (upper panel) and low versus high resolution (lower panel). The calibration C-statistics for all curves are similar (.01–.03).


Correct (%)
Confidence (%)
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100
Neither over- or under-confident
Over-confident
Under-confident


Correct (%)
Confidence (%)
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100
Low resolution
High resolution