Pitfalls in using eyewitness confidence to diagnose the accuracy of an individual identification decision.

James D. Sauer & Matthew A. Palmer

University of Tasmania

Neil Brewer

Flinders University

Supported by funding from Australian Research Council grants DP150101905 to N. Brewer et al. and DP140103746 to M. Palmer et al. We thank Scott D. Gronlund, Stacy Wetmore, and Ines Sučić who allowed us to access and re-analyze the raw data from previously published manuscripts.

Corresponding author: James D. Sauer

Division of Psychology

University of Tasmania

Bag 30, Hobart

Tasmania. 7001

Australia

Email: Jim.Sauer@utas.edu.au

Abstract

Recently, a number of authors have made strong claims about the likely very high accuracy of identifications made with very high levels of confidence when identification testing conditions are pristine. We argue that although these strong claims about the confidence-accuracy relation are justifiable at the aggregate level, they may be misleading when attempting to evaluate the accuracy of an individual identification. First, we consider the recent evolution of conclusions drawn about the confidence-accuracy relationship, and the implications of these conclusions for the utility of confidence for evaluating individual identifications. Next, we highlight factors that may undermine the generalizability of conclusions at the aggregate level to individual cases. Finally, we present re-analyses of published data demonstrating conditions where conclusions based on aggregate data would be misleading for practitioners evaluating an individual identification. We maintain that, when appropriately collected, confidence can be a useful guide when assessing the reliability of identifications. However, we argue that when police and triers of fact attempt to evaluate the likely accuracy of an individual identification decision it will often be impossible to know if one of the key prerequisites for assessing whether a high confidence identification indicates an accurate identification—namely, a fair lineup—has been met.

Keywords: eyewitness identification, confidence, accuracy, metacognition, calibration, CAC

Pitfalls in using eyewitness confidence to diagnose the accuracy of an individual identification decision

Identification of a suspect by a witness has important implications for some police investigations and court proceedings. During the investigation, a suspect identification may confirm an investigator's hypothesis about the suspect's guilt. This may affect the direction of further investigative efforts, or lead the police to charge the suspect and initiate the trial process. At trial, jurors find identification evidence compelling (Semmler, Brewer, & Douglass, 2011). Thus, prosecutors will be keen to highlight the available identification evidence. However, eyewitness identification is prone to error (e.g., Wells et al., 1998). A witness identifying an innocent suspect potentially undermines investigative efforts by encouraging officers to devote time and resources to pursuing the wrong suspect, allows the culprit to remain undetected and, at trial, increases the risk of a wrongful conviction (e.g., Innocence Project, 2018). Thus, at both the investigative and trial stages, evaluating the reliability of an obtained identification before proceeding any further may help avoid undesirable outcomes. As a potential marker of identification accuracy, eyewitness confidence is highly researched, recommended in some judicial guidance (e.g., "Neil v. Biggers," 1972), intuitively plausible, and considered by police, lawyers, and mock jurors alike to be diagnostic (Deffenbacher & Loftus, 1982; Potter & Brewer, 1999). Therefore, a police officer or prosecutor might consider a witness's confidence as a useful starting point for evaluating the reliability of an obtained identification. However, if a police officer, prosecutor, or defense attorney consulted the research literature on the confidence-accuracy relationship for eyewitness identifications, or asked eyewitness memory researchers about that relationship, what would they learn about the diagnostic value of the witness's expressed level of confidence for the particular identification they are evaluating?

Until the early years of this century, consumers of the psychological literature would have been persuaded that a witness's confidence bore little relation to the accuracy of the identification decision. In contrast, readers of the most recent literature might conclude that confidence can provide near-definitive guidance about the accuracy of an individual identification if the lineup was conducted under pristine conditions (viz, only one suspect in the lineup, the suspect did not stand out, the witness was cautioned that the culprit may not be present, double blind testing was used, and the confidence statement was obtained at the time of testing; Wixted & Wells, 2017). Here, we make a case for a more nuanced interpretation of the available evidence. We make a key distinction between what the literature says about the confidence-accuracy relation in general, or at what some have termed the aggregate level (e.g., Wixted & Wells, 2017), and what the literature says about how confidence might be used to evaluate the likely accuracy of an individual identification decision. We will argue that this is a very important distinction. Knowing that there is robust evidence indicating that—provided appropriate procedures are followed in constructing and conducting lineups—an extremely confident identification is likely to be accurate should allow police to feel quite comfortable that they are charging the right person, prosecutors to proceed enthusiastically with a case against that suspect, and jurors to arrive at a guilty verdict without demure. But what if the existing evidential base is limited with respect to the probing of the boundary conditions of the confidence-accuracy relationship? Or, what if the nature of some aspect of lineup construction or conduct makes it virtually impossible to determine, in an individual case, if appropriate procedures had not been followed, either knowingly or unwittingly?

In the following sections we first provide a very brief historical perspective and overview of key findings from studies using confidence-accuracy calibration and confidence accuracy characteristic approaches. Second, we examine some important measurement issues and consider the extent to which research to date has clarified the boundary conditions of the

confidence-accuracy relationship. Third, we examine findings from a number of published studies that clearly contradict the conclusion that high confidence necessarily signals accuracy at the individual case level. These studies have two key features in common. One is that the researchers followed very systematic procedures when constructing their lineups to avoid creating an unfair lineup (i.e., a lineup biased against the guilty or innocent suspect). Indeed, the procedures followed were much more elaborate than would be expected from police in any individual case. The other is that, despite the careful lineup construction procedures followed, posthoc scrutiny of the data indicates that the lineups were indeed biased against the innocent suspect in the sense that the innocent suspect was more likely to be selected than any of the fillers. In other words, these studies illustrate how it is possible to create non-pristine lineup conditions inadvertently, despite using lineup construction procedures that follow best practice recommendations. Finally, we comment briefly on the recording of eyewitness confidence and the role witness confidence plays in the prosecution of cases.

A (Very) Brief Historical Perspective

In this overview we focus on conclusions drawn from studies in which confidence was recorded immediately after the identification and, thus, was not vulnerable to post-decisional social or metacognitive influences. Moreover, we only consider research in which the data analytic approach extended beyond the examination of confidence-accuracy correlations as the literature has since converged on the idea that (a) correlation is a sub-optimal measure of the confidence-accuracy relationship and (b) conclusions drawn from these studies typically underestimated the covariation between confidence and accuracy (e.g., Brewer & Wells, 2006; Juslin, Olsson, & Winman, 1996; Lindsay, Read, & Sharma, 1998)[1].

[1] However, we acknowledge the contribution made by Sporer, Penrod, Read, and Cutler's (1995) work in highlighting the importance of considering the relationship separately for chooser and non-choosers.

In the first large-scale study of confidence-accuracy calibration for identification decisions, Brewer & Wells (2006, p.11) concluded that "confidence assessments obtained immediately after a positive identification can provide a useful guide for investigators about the likely accuracy of an identification." Then, on p.25, the authors elaborate: "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". Sauer, Brewer, Zweck, and Weber (2010) offered a similar conclusion: "For choosers in both the delayed and immediate conditions, increased confidence was associated with increased probable accuracy."

In the 2012 *Pennsylvania Instructions*—a set of simplified instructions intended to help jurors interpret identification evidence—Elizabeth Loftus and colleagues wrote that although "…confident witnesses are somewhat more accurate than unconfident witnesses, scientific research shows that eyewitness's (sic) confidence generally <u>is not</u> a reliable indicator of accuracy" (Loftus, Francis, & Turgeon, 2012, emphasis in the original). This conclusion has particular relevance in the current context because, although it refers to the confidence-accuracy relation in general, it is intended to provide jurors with guidance on how to use confidence (or not, as the case may be) to evaluate an individual identification. These instructions reflect comments in *State v. Henderson* (2011) suggesting that "…eyewitness confidence is generally an unreliable indicator of accuracy…", but also acknowledging that "…highly confident witnesses can make accurate identifications 90% of the time."

Three years later, Wixted, Mickes, Clark, Gronlund, and Roediger (2015, p.524) concluded that "… low confidence implies low accuracy, and high confidence implies high accuracy". Similarly, Carlson, Dias, Weatherford, and Carlson (2017, p.88) asked "Can

identifications made by highly confident eyewitnesses (those most likely to make it to trial) be trusted? In other words, are these identifications highly accurate?" and answered: "they are." Wixted & Wells (2017) then qualified Wixted et al.'s (2015) claim by noting that "a low-confidence ID implies low accuracy, and a high-confidence ID implies high accuracy" when testing conditions are "pristine" (i.e., free from procedural bias, p.20). Semmler, Dunn, Mickes, and Wixted (2018) went a step further, proposing that "reliability for a given level of confidence is largely unaffected by estimator variables (p.3)."

The literature clearly converges on the position that, in general or at the aggregate level, confidence is related to identification accuracy: As witness confidence increases so, too, does the likely accuracy of the identification. However, these positions differ radically in terms of how confidence might be used to evaluate the likely guilt of a *particular* suspect-come-defendant: that is, whether very high confidence virtually guarantees accuracy for an individual identification. Below we provide a brief overview exploring how these divergent opinions emerged in the literature (noting that much of this information is covered in some form elsewhere in the literature; e.g., Brewer, 2006; Brewer & Weber, 2008; Brewer & Wells, 2006; Palmer, Brewer, Weber, & Nagesh, 2013; Sauer & Brewer, 2015; Sauer et al., 2010; Wixted et al., 2015). We then return to our key question: What pitfalls might officers, prosecutors, judges, and jurors encounter when using eyewitness confidence to diagnose the accuracy of an individual identification decision?

An Overview of Recent Key Findings

There is clear theoretical support for the existence of a meaningful confidence-accuracy relationship in recognition memory (e.g., Green & Swets, 1966; Macmillan & Creelman, 1991; Van Zandt, 2000), and the literature contains robust empirical demonstrations of meaningful (albeit imperfect) confidence-accuracy relations in a variety of basic memory and discrimination domains (e.g., Baranski & Petrusic, 1994; Björkman,

Juslin, & Winman, 1993; Kunimoto, Miller, & Pashler, 2001; Weber & Brewer, 2003, 2004). As Van Zandt (2000) noted, psychologists have been developing models to account for the empirical relationship between confidence, accuracy, and other behavioral variables across varied domains since the 1940s (e.g., Cartwright & Festinger, 1943; Festinger, 1943a, 1943b). But what do we know about the confidence-accuracy relationship for eyewitness recognition memory?

Confidence-accuracy calibration. Juslin et al. (1996) substantially advanced our understanding of the forensic utility of the confidence-accuracy relation when they applied calibration analysis—a tool commonly used in other judgment and decision-making domains to compare the objective and subjective probabilities of response accuracy—to eyewitness identification data. Calibration analysis speaks to both the linearity of the relation (the extent to which accuracy increases with confidence) and the realism of the relation (the degree of correspondence between the subjective [confidence] and objective [accuracy] probability of response accuracy across a sample of decisions). Thus, calibration provides useful information about the likely accuracy of a decision made with a certain level of confidence under a specific set of conditions and, *potentially*, across conditions. Juslin et al.'s work motivated a series of large-scale calibration studies demonstrating robust, positive relationships between the level of confidence expressed by the witness and the likely accuracy of the witness's identification, across a variety of theoretically- and forensically-relevant manipulations (e.g., instructional bias, filler similarity, retention interval, divided attention at encoding, and own- vs. other-race identification; Brewer & Wells, 2006; Carlson et al., 2016; Dodson & Dobolyi, 2016; Palmer et al., 2013; Sauer et al., 2010).

We should note, however, that the number of calibration studies in the eyewitness identification literature is still small (possibly reflecting the substantial time and resources

required to achieve adequate power for this form of analysis[2]). Further, results from calibration analyses highlight three important caveats associated with the conclusion that eyewitness confidence and accuracy are meaningfully related. First, this relationship seems to hold only for choosers (i.e., individuals who identify a lineup member). Second, for choosers, calibration analyses show that the confidence-accuracy relation is typically characterized by overconfidence in the upper half of the confidence scale (e.g., Brewer & Wells, 2006; Palmer et al., 2013; Sauer et al., 2010). Finally, the absolute level of accuracy associated with any given level of confidence varies between and within studies. Put another way, the confidence-accuracy relation is not fixed: The strength of the relationship is moderated by other estimator and system variables (see Brewer, 2006, for a review). To illustrate, consider the calibration curves presented by Brewer and Wells (2006). Absolute levels of accuracy for positive identifications associated with 90-100% confidence[3] vary from approx. 60% to approx. 95% depending on the similarity of the fillers to the suspect in the lineup, the target-absent base rate, the instructions given to the participant before they view the lineup, and the speed with which responses are made (see also Dodson & Dobolyi, 2016, for the moderating effects of response time on the confidence-accuracy relation). The important applied implication is that although confidence is meaningfully related to accuracy in a variety of conditions, the *absolute* level of accuracy associated with any level of confidence will presumably also vary in applied contexts according to the presence or absence of these (and other) moderators (i.e., consistent with demonstrations of the hard-easy—where overconfidence increases with task difficulty—effect in other domains; Gigerenzer, Hoffrage, & Kleinboelting, 1991; Juslin,

[2] There are, however, also studies using a more economical face recognition paradigm in which participants study, and make recognition decisions about, a series of faces. This multi-trial format allows for the calculation of relevant calibration statistics within-participants, and produces findings generally consistent with the studies using an eyewitness identification paradigm (e.g., Cutler & Penrod, 1989; Weber & Brewer, 2003, 2004).

[3] Here we consider only point estimates of accuracy. The plausible range of accuracy values at these intervals would obviously be even greater if we considered 95% confidence intervals around these point estimates.

Winman, & Olsson, 2000). Thus, although the calibration literature is consistent in demonstrating that confidence and accuracy are related across a sample of participants' decisions, the apparent inconsistency in absolute levels of accuracy associated with any given level of confidence precludes a strong conclusion, based on confidence, about likely accuracy of any particular identification decision.

Confidence-accuracy characteristics. The Confidence Accuracy Characteristic (CAC; Mickes, 2015) has recently been presented as a more generalized alternative to calibration. Like calibration, this approach examines variations in accuracy as a function of confidence. However, unlike calibration, it does not require confidence to be measured on a probabilistic scale (i.e., 50-100% or 0-100%): any ordinal confidence scale is sufficient. This approach is supported by Tekin and Roediger's (2017) finding that varied confidence scales produce comparable confidence-accuracy relations. Further, in cases where confidence is initially assessed on a 0-100% scale, published CACs typically plot confidence data collapsed into three "bins"[4]. For example, a 0-100% confidence scale might be reduced to scale with low (0-60%), moderate (70-80%), and high (90-100%) bins (e.g., Carlson et al., 2016; Mickes, 2015). In other cases, researchers have collapsed scales down to two confidence categories (e.g., reducing a five-point scale to low [ratings of 1-4] and high [ratings of 5] confidence; Wixted, Read, & Lindsay, 2016, see Figure 1D in the original). This approach has the benefit of being more flexible in assessing the confidence-accuracy relationship—allowing researchers to assess variations in accuracy as a function of the level of confidence expressed, regardless of the scale on which confidence was recorded—but it is not without limitations. For example, one cannot talk sensibly about the effects of relevant variables on over/underconfidence (the realism of confidence judgements) when confidence is not

[4] Initially, Mickes' (2015) collapsed confidence data to three bins to offset the relatively low number of data points in the lower confidence categories, and gain a more stable estimate of the relationship. This is also commonly done with calibration curves, with 11 point scales (0-100% with decile response options) often collapsed to 5 point scales.

recorded on a probabilistic scale. When the confidence scale is not probabilistic, the scale options have no a priori values. Therefore, one cannot draw any conclusions about realism (in an absolute sense in any given condition, or in terms of between-subjects differences across conditions). This limitation is important if one wishes to generalize conclusions about the predictive value of confidence, at any but the highest and lowest levels of confidence, across conditions (e.g., from the laboratory to applied settings).

A second important area of departure between the CAC and calibration approaches is that CACs plot only suspect identifications. Calibration curves (in the identification literature) typically omit target-present filler identifications (see, for example, Brewer & Wells, 2006) but include target-absent filler identifications (providing a conservative estimate of the realism of participants' confidence judgments)[5]. In contrast, CACs typically include only identifications of the target (from target-present lineups) and target-replacement (from target-absent lineups), excluding all filler identifications in an approach somewhat analogous to considering only suspect identifications in applied contexts (though, for some CAC research, the innocent suspect identification rate is estimated by dividing the total number of identifications by the number of lineup members; e.g., Seale-Carlisle & Mickes, 2016; Wilson, Seale-Carlisle, & Mickes, 2018).

Using the CAC approach, researchers have found results generally consistent with those of previous calibration analyses (i.e., for choosers, likely accuracy increases with confidence), but with one important departure: CAC curves show very high levels of accuracy (often over 90%, and sometimes approaching 100%) at the highest confidence level (e.g., Carlson et al., 2017; Mickes, 2015; Wixted et al., 2016). These accuracy rates are

[5] Carlson et al. (2017) adopted a different approach, estimating innocent suspect identification rates by dividing the total number of target-absent identification by the number of lineup members (i.e., six). This approach replicates the approach taken by Brewer and Wells (2006) when computing diagnosticity ratios for decisions made with different confidence levels.

notably higher than those typically reported in the relevant calibration literature (cf. Carlson et al., 2017). Based on CAC analyses, and some calibration analyses, researchers have recently concluded that "…participant-eyewitnesses who indicate 90–100% confidence tend to be accurate within that range as well" (Carlson et al., 2016, p.907), that "…low confidence implies low accuracy, and high confidence implies high accuracy…" (Wixted et al., 2015, p. 524), and that "…high-confidence suspect ID accuracy exceeds 95% correct…" (Wixted et al., 2016, p. 199). These statements obviously represent somewhat stronger conclusions, compared to those that have emerged from the calibration literature, about the relationship between a very high level of identification confidence and the likely accuracy of any individual identification.

Some Measurement Issues

Researchers reporting calibration analyses have typically limited their conclusions to (a) identifying a meaningful, but typically overconfident, positive relationship between confidence and accuracy, and (b) suggesting that confidence may therefore be informative when assessing the reliability of an identification. Because the large sample calibration data that exist typically include all target-absent picks, some proportion of which will obviously be known-innocent filler picks, conclusions about likely suspect guilt in an individual case need to be somewhat coarse-grained in nature. For example, witnesses who make a positive identification with 90-100% confidence are likely to be accurate 60-95% of the time. In contrast, conclusions based on CAC analyses appear to offer decision-makers in the criminal justice system less ambiguous, or more fine-grained, guidance for evaluating an individual identification: A high-confidence identification is probably correct and a low-confidence identification is probably incorrect. Are these more definitive conclusions warranted and, by extension, is the apparently simple message to be gleaned by investigators, legal practitioners and jurors justified? We argue not. Although they may reflect the confidence-accuracy

relation at the aggregate level, they do not provide sufficient guidance for a decision-maker seeking to evaluate an individual identification. Three important limitations on our current knowledge preclude such definitive conclusions, even when the lineup is conducted under pristine conditions. First, there are limits on the methods commonly used to estimate accuracy when assessing the confidence-accuracy relationship. Second, there has been limited exploration of the boundary conditions for the confidence-accuracy relationship. Third—and by far the most critical limitation—based on our reanalysis of four published datasets (see below), we show that when boundary conditions change in certain ways, the confidence-accuracy relationship breaks down badly. In such cases, high confidence no longer indicates a high likelihood of accuracy. Critically, however, these cases—and the consequent breakdown of the confidence-accuracy relationship—(a) occurred despite the researchers following (current) best practice recommendations for constructing lineups that almost certainly would never be matched by even the most conscientious police lineup administrators, and (b) cannot always be identified in advance. Thus, a decision-maker will be unaware that the "high-confidence, high-accuracy" proposition is not applicable for the particular identification in question.

Estimating Accuracy

As identified previously, there is an important difference in the calculation of accuracy rates for the calibration and CAC approaches. Although this issue may seem to simply reflect an analytical preference, it has potentially significant implications when attempting to draw generalizable conclusions about the absolute levels of accuracy associated with different levels of confidence. As a reminder, in contrast to the typical calibration approach which excludes filler identifications from target-present lineups but retains filler

identifications from target-absent lineups[6] (e.g., Brewer & Wells, 2006; Palmer et al., 2013; Sauer et al., 2010), CAC analyses consider only suspect identifications from target-absent lineups, reducing the overall number of errors included in the calculation and increasing accuracy. Although approaches that ignore all filler identifications obviously constrain theoretical understanding of confidence judgments, the decision to focus only on suspect identifications has intuitive appeal, and provides a clear analogue to the primary issue of concern when evaluating identifications in applied settings: Given the suspect has been identified, how likely is it that the identification is correct and the suspect is guilty? From this purely applied perspective, it makes sense to consider only suspect identifications when assessing the confidence-accuracy relationship (i.e., to adopt the CAC approach). However, despite providing a neat analogue of the applied situation, attempting to draw general conclusions about the confidence-accuracy relationship based only on suspect identifications in lab settings may be problematic. To understand why, we need to consider how researchers typically construct lineups in experimental settings.

Typically, when constructing lineups in research settings, experimenters select fillers for their target from a pool of photographs based on some combination of the potential fillers' match to the description of the target (often obtained from pilot participants or lab co-inhabitants) and their general level of physical similarity to the target (determined by obtaining similarity ratings from pilot participants, or a highly scientific "eye-ball test"). The target is placed among a number of plausible fillers to create a target-present lineup. For maximal experimental control, target-absent lineups typically use the "same fillers design"; replacing the target with a designated "innocent suspect" selected from the original pool of potential fillers, and often bearing a relatively high level of similarity to the target (see Clark

[6] This treatment of target-absent filler identifications is not an inherent feature of the calibration approach. Instead, it reflects the fact that often these studies had no rigorous, a priori justification for designating any particular filler as the innocent suspect.

& Tunnicliff, 2001, for a review). Clark and Tunnicliff argue quite rightly that although this approach promotes experimental control, it does not faithfully represent the process for creating lineups for innocent suspects in applied settings. Specifically, selecting fillers based on their match (in appearance or description) to the target produces target-absent lineup fillers selected based on their match to someone *other than the suspect* in the lineup. More importantly, Clark and Tunnicliff demonstrated that using fillers selected to match the target produced a lower false identification rate from target-absent lineups (≈5%) than using fillers selected to match the innocent suspect (≈25%). Further, the conditional probability of an innocent suspect identification—the likelihood the suspect was identified, given the witness picked someone from the lineup—was reliably above chance level when using suspect-matched fillers, but not when using target-matched fillers. These findings suggest that typical experimental methodologies may underestimate innocent suspect identification rates in applied settings. If this is true, when analyzing the confidence-accuracy relation, focusing only on innocent suspect identifications from target-absent lineups may systematically overestimate accuracy at any given level of confidence.

It remains to be seen whether the patterns reported by Clark and Tunnicliff (2001) will occur consistently. Oriet and Fitzgerald (2018, Experiment 2) provide evidence of *lower* innocent suspect identification rates for suspect-matched (5%) compared to target-matched (27%) fillers. Yet, inspection of the data from their Experiment 3 reveals that this pattern was less evident with low similarity fillers and reversed with high similarity fillers. Nonetheless, what is known at present is that there is some evidence that typical lab procedures produce systematically lower innocent suspect identification rates than more realistic lineup construction methods. Moreover, we do not understand why this phenomenon emerges sometimes and not others. Therefore, we contend that this caveat (relating to the effects of varied approaches to estimating accuracy) remains important when generalizing levels of

accuracy obtained in the lab to applied settings in order to evaluate the likely accuracy of an individual identification made with a given level of confidence. Although the typical calibration approach deliberately over-estimates error rates by including target-absent filler identifications, the CAC approach may underestimate error rates by considering only suspect identifications from lineups constructed in a manner that may underestimate innocent suspect identification rates. The truth almost certainly lies somewhere between the two approaches, and may be better approximated using a "worst case scenario" approach (where the most frequently identified target-absent lineup member serves as the designated innocent suspect; Pryke, Lindsay, Dysart, & Dupuis, 2004). At present it is not clear that either the calibration or CAC approach allows for a reliable and generalizable estimate of the *absolute* level of accuracy associated with individual levels of confidence. Thus, decision-makers must exercise caution when evaluating an individual identification based on the witness's expressed level of confidence.

Understanding Boundary Conditions

Through experimentation, researchers have gained some understanding of the boundary conditions for the confidence-accuracy relation for eyewitness identification. As discussed above, both calibration and CAC analyses suggest a positive and generally linear confidence-accuracy relation that, based on research to date, appears robust against manipulations of forensically- and theoretically-relevant variables (e.g., Brewer & Wells, 2006; Carlson et al., 2017; Dodson & Dobolyi, 2016; Palmer et al., 2013; Sauer et al., 2010; Wixted et al., 2016).

However, our understanding of the boundary conditions for the confidence-accuracy relation is, at present, limited in three important ways. First, although the effects of experimental manipulations on identification accuracy vary in magnitude across studies, these manipulations often produce what might be characterized as modest changes to overall

accuracy. Reported effects are generally statistically significant and non-trivial in size. Nonetheless, they represent only a tentative probing of the boundary conditions for the confidence-accuracy relationship. For example, when looking at effects on identification accuracy, Sauer et al.'s (2010) manipulation of retention interval reduced accuracy for choosers from 62% in the immediate condition to 45% in the delayed testing condition, and Palmer et al.'s (2013) exposure duration manipulation reduced choosers from 57% to 39% in the long and short exposure conditions, respectively. When considering effects in terms of ability to identify a guilty suspect or reject a target-absent lineup, Brewer and Wells (2006) found accuracy rates between 64% and 72% for target-present lineups and 57% and 70% for target-absent lineups, depending on the particular combination of instructional bias, filler similarity, and target stimulus considered. Carlson et al.'s (2017) weapon-presence manipulations produced relatively large effects: A visible weapon (cf. no weapon control) reduced accuracy from 50% to 25%, and from 25% to 16%, for target-present and -absent lineups, respectively. However, even in this case, effects were only in the 10-25% range. This is by no means an exhaustive list of findings, but it is indicative. In the confidence-accuracy calibration literature, effects on overall accuracy generally range from somewhere around 5% to somewhere around 20%. What do we know about changes in the confidence-accuracy relationship when manipulations produce large reductions in accuracy? The answer is very little.

Perhaps conditions that contribute to low decision accuracy do so primarily by increasing filler picks. As filler picks are excluded from CAC analyses, the diagnostic value of very high confidence identifications would, of course, not be undermined. Alternatively, under conditions that undermine accuracy, might there be some individuals who identify an innocent suspect but are simply unlikely to adjust confidence appropriately? This might apply to the dispositionally confident witness, to one who simply does not reflect sufficiently on the

encoding and test conditions, or one who does not entertain hypotheses about why they might be wrong (cf. Brewer, Keast, & Rishworth, 2002). We do not know the answers to these questions. Perhaps the high confidence-high accuracy relationship at the aggregate level would be little affected. However, we believe these are important issues when it comes to evaluating the likely accuracy of an individual identification, and it may be issues such as these that underpinned the concerns expressed in Loftus et al.'s 2012 Pennsylvania Instructions to which we referred earlier.

Indeed, several theoretical frameworks hold that the accuracy of metacognitive judgments will weaken when memory quality is reduced. These include the *optimality hypothesis* (Bothwell, Deffenbacher, & Brigham, 1987; Deffenbacher, 1980) and *memory constraint hypothesis* (Hertzog, Dunlosky, & Sinclair, 2010; see also Perfect & Stollery, 1993). Central to these frameworks is the idea that metacognitive judgments are influenced by various cues and heuristics, some of which are more diagnostic than others of memory accuracy. Crucially, memory strength influences the degree to which diagnostic information is available during metacognitive judgments: When memory is stronger, more diagnostic information is available, and confidence judgments will better reflect accuracy. However, in the eyewitness identification literature, empirical support for this notion is mixed; some results align with predictions based on the optimality hypothesis (e.g., Bothwell et al., 1987; Brigham, 1990; Krafka & Penrod, 1985) and some do not (e.g., Palmer et al., 2013 and Sauer et al., 2010, both found superior confidence-based discrimination in some conditions associated with lower overall accuracy).

Perhaps, as suggested by recent CAC papers, the confidence-accuracy relation for suspect identifications—or at least the accuracy of high-confidence suspect identifications—is robust against even extreme boundary conditions. Indeed, Semmler et al. (2018) recently proposed a theoretical explanation for the apparently robust accuracy rates associated with

high levels of confidence. Essentially, Semmler et al. argued that a constant likelihood ratio signal detection model—in which confidence criteria "fan out" across a memory strength continuum—predicts that suspect identification accuracy at high levels of confidence should remain stable despite changes in discriminability or overall accuracy. This is a plausible theoretical account, and could go some way to assuaging the concerns we have expressed in this section. Yet, it is notable that Semmler et al.'s conclusion was only supported by one of the three judgment conditions they examined.

Second, experimental studies understandably avoid confounding manipulations. This practice is entirely justifiable, but places important limitations on the generalizability of conclusions to applied settings where such controls are absent. Thus, although some experimental manipulations (e.g., retention interval, exposure duration, administrator influence) may produce modest effects on performance and/or the confidence-accuracy relation in lab settings, this is not to say that these variables will not be associated with larger effects in applied settings. For example, reducing memory quality via manipulations of encoding duration or retention interval may not in itself nullify the confidence-accuracy relation. However, in applied settings, a very long (and not atypical) retention interval or very dim illumination conditions (and the associated reductions in memory quality) may, for example, interact with witnesses' assumptions about the likelihood of the target being present in the lineup to influence choosing and confidence in ways we cannot necessarily predict. As an example, we draw on the demonstration by Brewer and Wells (2006) of how variations in the target-absent base rate affected the confidence-accuracy relation. Witnesses' assumptions about the likelihood the target will be present affect their decision criterion placement (i.e., if a witness expects the target to be present they are likely to set a more lenient response creation compared to a witness who expects the target to be absent). As the target-absent base rate increases, a lenient decision criterion becomes increasingly problematic. According to

theorizing about confidence judgements grounded in signal detection and accumulator models, a lenient (cf. conservative) criterion will produce more false identifications made with higher levels of confidence (e.g., Green & Swets, 1966; Van Zandt, 2000; Vickers, 1979). The effect this will have on the accuracy of high-confidence suspect identifications will likely interact with suspect plausibility and filler similarity in ways we can speculate about in a general sense, but not necessarily anticipate in an individual case. As a further example, Eisen, Smith, Olaguez, and Skerritt-Perta (2017) demonstrated that participants who were led to believe they were part of an actual criminal investigation were more likely to pick from a showup (including when the suspect was innocent), and showed greater overconfidence than participants who completed the identification test under standard laboratory conditions. Furthermore, admonitions intended to address problematic assumptions about the likely guilt of a suspect attenuated this effect (though were less effective for more plausible innocent suspects). Thus, while researchers are developing an understanding of the boundary conditions for the confidence-accuracy relationship in controlled settings, we must be cautious when generalizing findings (especially about the likely accuracy of any individual identification made with a given level of confidence) to applied settings.

Finally, even though some analyses suggest that accuracy for highly confident identifications is less volatile than accuracy at lower levels of confidence, the current literature provides little guidance on how stable this phenomenon is across variations in the way lineups are constructed and suspects are selected (i.e., factors that may affect innocent suspect identification rates) even given otherwise pristine testing conditions. The re-analyses we present below, however, do speak to this issue.

Should CAC Findings Guide Evaluations of Individual Cases?

In several recent papers, researchers have shown that high levels of confidence indicate high levels of accuracy (e.g., Carlson et al., 2017; Mickes, Clark, & Gronlund, 2017; Wixted et al., 2015; Wixted & Wells, 2017), but also noted this only holds when testing conditions are pristine. Supporting their claims about the robust accuracy of high-confidence identifications, they provide CAC curves based on new data, and on re-analyses of previously published calibration data. The curves presented do indeed show consistently high accuracy at the highest levels of confidence. However, here we present re-analyses of several published datasets that challenge the generality of the "high confidence implies high accuracy" conclusion to situations where decision-makers must evaluate an individual identification. Before presenting these re-analyses, we emphasize the following. We certainly do not dispute the existence of a meaningful confidence-accuracy relation; in fact, we argue strongly in support of that claim. Nor do we quarrel with the conclusion that, at the aggregate level, highly confident suspect identifications are highly likely to be accurate.

The datasets we re-analyze are large enough to provide stable estimates of the confidence-accuracy relation, though the specific conditions we focus on have not previously been subjected to this analysis. We selected these datasets because they demonstrate conditions under which the "high-confidence, high-accuracy" proposition breaks down. We stress that the datasets are not representative of the literature in aggregate. In fact, they come from studies or conditions that violate the pristine conditions referred to by Wixted and Wells (2017) and would be disregarded by those authors. But we will argue that these violations (a) have occurred despite the researchers following best practice lineup construction procedures, (b) cannot necessarily be anticipated, and (c) can severely affect the accuracy of highly confident suspect identifications. In other words, these datasets have important implications for decision-makers who need to evaluate individual identifications, and might draw

inappropriate conclusions based on the literature detailing the confidence-relation in aggregate (Wixted & Wells, 2017).

Wixted & Wells (2017) note that the "high-confidence, high-accuracy" proposition will break down when lineups are biased so that the suspect stands out in some way. What does it mean for a lineup to be biased? One obvious example of lineup bias occurs when fillers are not selected based on their match to a description of the target (or their physical resemblance to the target) and, consequently, the suspect is clearly the only lineup member matching that description (or resembles the target). This form of bias is likely to be detectable based on a visual inspection of the lineup by an experienced researcher. A less obvious case of bias may be detectable only after a significant amount of data has been collected in a lab setting, but is unlikely to be detected in applied settings. This bias occurs when, despite the researchers' conscientious and systematic efforts to match fillers to the witness's description (or target's physical appearance) and achieve suitable functional lineup size, it becomes clear after a significant amount of lab data have been collected that one person in the target-absent lineup was selected much more often than others. Recent work by Tardif et al. (2019) highlights an alternative avenue—other than coincidental or unusual resemblance—through which an innocent suspect might stand out as distinctive in a lineup, despite efforts to follow best practice guidelines. Tardif et al. (2019) demonstrated that most of the variance in face recognition performance (between super-recognizers, "normal" participants, and prosopagnosics) can be predicted by participants' use of information relating to the eyes/eyebrows and the mouth of the target stimuli. What is the likelihood of such features being captured, in detail, in a witness's description and then being sufficiently replicated in the selected fillers to avoid a suspect who possesses these features standing out? Lindsay, Martin, and Webber (1994) found that details relating to a culprit's eyes, eyebrows, and mouth were included in ~3%, 0%, and 0%, respectively, of the 105 descriptions they sampled

from real crimes. Particularly distinctive examples of these features might be mentioned, but would descriptions also include relevant information relating to spatial relations between features? If not, one can imagine an innocent suspect might possess identifiable features that would not be captured in fillers, and yet the innocent suspect would be most unlikely to be recognized as standing out based on those features.

Thinking in more concrete terms, how this might play out in a field setting? The APLS white paper recommends selecting fillers who match a description of the target. Let's assume that, having done this before constructing the lineup, the officer/s constructing the lineup put these filler photos alongside the photo of the suspect and compare them carefully for physical resemblance (selecting the best of the bunch to serve as fillers in the lineup). If the officers did this meticulously they might be able to match for eye color (if the photo were clear enough, which is often not the case), and potentially for very distinctive (e.g., bushy) eyebrows, and possibly for a distinctive (e.g., very wide) mouth. Leaving aside the most striking examples of these features, would the officers be likely to match the angles of those features, their width, the distance between them, their positioning on the face (i.e., factors Tardif et al. suggest may be very important)? Maybe, maybe not. However, those features are ones that somehow, in some cases, only the witness has picked up on (though they probably wouldn't verbalize them) and, in some cases, maybe only a very small proportion of (a very large number of potential) cases, the innocent suspect might also possess. Of course, the same logic applies if fillers are selected according to resemblance to image of the culprit obtained from CCTV footage. Would CCTV images allow the officers to discern the key features, appreciate the information that may have been distinctive to a particular witness who saw the culprit live, and then replicate the necessary diagnostic features across fillers? Maybe, maybe not.

Given such cases, how might police avoid constructing a biased lineup? Properly matching fillers to a description or an obtained image of the culprit should avoid the first source of bias, but not necessarily the second. How could an officer constructing a lineup reasonably ensure the suspect, if innocent, is no more plausible than the fillers? Without replicating the original encoding event and collecting a significant amount of lineup data, this seems impossible. That means if a lineup is constructed using a match description (and/or match resemblance) strategy and has high functional size, any classification of the lineup as biased due to unusual suspect plausibility must be post hoc. The necessary conclusion is that it cannot be anticipated and a decision-maker evaluating an individual identification obtained under such conditions cannot know in advance that the general "high-confidence, high-accuracy" proposition will not hold in this specific case. Wixted and Wells (2017) were apparently sensitive to this dilemma when they noted: "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).

The data we present shortly are simply cases that demonstrate this point. We are not claiming these data are representative of the aggregate confidence-accuracy relation; rather we are saying these situations can arise despite careful lineup construction. We are not saying that the innocent suspects in these datasets did not have more chance of being selected than other lineup members; rather we are saying that sometimes this can only be known post hoc. We are not claiming that such cases are likely to be common in field settings; rather, as noted by Wixted & Wells (2017), we suggest they could happen. Thus, we argue, it is very risky to make strong recommendations, albeit with provisos, that the police and the courts may pick up on as applying directly to an individual case where in fact one of the critical provisos (namely, the suspect did not stand out) may not be verifiable.

When considering what steps officers might reasonably be expected to take to avoid lineup bias, we refer to the following quote from the published working draft of the updated version of the *APLS Scientific review of eyewitness identification procedures*:

> "We are not suggesting that police have to conduct a mock witness test on each lineup in order to know if they have a good lineup. Instead, we believe that a conscientious and objective detective would have a good sense of whether the lineup was fair without conducting a mock witness test with a large number of people. However, we recommend that a non-blind police officer building the lineup have at least one or two other people (ideally, blind as to which person is the suspect) look at the witness description and the lineup to get a second opinion on whether it would pass a mock witness test." (Wells et al., 2018, p.45)

When deciding whether or not to include each of the datasets examined below, our key question was not: Did the final data set indicate that the innocent suspect stood out from the other lineup members? Rather, it was: Did the details provided on how the lineups were constructed in the manuscript methods' sections indicate that the researchers reached this minimum standard? If they did, we argue that the data speak to conditions under which, despite the ostensible fairness of the lineup, the confidence-accuracy relation might breakdown and conclusions based on the aggregate confidence-accuracy relation might lead to erroneous evaluations of an individual identification.

First, we re-analyzed data from Gronlund, Carlson, Dailey, and Goodsell (2009) to produce CAC curves. This study originally compared identification performance from simultaneous and sequential lineups and, although some information was presented relating to the confidence-accuracy relation, no conclusions relevant to the present article were drawn.

Participants viewed a simulated crime video and, after a 10 minute distractor task, made an identification from either a sequential or simultaneous 6-person lineup (i.e., one suspect and five fillers, with the target-absent lineup including a designated innocent suspect). Participants then provided a confidence rating on 1-7 scale (1 = *not all confident*; 7 = *very confident*). Note that these data have previously been excluded from some meta-analyses (e.g., Palmer & Brewer, 2012, and the "gold standard" subset reported by Steblay, Dysart, & Wells, 2011) because they produced idiosyncratic innocent suspect identification rates, although they have been included in other meta-analyses (Fitzgerald, Price, Oriet, & Charman, 2013, and the overall analyses reported by Steblay et al., 2011). Although Gronlund et al.'s data show idiosyncratic patterns of results, and were designed to create a situation in which the innocent suspect was highly plausible, Gronlund et al.'s (2009) procedure for lineup construction, under careful scrutiny, appears meticulous and thoughtful (see p.143 of the original article). Briefly, when selecting "good" fillers (i.e., for their fair lineup conditions[7]), two research assistants who had not seen the target event each identified a pool of 50 potential fillers who all matched the sex, ethnicity, and five key descriptors of the target (distilled from descriptions provided by 27 pilot participants), and had no distinctive characteristics (tattoos, beards, or bald or shaved heads). The first author (Gronlund) then examined this pool of fillers, and excluded any he judged to insufficiently resemble the target (thus, good fillers needed to match core components of the description *and* look sufficiently similar to the target). This produced a pool of 50 "good" fillers from which lineups were constructed. These lineups were shown to 76 mock-witnesses, who had

[7] We do not use Gronlund et al.'s original condition labels referring to "fair" and "biased" lineups. Suspect identification rates in Gronlund et al.'s "fair" lineup conditions indicate a bias toward the highly plausible innocent suspect despite the quality of fillers of selected. Thus, we refer to "good" vs. "poor" filler conditions. "Good" fillers needed to match ethnicity, sex, and five other descriptors (**and be judged as sufficiently similar in physical appearance to the target**) whereas poor fillers only matched ethnicity, sex, and one other descriptor (and were removed if judged to be too similar to the target. The good and poor fillers as referred to in the current manuscript relate to the "fair" and "biased" lineups reported in Gronlund et al.'s original paper.

not seen the video but had learned the description of the target, and identified the lineup member who best matched that description. Based on data from these pilot participants, Gronlund et al. excluded fillers selected at a rate lower than chance. Some lineups were altered as a result of this initial piloting, and the process was repeated with second group of 55 mock-witnesses to produce the lineups used in the study. Clearly, regardless of the eventual outcome, the care taken by these researchers is likely to exceed the capacity of officers constructing lineups in field settings. It is extremely important that researchers do not neglect datasets or conditions that are characterized by high functional size but do not conform to the broader confidence-accuracy pattern. This is particularly true given that the datasets we currently have are likely derived from a very limited sampling of the encoding and test conditions likely to prevail in real crimes and lineups. However, for those who remain unconvinced that these data are informative about the confidence-accuracy relation in field settings, their inclusion can instead serve to highlight the need for researchers to provide more detailed reporting and closer scrutiny of response patterns to check assumptions relating to lineup fairness, and as indicating that summary lineup fairness indices may conceal important biases that nonetheless manifest in effects on accuracy and the confidence-accuracy relation.

We re-analyzed only the data from the simultaneous lineup conditions (N = 1,279). The key manipulations were (1) the degree of match between the target as seen in the video and the image of the target shown in the lineup (producing a strong vs. weak match condition; where the image for the strong match condition was taken on the same day as the encoding stimulus was filmed, and the image for the weak match condition was taken several weeks later, after the target had grown facial hair and changed his hairstyle), (2) the plausibility of the innocent suspect (strong vs. weak; as determined by identification rates

during pilot testing), and (3) the quality of fillers in the lineup (good vs. poor)[8]. We consider only data from the lineups including good fillers. Notably, in the good fillers condition, Tredoux's (1998) *E'* indicated high functional size (e.g., 3.75 – 4.51). However, even in the good filler lineups, the identification rate for strong innocent suspect (75%) was higher than for the weak innocent suspect (27%).

Consistent with the CAC approach, our CAC curves include only identifications of the target and innocent suspect. However, we took two approaches to collapsing raw confidence rating into bins for the CAC analyses. First, to provide a clean break between the highest level of confidence and all other levels of confidence, and to provide the best chance to observe the high levels of accuracy commonly reported at the highest level of confidence, we adopted Wixted et al.'s (2016) approach of treating the highest level of confidence as "high confidence" and everything else as low confidence.[9] Second, as per Mickes (2015) and Carlson et al. (2016), we collapsed confidence into three bins: low (0-60%), moderate (70-80%), and high confidence (90-100%). To do this, we converted ratings from the 7-point scale into percentages expressing the given rating as a function of the maximum confidence level. Thus, raw confidence ratings of 1, 2, 3, and 4 were classified as low confidence, a rating of 5 was classified as moderate confidence, and ratings of 6 and 7 were classified as high confidence. Obviously, there is some noise in this conversion. Figure 1 shows the CAC curves produced by this re-analysis, with the 3-level and 2-level CAC curves shown in the upper and lower panels, respectively.

Three findings are clear. First, the level of accuracy associated with the highest level of confidence varies substantially across conditions. For example, consider the accuracy rates

[8] Gronlund et al. also included a manipulation of encoding quality, but collapsed data across the levels of this variable because the manipulation had non-significant effects on performance.

[9] A 2-point function obviously provides only very limited information about the full confidence-accuracy relation, but our purpose here is not to speak to the full relation, but test the robustness of the claim that high confidence implies high accuracy.

displayed in Figure 1's lower panel. When the degree of match between the witness's memory of the culprit and the target as seen in the lineup is high, and the plausibility of the innocent suspect is low, the point estimate for accuracy at the highest level of confidence is ≈80% (with SE bars including values over 90%). However, when the degree of match between the witness's memory of the culprit and the target as seen in the lineup is low, and the plausibility of the innocent suspect is high, the point estimate for accuracy at the highest confidence level is extremely low: ≈20% (with SE bars including values approaching 10%). Second, at the highest confidence level, the vast majority of these curves show accuracy substantially below the commonly reported 90-100% level. Third, accuracy at the highest confidence level appears to vary systematically according to the plausibility of the designated innocent suspect. When the innocent suspect is highly plausible (functions with the circle markers), accuracy is lower – *even at very high levels of confidence* – than when the plausibility of the suspect is low. Importantly, in this case, we cannot tell what made one suspect highly plausible and the other less plausible. However, even in the lineups with good fillers and high functional size, at some point the plausibility of the suspect changes, and the confidence-accuracy relationship breaks down. Other than matching fillers to the target's description and trying to ensure high functional size, and carefully examining the photo of each filler against that of the suspect to ensure reasonable resemblance, how might an officer who cannot know in advance how plausible his or her suspect is relative to other lineup members avoid creating a lineup that may be biased against the suspect? Likewise, how can judges and jurors know if a description-matched and apparently high functional size lineup, with fillers who resembled the suspect, was biased against the suspect? In other words, how can police, judges or jurors avoid incorrectly evaluating the likely accuracy of a highly confidence suspect identification?

Next, we re-analyzed data reported by Wetmore et al. (2015). Wetmore et al. examined the effects of retention interval (immediate vs. 48hr delay), suspect type (guilty, high plausibility innocent suspect, and low plausibility innocent suspect), and lineup type (fair lineup, biased lineup, show-up) on identification performance, using one of Gronlund et al.'s (2009) encoding stimuli, their "strong match" target, and their high and low plausibility innocent suspects (described earlier). Thus, these stimuli are again the products of conscientious lineup construction efforts, even for highly plausible suspects. Although there is some overlap in the stimuli used by Wetmore et al. and Gronlund et al., we include both datasets because (a) the data came from separate samples of participants, and (b) Wetmore et al. included additional manipulations of theoretical relevance (i.e., a manipulation of memory strength). They made no conclusions relating to the confidence-accuracy relationship. Participants (N = 1,584) viewed the crime video and, after the retention interval (solving 20 anagrams or leaving and returning 48hrs later), made an identification from a 6-member lineup and provided a confidence rating on the same 1-7 scale used by Gronlund et al.

As with our re-analysis of Gronlund et al.'s (2009) data, we adopted two approaches for our CAC analyses. The results are plotted in Figure 2, again with the 3-level and 2-level CAC curves shown in the upper and lower panels, respectively. Two features of these CAC curves bear mention. When lineups include good fillers and the innocent suspect is low in plausibility, the accuracy rate for the highest level/s of confidence is consistent with the levels typically reported in the CAC literature. However, as with the Gronlund et al. data, it is clear that the plausibility of the innocent suspect affects accuracy, even at the highest level of confidence. For example, consider the accuracy rates for good filler lineups displayed in Figure 2's lower panel. At the highest level of confidence, point estimates of accuracy are as low as ≈60% (with SE bars including values below 50%) or as high as 100%, depending on

the plausibility of the innocent suspect, and the delay between encoding and testing. However, in this dataset, this effect is only apparent for lineups with good fillers.

We then reanalyzed data from Colloff, Wade, Wixted, and Maylor (2017). This experiment examined age-related differences in identification performance from fair and biased lineups, when the suspect possessed a distinctive feature. We re-analyzed only the data from the fair lineup conditions. For fair lineups, fillers were selected by (a) creating modal descriptions of the culprits based on descriptions provided by 18 participants and (b) identifying, for each culprit, 40 potential fillers who matched the modal description. Photos of fillers were edited to remove differences in background, standardize any visible clothing, and remove distinctive facial features (see description in Colloff, Wade, & Strange, 2016). Lineups were then randomly generated, for each participant, by drawing fillers from the pool of 40 potential fillers for the given culprit. Again, it is clear that the researchers were conscientious in the processes followed to select fillers and standardize lineup materials. Fair lineups involved either replicating the distinctive feature across all lineup members, pixelating the distinctive feature on the suspect (and the corresponding area of the face on fillers), or concealing the distinctive feature on the suspect (and the corresponding area of the face on fillers)[10]. The initial sample included, 1,570 young participants (aged 18-30), 1,570 middle-aged participants (31-59), and 1,570 older participants (aged 60 or over). Participants viewed one of four simulated crime events and, after completing an 8 min filler task, completed an identification task from a 6-member lineup, and provided a confidence rating in the accuracy of their decision. Data from two of these simulated crime events were reported in the original paper, while data from the other crime events were excluded from analyses but presented in the supplemental materials. The authors presented confidence-accuracy curves in

[10] Although the inclusions of conditions where (a) suspect have distinctive features and (b) researchers digitally manipulate lineup images might appear to lack ecological validity, this practice is not uncommon in the lab (Zarkadi, Wade, & Stewart, 2009) and occurs frequently in police lineup construction due to the prevalence of scars, tattoos, and other distinctive features. Moreover, it is a recommended practice (Wells et al., 2018, p. 44).

the original article, but drew no strong conclusions about absolute levels of accuracy at any confidence level. We analyzed data from all four stimulus sets, but present results separately for the data included and excluded from original paper. Colloff et al. excluded these data because performance was "very low for young participants, and at floor for older subjects." (p. 246). However, we reiterate our earlier point about understanding the boundary conditions of the confidence-accuracy relationship, particularly given recent suggestions that accuracy at very high levels of confidence is robust against changes in task difficulty (Semmler et al., 2018). Clearly these excluded data are highly informative because they are obtained from fair lineups and under conditions indicating a difficult discrimination task, probably not very different from many 'real world' situations.

Before considering the results, three points are worth noting. First, like Colloff et al., we collapsed data across the three versions of fair lineup. Second, this study did not include a designated innocent suspect for its fair lineup conditions. Thus, for fair lineups, the innocent suspect identification rate is estimated by dividing the total number of identification from target-absent lineups by the number of the lineup members (i.e., 6). Third, despite the impressive size of the initial dataset (N = 4,710), some individual points on the CAC curves are based on a low number of observations and the patterns of results must interpreted with caution (Table 1 presents, for each CAC, the number of datapoints at the highest confidence level and the estimated innocent suspect identification rate for each condition).

We calculated our CAC curves based on the data available in Colloff et al.'s manuscript and supplemental materials. Confidence data were binned (by the original authors) into 5 categories: 0-20%, 30-40%, 50-60%, 70-80%, and 90-100%. We used these counts to collapse confidence data into 3-category CAC curves, as per the previously described analyses (see Figure 3; data included in and excluded from the original paper are presented). Despite the potential noisiness of the curves, one finding is clear: Although one

curve (young participants viewing fair lineups in the originally included data) shows accuracy in the 90-100% range for the highest level confidence, and a couple of other curves show accuracy rates of approximately 90% at the highest confidence level, most of the curves do not. For the data excluded from the main part of the paper (i.e., the more difficult conditions), accuracy rates at the 90-100% confidence levels do not match the level of confidence level expressed. Thus, counter to Semmler et al.'s (2018) argument, Colloff et al.'s data show that increases in task difficulty were associated with reduced accuracy at the highest level of confidence; even in conditions where the lineups were constructed to be fair, and the innocent suspect identification rate was set to chance (i.e., estimated by dividing the total number of target-absent identifications by the number of the lineup members).

Finally, we reanalyzed data reported by Sučić, Tokić, and Ivešić (2015). We note that Sučić et al. (2015) purposefully selected the most similar filler (determined by pilot ratings) as their designated innocent suspect for target-absent trials. We appreciate that this approach could, in line with Wixted & Wells' (2017) caveat relating to "unusual" levels of resemblance, inflate innocent suspect identification rates, although using subjective similarity ratings in this way in no way guarantees this will occur (see, for example, Brewer, Weber, & Guerin, 2019). However, the researchers were clearly conscientious in their efforts to construct fair lineups for their target/suspect. First, 13 participants provided a description of the target based on a brief (5-7 s) exposure. These descriptions were used to produce a modal description, including features about which at least 50% of participants agreed. Based on this modal description, a pool of potential fillers was identified. One group of 17 participants rated the similarity of each pair of potential fillers. A second group of 27 participants rated the similarity of each potential filler to the description of the target. The fillers selected were those that were top-ranked for inter-photograph similarity and match to description. The filler with the highest similarity rating was selected as the designated innocent suspect. The target-

absent lineup was pilot-tested with 39 new mock-witnesses, and produced a Tredoux's *E* of 5.14. We note that, although the innocent suspect identification rate in the study proper was 35% (i.e., above chance), when selecting their designated innocent suspect the researchers followed a procedure that is probably common to many studies and that, while increasing the likelihood of an innocent suspect identification, did not produce an innocent suspect rate much higher than the next-most selected filler (24%).

The original study investigated the confidence-accuracy relation for sequential and simultaneous lineups in a field setting. A confederate approached a potential participant, and interacted with them for 15-60s, showing the participant both front-on and side views of their face during the interaction. Thirty seconds after the interaction, the experimenters approached the potential participant and those who consented completed an identification task and provided a confidence rating in the accuracy of their decision. Based on their analyses, which collapsed across lineup type, Sučić et al. reported a confidence-accuracy relation that was meaningful but imperfect. We re-analyzed their data looking only at decisions from simultaneous lineups and, as per our previous analyses, include only identifications of the designated innocent suspect (see Figure 4). Again, despite the researchers' conscientious efforts to ensure lineup fairness—using match-description and match-to-culprit strategies, and multiple rounds of pilot testing for similarity to produce a lineup with high functional size and high filler similarity—the accuracy of highly confident suspect identifications is well below that typically reported in the CAC literature.

What should we make of the patterns obtained from these four re-analyzed datasets? First, high confidence does not consistently imply high accuracy. Second, there are factors that affect accuracy rates at even the highest levels of confidence, but these effects are not always consistent in direction. Third, although we can identify manipulations that produce these effects (e.g., the plausibility of the innocent suspect), we do not necessarily understand

the mechanism/s through which these effects emerge. It is clear that the innocent suspect is the most plausible member in some of the cases we have highlighted and, following Wixted and Wells' (2017) approach, these data would be discarded. But it is also clear that the researchers in the four studies have followed procedures in lineup composition that are systematic, appropriate and, importantly, much more sophisticated than police are likely to follow or would be expected to follow. Moreover, the bias, or unfair nature of the lineup, in these cases has only come to light after lineup data had been obtained from large samples and, indeed, long after peer review and publication.

Are cases with highly plausible innocent suspects likely to occur in practice?

We acknowledge Wixted and Wells' (2017) caveat about the impact of unusual resemblance on the diagnostic value of very high confidence identifications, and their claims that cases of coincidental and unusual resemblance are likely to be rare (or, in the case of unusual resemblance, predictable and that appropriate filler selection strategies will preserve pristine testing conditions). Some may argue that cases of coincidental resemblance are inherently rare enough that they have little or no bearing on the applicability of "high-confidence, high-accuracy" conclusion to individual cases. We are not sure how rare such cases are likely to be, or how rare they would need to be in order to be dismissed out of hand. However, we believe the implications of such cases are non-trivial when considering the generalizability of the high-confidence, high-accuracy conclusion.

Are coincidences inherently rare? Statisticians are aware that extremely improbable events are commonplace (Hand, 2014). To get a sense of how rare such occurrences are likely to be, and whether rare occurrences are likely to be important, a first step might be to consider how often they have the opportunity to occur. Wixted & Wells note that none of the Innocence Project's DNA exoneration cases involved coincidental resemblance. There are currently 259 Innocence Project exonerations involving mistaken identification. This sounds

like a big number. Is it though, relative to the number of identification parades being conducted? It is difficult to find clear and comprehensive estimates of the frequency of identification procedures in field settings. Here we present data, albeit imperfect, that speak to this issue. In the Police Executive Research Forum (PERF, 2013) report submitted to the National Institute of Justice, researchers contacted a random stratified sample of 1,377 law enforcement agencies throughout the US, and 619 responded. Of the 316 agencies that reported their use of lineups, the average number of lineups for 2010 was 41. Thus, based only on the responses from this sample, there were over 12,000 lineups conducted in the US in 2010. Regarding lineups in the UK, Horry, Halford, Brewer, Milne, and Bull (2014) reported 833 lineups conducted over an 8 year period (1992-2000) in Hampshire alone (i.e., one of 45 territorial police forces in the UK, and the 14th largest in terms of no. of officers employed and area covered), and Valentine, Hughes, and Munro (2009) estimated that 80,000 lineups were conducted in 2006 alone, across England and Wales. These data are clearly incomplete, but nonetheless indicate that there are likely to be thousands of identification tests run each year, under varying conditions in the US alone, and many thousands more internationally. This seems to provide a reasonable opportunity for rare events to occur.

Will best practice lineup construction methods prevent this problem?

From the perspective of evaluating a particular identification we see no reason why this situation could not arise when police construct lineups. Moreover, we see no guaranteed method for preventing it, regardless of how conscientious officers might be in their efforts to construct unbiased lineups. As argued above, Gronlund et al. used both match-description and match-resemblance protocols when selecting their "good" fillers. Thus, they used an approach generally regarded as best practice (match-description) and augmented this with a match-resemblance approach (as recommended for cases where the suspect is likely to strongly resemble the culprit for non-coincidental reasons; e.g., because they became a

suspect based on their resemblance to CCTV footage of the target; see Wixted & Wells, 2017). This conscientious approach did not preclude an adverse effect on the accuracy of high-confidence responses. Critically, in cases where it happens, the investigating officer, the judge, and the jurors will have no basis for knowing that the "high-confidence, high-accuracy" proposition does not apply to the suspect identification under consideration. As Wixted & Wells note, some methods of arriving at a suspect (e.g., if the suspect becomes a suspect *because* they resemble a CCTV image of the perpetrator) might be more likely than other methods (e.g., if the suspect becomes a suspect because they have committed a similar crime on a previous occasion) to produce suspects that, when innocent, are nonetheless highly similar to the culprit. However, there may also be situations where a given suspect appears highly plausible to a given witness based on factors that cannot necessarily be recognized or quantified (cf. Tardif et al., 2019).

As already noted, Wixted & Wells' (2017) clearly warned that the "high-confidence, high-accuracy" proposition will break down when lineups are biased; where the suspect stands out because the fillers in the lineup are not sufficiently plausible. However, these authors also acknowledge that the criteria for establishing fairness are not well-defined. Although lineup bias may be obvious in some cases, this will not always be true and the absence of an obvious bias does not entail fairness. Moreover, although the literature reports a variety of metrics designed to measure lineup fairness, these indices may not be robust enough to guide decision-making in applied settings. This point is borne out in a recent paper by Mansour, Beaudry, Kalmet, Bertrand, and Lindsay (2017). Using a mock-witness paradigm and different types of target description (i.e., modal descriptions, descriptions provided by single witnesses, etc.), Mansour et al. assessed the reliability and validity of various approaches to assessing lineup bias (e.g., measures of functional size and of bias against the suspect or defendant), with a sample of over 1,000 participants. The authors

concluded that "lineup fairness measures cannot be accepted at face value as reflecting the properties of the lineups they are used to measure" (p. 112), and "do not meet the *Daubert* criteria that would justify presenting them as evidence, at least for lineups constructed to be fair." (p. 113) How, then, might prosecutors, defense attorneys, judges, and jurors determine whether the lineup from which a suspect was identified meets the fairness threshold required for high confidence to indicate a high likelihood of accuracy?

In sum, two critical questions seem to us to be unresolved. First, if researchers following meticulous procedures in controlled lab conditions cannot guarantee a lineup will not be biased against the suspect, what can reasonably be expected of officers in field settings (who don't have the advantage of experimental data to assess the plausibility of their suspect)? Second, without reliable and valid measures of lineup fairness, how are triers of fact to know whether, for a given identification made with high confidence, the "fairness" requirement for the "high-confidence, high-accuracy" proposition has been met?

Some Final Considerations on the Utility of Confidence

As a general principle, researchers favoring both the calibration and CAC approaches agree that confidence is only likely to be diagnostic of accuracy if measured immediately following the decision (i.e., prior to the witness being exposed to any social influence; e.g., Brewer & Palmer, 2010; Wells & Bradfield, 1998; Wells et al., 1998). However, there are some other points to consider when discussing the applied utility of witnesses' expressed confidence for evaluating the likely accuracy of an identification.

First, if confidence is to be useful, how should it be recorded? In applied settings, current protocols (where they exist) typically suggest recording confidence "in the witness's own words". Verbal expressions of uncertainty do, in some contexts, better capture underlying psychological uncertainty and provide better indications of future behavior (Windschitl & Wells, 1996). However, based on a very limited literature for recognition

decisions, there is little evidence of any difference in the diagnostic value of confidence measured on numerical and verbal scales (Weber, Brewer, & Margitich, 2008). However, the verbal expressions of confidence in these comparisons are based on standardized, ordinal confidence scales with verbal labels. These constrained verbal expressions of confidence, and the systematic confidence-accuracy relations they produce, may not generalize to applied settings where a witness expresses confidence "in their own words" (i.e., in a format that lacks an inherent ordinal structure). Unless the witness's own words express an extreme level of confidence (e.g., "I'm certain it's number six" or "It might be number six, but I'm really not sure"), such spontaneous utterances may be ambiguous, and of limited value when assessing the reliability of an individual identification. Thus, our preference is for the use of standardized, numerical confidence scales, a suggestion also made by Sauer and Brewer (2015). Specifically, although Tekin and Roediger (2017) demonstrate that different confidence scales produce similar confidence-accuracy relations, we advocate the use of a 0-100% scale. There are sufficient data to demonstrate that adult witnesses can use such scales to effectively discriminate between instances where their decisions are likely (vs. unlikely) to be correct and, although in the case of confidence these scales may not possess ratio scale properties, they do provide a less ambiguous metric for interpreting the expressed level of confidence. These scales confer additional benefits for researchers, in that they are uniquely amenable to a variety of analyses useful for assessing the confidence-accuracy relationship (e.g., assessing over/underconfidence). In both applied and lab settings, 0-100% scales also have the advantage that they are less vulnerable than verbal scales to post-hoc massaging or misperception by those using the information. For example, a verbal expression of confidence such as "likely" could plausibly be interpreted as meaning highly confident or moderately confident. In contrast, although an expression of "70%" confidence might not entail a 70%

chance of accuracy, its meaning relative to upper and lower extremes of the scale is less ambiguous.

Second, we reiterate the suggestions made by previous researchers that only confidence recorded immediately after the identification is made should be presented in court (e.g., Brewer, 2006; Brewer & Wells, 2009; Sauer & Brewer, 2015; Wixted et al., 2015; Wixted & Wells, 2017). Given that various social influences (e.g., post-identification feedback and pre-trial preparation) can dissociate confidence from accuracy, and that this dissociation tends overwhelmingly to manifest in confidence inflation, in-court expressions of confidence are likely to systematically undermine the evaluation of identification evidence (e.g., Semmler & Brewer, 2006; Wells & Bradfield, 1998; Wells et al., 1998).

The final point we would like to make relates to the oft-repeated suggestion that high confidence identifications are the ones most likely to end up at trial. For example, Carlson et al. (2017) noted that "identifications made by highly confident eyewitnesses [are] most likely to make it to trial. (p.8)." Similarly, Roediger, Wixted, and DeSoto (2012) argued that "Most cases of eyewitness identification come from people who are highly confident and believe they are correctly identifying the right person. Identifications made with low confidence generally never make it to a court of law and are given little weight if they do. (p. 88)" When this claim is made, we suspect most readers would interpret it along the following lines: High confidence identifications are both highly likely to be accurate and more likely (cf. low confidence identifications) to end up in court; thus, low confidence identifications are unlikely to contribute to wrongful convictions, and identification evidence as presented in court is highly likely to be reliable.

We suggest this claim is problematic for several reasons. First, to the best of our knowledge, at present there are no systematic protocols in place for collecting confidence following an identification in many jurisdictions. Thus, there is no guarantee confidence is

even collected, let alone collected in a way that can be sensibly interpreted (cf. *in the witness's own words*). And, as noted earlier, although we know that police and lawyers find confident testimony persuasive, we are not aware of *data* showing that the precise level of confidence a witness expresses influences the decision to proceed with a prosecution. Second, when confidence is collected, we are not aware of hard evidence that clearly shows that police or prosecutors are likely to "dismiss" low confidence suspect identifications. We do know, however, that suspect identifications made by less-than-certain witnesses have contributed to wrongful convictions (Garrett, 2011). We also know that witnesses who were not confident when making their identification can appear very confident by the time they reach court (Garrett, 2011; Wixted & Wells, 2017). Finally, the Innocence Project's DNA exoneration cases clearly indicate that mistaken identifications contribute to wrongful convictions[11]. Thus, in these cases, either (a) the witness was highly confident but wrong at the identification, (b) the witness was low in confidence but the prosecution proceeded anyway, or (c) the witness's confidence did not affect the decision to prosecute the suspect, or was not recorded. Regardless, the suggestion that it is very confident and highly likely to be accurate identifications which are most likely to end up at trial is one that we believe does not yet have robust empirical support.

Conclusion

In sum, this review leads to the following conclusion: The extent of variation in the confidence-accuracy relation precludes us from making strong, generalized claims about the accuracy of high confidence identification decisions, even under "pristine" conditions, *when evaluating individual identifications* (cf. considering the confidence-accuracy relation in aggregate). It is true that a growing body of theoretical and empirical literature converges on the conclusion that, as per the basic memory and decision-making literature, confidence can

[11] www.innocenceproject.org/

be informative about likely accuracy in the eyewitness identification domain. However, as discussed above, the limitations of common approaches to measuring identification accuracy and our understanding of the boundary conditions for the confidence-accuracy relation, combined with the data presented here, demonstrate that we cannot confidently draw decisive conclusions about the accuracy of an individual identification made with a particular level of confidence. For example, as we have shown, in experimental conditions where lineups were ostensibly fair—or, at least, where fillers were selected to provide a good match to both the description and physical resemblance of the target—but innocent suspects were highly plausible, anywhere between 4 (Wetmore et al., 2015) and 8 (Gronlund et al., 2009) out of every 10 extremely confident identifications might be wrong. How likely are such cases to emerge in the field? We cannot say. Some may argue that such cases are likely to occur so rarely that we have simply created a red herring. We leave that for readers to judge. We take the view that the consequences that flow from a positive identification in an individual case are generally so significant that we are reticent to endorse a position that states or implies that triers of fact can assume a suspect identification is accurate if it is made with high confidence.

In the studies we have re-examined, the researchers may have deliberately chosen a standout innocent suspect (as Gronlund did) and then deliberately filled the lineup with poor fillers. But we take the researchers' descriptions of what they did at their word, and their descriptions indicate thoughtful attempts to select suitable fillers. However, as with our own experience constructing and testing lineups, even when we carefully assess rated similarity, it often turns out that someone other than the most similar-rated non-target lineup member is the most popular pick from the target-absent lineup. In other words, being systematic and attempting to follow best practice guidelines can break down because we as researchers, and more importantly police officers in the field, cannot see what the witness has seen at the

crime. Thus, the question becomes: Although we accept the importance of using confidence to investigate evaluate identification evidence, do we want to tell the courts that a very high confidence identification of the accused is just about guaranteed to be accurate? To some degree, this is a matter of individual judgment about what constitutes sufficient evidence. Researchers will differ on this issue. Clearly, we fall on one side, but all we hope is that researchers and practitioners think about these possibilities.

This certainly does not mean we are abandoning confidence as a useful tool for evaluating identification evidence. Rather, we are issuing a warning that, regardless of what the data indicate at the aggregate level, high confidence may be very misleading in an individual case. Moreover, we likely will not know when this caveat applies. Nor will the most conscientious police procedures be able to prevent the problem occurring.

We maintain that confidence can certainly be informative (if measured appropriately), and that higher levels of confidence indicate an increased likelihood of accuracy. However, to close, we reiterate a point made by Brewer & Wells (2006) which we believe captures the limits of what the literature to date permits in terms of conclusions:

> "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)

References

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

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics, 54*(1), 75-81. doi:10.3758/bf03206939

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.

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

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

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology, 15*, 77-96. doi:10.1348/135532509x414765

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

Brewer, N., Weber, N., & Guerin, N. (2019). Police lineups of the future? *American Psychologist*. doi:doi.org/10.1037/amp0000465

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.

Brewer, N., & Wells, G. L. (2009). Obtaining and interpreting eyewitness identification test evidence: The influence of police-witness interactions. In T. Williamson, R. Bull, & T. Valentine (Eds.), *Handbook of psychology of invetigative interviewing: Current developments and future directions* (pp. 205-220). Chichester: Wiley-Blackwell.

Brigham, J. C. (1990). Target person distinctiveness and attractiveness as moderator variables in the confidence- accuracy relationship in eyewitness identifications. *Basic and Applied Social Psychology, 11*, 101-115.

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*(1), 82-92. doi:https://doi.org/10.1016/j.jarmac.2016.04.001

Carlson, C. A., Young, D. F., Weatherford, D. R., Carlson, M. A., Bednarz, J. E., & Jones, A. R. (2016). The influence of perpetrator exposure time and weapon presence/timing on eyewitness confidence and accuracy. *Applied Cognitive Psychology, 30*(6), 898-910. doi:10.1002/acp.3275

Cartwright, D., & Festinger, L. (1943). A quantitative theory of decision. *Psychological Review, 50*(6), 595-621. doi:http://dx.doi.org/10.1037/h0056982

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.

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

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology & Aging, 32*(3), 243-258.

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

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law & Human Behavior, 4*, 243–260. doi:doi:10.1007/BF01040617

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

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology, 30*(1), 113-125. doi:10.1002/acp.3178

Eisen, M. L., Smith, A. M., Olaguez, A. P., & Skerritt-Perta, A. S. (2017). An examination of showups conducted by law enforcement using a field-simulation paradigm. *Psychology, Public Policy, & Law, 23*(1), 1-22.

Festinger, L. (1943a). Studies in decision: I. Decision time, relative frequency of judgment, and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology, 32*, 291-306.

Festinger, L. (1943b). Studies in decision: Ii. An empirical test of a quantitative theory. *Journal of Experimental Psychology, 32*, 291-306.

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law, 19*(2), 151-164. doi:10.1037/a0030618

Garrett, B. (2011). *Convicting the innocent: Where criminal prosecutions go wrong. .* Cambridge, MA: Harvard University Press.

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

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

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 15*(2), 140-152. doi:10.1037/a0015082

Hand, D. J. (2014). Never say never. *Scientific American, 310*(2), 72-75. doi:10.1038/scientificamerican0214-72

Hertzog, C., Dunlosky, J., & Sinclair, S. M. (2010). Episodic feeling-of-knowing resolution derives from the quality of original encoding. . *Memory & Cognition, 38*, 771-784. doi:10.3758/MC.38.6.771

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*(1), 94-108. doi:http://dx.doi.org/10.1037/lhb0000060

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.

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

Krafka, C., & Penrod, S. (1985). Reinstatement of context in a field experiment on eyewitness identifications. *Journal of Personality and Social Psychology, 49*, 58-69.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition, 10*(3), 294-340. doi:http://dx.doi.org/10.1006/ccog.2000.0494

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., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions: A problem for the match-to-description lineup foil selection strategy. *Law & Human Behavior, 18*(5), 527-541.

Loftus, E. F., Francis, E., & Turgeon, J. (2012). *Pennsylvania instructions*. Retrieved from http://www.dauphincounty.org/government/Court-Departments/Offices-and-Departments/Court-of-Common-Pleas/Documents/Turgeon/Model-Eyewitness-Identification-Jury-Instructions.pdf

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Mansour, J. K., Beaudry, J. L., Kalmet, N., Bertrand, M. I., & Lindsay, R. C. L. (2017). Evaluating lineup fairness: Variations across methods and measures. *Law and Human Behavior, 41*(1), 103-115. doi:10.1037/lhb0000203

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*(2), 93-102. doi:http://dx.doi.org/10.1016/j.jarmac.2015.01.003

Mickes, L., Clark, S. E., & Gronlund, S. D. (2017). Distilling the confidence-accuracy message: A comment on wixted and wells (2017). *Psychological Science in the Public Interest, 18*(1), 6-9. doi:10.1177/1529100617699240

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

Oriet, C., & Fitzgerald, R. J. (2018). The single lineup paradigm: A new way to manipulate target presence in eyewitness identification experiments. *Law & Human Behavior, 42*(1), 1-12.

Palmer, M. A., & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior, 36*(3), 247-255. doi:10.1037/h0093923

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*(1), 55-71. doi:10.1037/a0031602

Perfect, T. J., & Stollery, B. (1993). Memory and metamemory performance in older adults: One deficit or two? *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 46*(1), 119-135.

Police Executive Research Forum [PERF]. (2013). A national survey of eyewitness identification processes in law enforcement agencies. Washington, dc: Author. Retrieved from http://policeforum.Org/library/eyewitness-identification/nijeyewitnessreport.Pdf.

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

Pryke, S., Lindsay, R. C. L., Dysart, J. E., & Dupuis, P. (2004). Multiple independent identification decisions: A method of calibrating eyewitness identifications. *Journal of Applied Psychology, 89*(1), 73-84. doi:10.1037/0021-9010.89.1.73

Roediger, H. L., Wixted, J., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. P. Sinnott-Armstrong (Eds.), *Oxford series in neuroscience, law and philosophy. Memory and law* (pp. 84-118). New York, NY: Oxford University Press.

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. In T. Valentine & J. P. Davis (Eds.), *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and cctv* (pp. 185-208). Chichester: Wiley Blackwell.

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.

Seale-Carlisle, T. M., & Mickes, L. (2016). Us line-ups outperform uk line-ups. *Royal Society Open Science, 3*(9). doi:10.1098/rsos.160300

Semmler, C., & Brewer, N. (2006). Postidentification feedback effects on face recognition confidence: Evidence for metacognitive influences. *Applied Cognitive Psychology, 20*, 895-916.

Semmler, C., Brewer, N., & Douglass, A. B. (2011). Jurors believe eyewitnesses. In B. L. Cutler (Ed.), *Conviction of the innocent: Lessons from psychological research* (pp. 185 - 209). Washington, D.C.: APA Books.

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied, 24*(3), 400-415.

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.

State v. Henderson, WL 3715028 (2011).

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*(1), 99-139. doi:10.1037/a0021650

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*(8), 798-819. doi:10.1080/1068316X.2015.1054383

Tardif, J., Morin Duchesne, X., Cohan, S., Royer, J., Blais, C., Fiset, D., . . . Gosselin, F. (2019). Use of face information varies systematically from developmental prosopagnosics to super-recognizers. *Psychological Science, 30*(2), 300-308. doi:10.1177/0956797618811338

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*(1), 49. doi:10.1186/s41235-017-0086-z

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law & Human Behavior, 22*, 217–237.

Valentine, T., Hughes, C., & Munro, R. (2009). Recent developments in eyewitness identification procedures in the united kingdom. In R. Bull, T. Valentine, & T. Williamson (Eds.), *Handbook of psychology of investigative interviewing: Current developments and future directions* (pp. 221-240). Chichester, UK: Wiley-Blackwell.

Van Zandt, T. (2000). Roc curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and 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 judgements. *Journal of Experimental Psychology: Applied, 10*, 156-172.

Weber, N., Brewer, N., & Margitich, S. D. (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 Publishers,.

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., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. (2018). *Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence (draft)*. Retrieved from http://ap-ls.wildapricot.org/resources/Documents/Scientific_Review_Paper/APLS%20Scientific%20Review%20Paper%20initial%20draft%20July%2030%202018.pdf

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 photo spreads. *Law and Human Behavior, 22*, 603-647.

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal*

of Applied Research in Memory and Cognition, 4(1), 8-14. doi:https://doi.org/10.1016/j.jarmac.2014.07.003

Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. (2018). The effects of verbal descriptions on performance in lineups and showups. *Journal of Experimental Psychology: General, 147*(1), 113-124.

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., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L., III. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*(6), 515-526. doi:http://dx.doi.org/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*(2), 192-203. doi:http://dx.doi.org/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*(1), 10-65. doi:doi:10.1177/1529100616686966

Zarkadi, T., Wade, K. A., & Stewart, N. (2009). Creating fair lineups for suspects with distinctive features. *Psychological Science, 20*(12), 1448-1453.

Table 1.

Number of Datapoints at the Highest Confidence Level for CACs, for Fair Lineups based on Colloff et al.'s (2017) data (Figure 3).

	Sample size at high confidence[a]	
	Included Data	Excluded Data
Young	42	12
Middle-aged	28	14
Older	18	5

[a] Sample sizes for the fair lineup conditions are approximations because, given no designated innocent suspect, we estimated innocent suspect identification rates (dividing the total target-absent identification rate by nominal size). Thus, for example, 41.5 is entered as 42 (Young witness, Fair lineup, Included data).

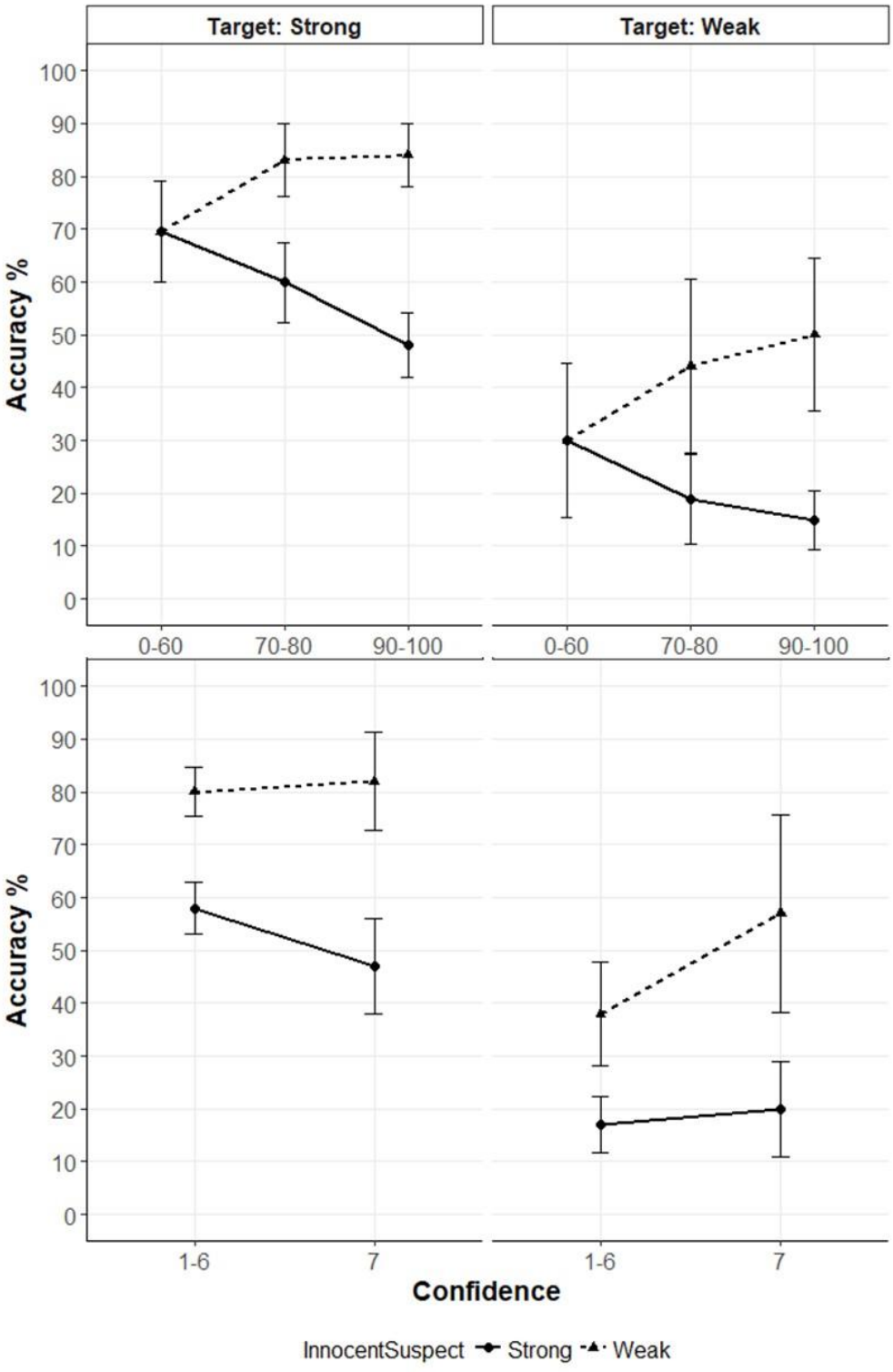

Figure 1. CAC curves according to the strength of match between the target at encoding and test (Target: Strong vs. Target: Weak) and the plausibility of the Innocent Suspect (Strong vs. Weak) based on Gronlund et al.'s (2009) data. The 3-level and 2-level CAC curves are shown in the upper and lower panels, respectively. The values on the x-axis for the 3-level CAC reflect

Mickes' (2015) and Carlson et al.'s (2016) approach of collapsing confidence into three bins: low (0-60%), moderate (70-80%), and high confidence (90-100%). Raw confidence ratings of 1, 2, 3, and 4 were classified as low confidence, a rating of 5 was classified as moderate confidence, and ratings of 6 and 7 were classified as high confidence. The values on the x-axis for the 2-level CAC reflect Wixted et al.'s (2016) approach of treating the highest level of confidence as "high confidence" and everything else as "low confidence". Errors bars represent SEs.

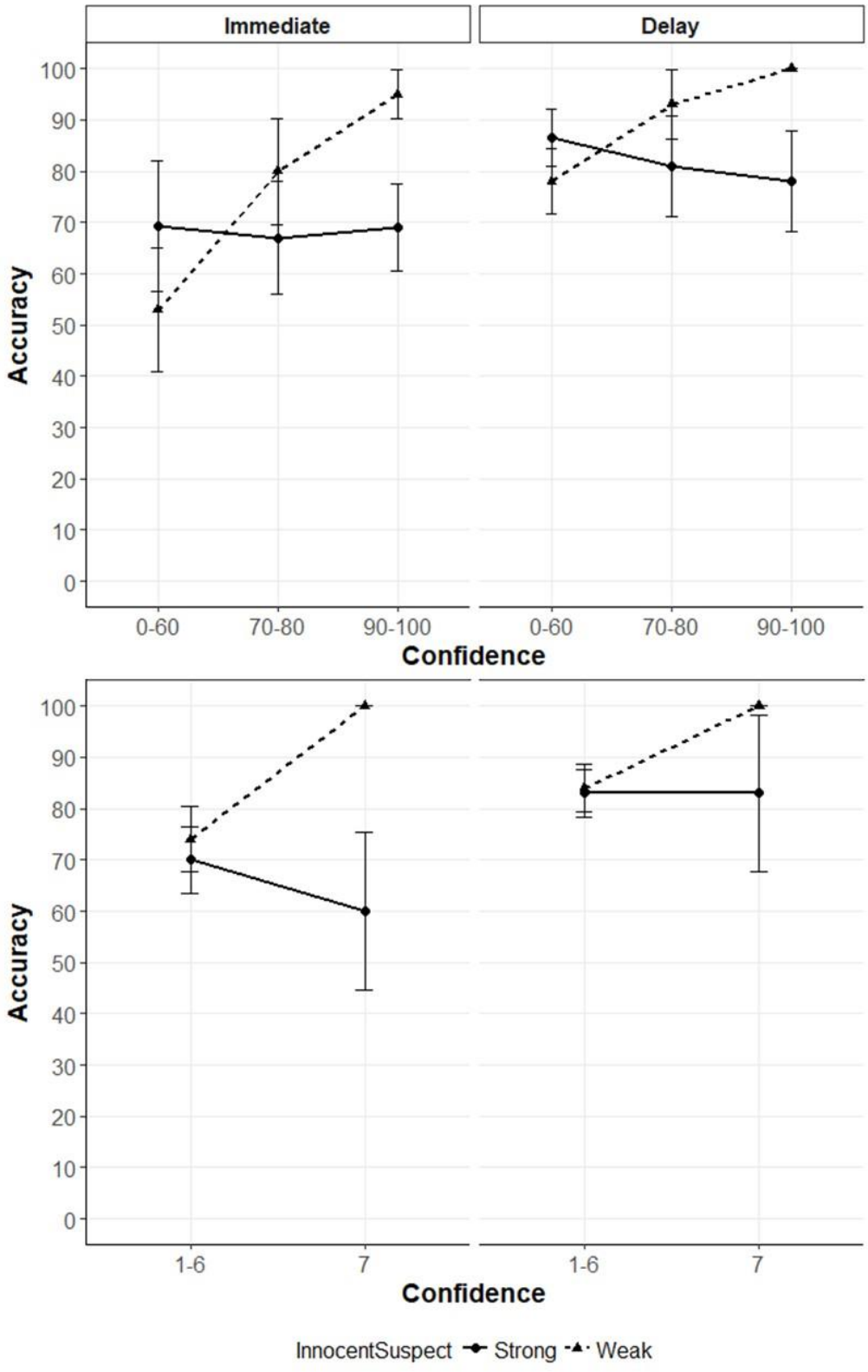

Figure 2. CAC curves according to retention interval; and the plausibility of the Innocent Suspect (Strong vs. Weak) based on Wetmore et al.'s (2015) data. The 3-level and 2-level CAC curves are shown in the upper and lower panels, respectively. Errors bars represent SEs.

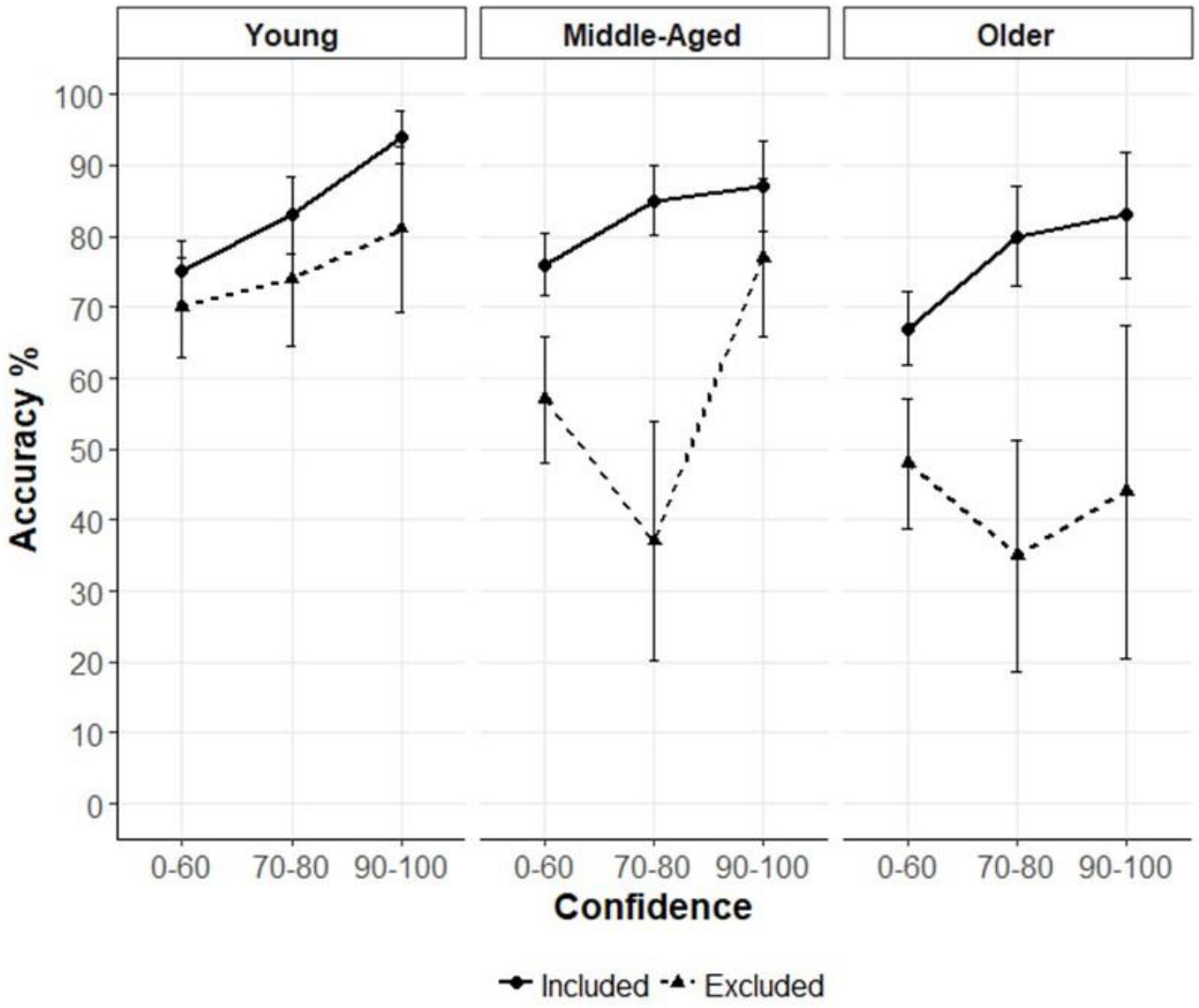

Figure 3. CAC curves according to participant age (18-30 = young; 31-59 = Middle-aged; 60+ = Older), based on data included in, and excluded from, analysis in Colloff et al. (2017) data. Errors bars represent SEs.

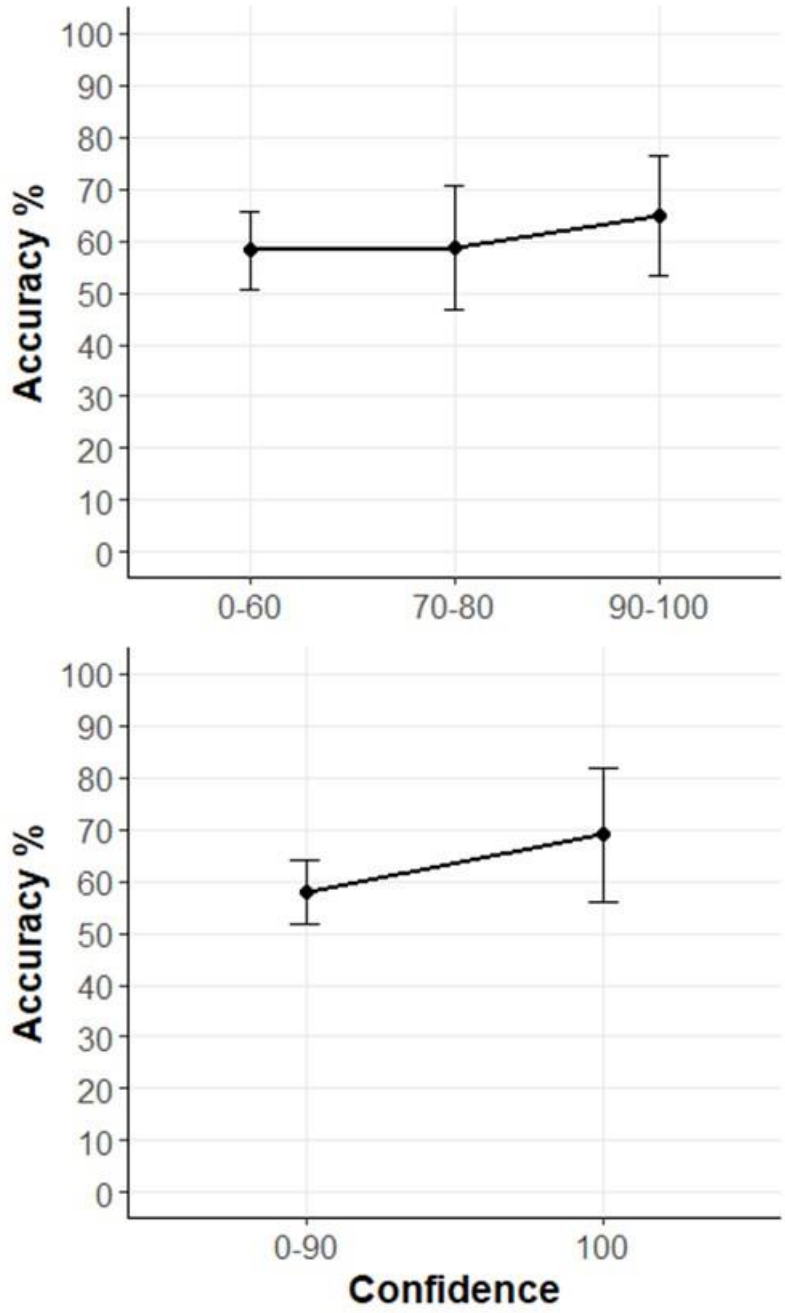

Figure 4. CAC curves Sučić et al.'s (2015) data. The 3-level and 2-level CAC curves are shown in the upper and lower panels, respectively. Errors bars represent SEs.