Eyewitness memory is widely believed to be unreliable because (a) high-confidence eyewitness misidentifications played a role in over 70% of the now more than 300 DNA exonerations of wrongfully convicted men and women, (b) forensically relevant laboratory studies have often reported a weak relationship between eyewitness confidence and accuracy, and (c) memory is sufficiently malleable that, not infrequently, people (including eyewitnesses) can be led to remember events differently from the way the events actually happened. In light of such evidence, many researchers agree that confidence statements made by eyewitnesses in a court of law (in particular, the high confidence they often express at trial) should be discounted, if not disregarded altogether. But what about confidence statements made by eyewitnesses at the time of the initial identification (e.g., from a lineup), before there is much opportunity for memory contamination to occur? A considerable body of recent empirical work suggests that confidence may be a highly reliable indicator of accuracy at that time, which fits with longstanding theoretical models of recognition memory. Counterintuitively, an appreciation of this fact could do more to protect innocent defendants from being wrongfully convicted than any other eyewitness identification reform that has been proposed to date.

Keywords: eyewitness memory, confidence, accuracy

Articles attesting to the unreliability of eyewitness memory appear on a regular basis, both in the popular news media and in the scientific literature. To cite one recent example, a December 2, 2013, article in the Washington Post observed that “Eyewitness misidentifications played a role in the majority of more than 300 convictions that have been overturned because of DNA evidence since 1989” (Hsu, 2013). Indeed, a detailed analysis of those DNA exoneration cases reveals that more than 70% involved eyewitness misidentifications—misidentifications that were virtually always made with high confidence in a court of law (Garrett, 2011a). These unsettling facts seem to comport with forensically relevant laboratory studies that have often reported that the correlation between confidence and accuracy is weak, at best. Krug (2007), for example, summarized research on the confidence–accuracy (CA) relationship in the following way: “Most scientific studies have found the CA relationship to be relatively weak or nonexistent; in fact, this is one of the most consistent findings in the memory research literature” (p. 31). Similarly, Simons and Chabris (2011) recently wrote that “…most memory experts agree that an isolated expression of confidence is at best a limited predictor of memory accuracy” (p. 5). When findings like these are considered together with other research showing that it is possible to experimentally induce false memories of events that never happened (e.g., Loftus, 2005; see Roediger & Gallo, 2002, for a review), the conclusion seems obvious: Eyewitness memory is unreliable.

Despite the unreliability of eyewitness memory, research has shown that jurors find high-confidence eyewitness IDs to be particularly compelling evidence of guilt (e.g., Cutler, Penrod, & Stuve, 1988), the conclusion seems obvious: Eyewitness testimony is fickle and, all too often, shockingly inaccurate,” Arkowitz and Lilienfeld (2010) underscored this point:

For example, jurors tend to give more weight to the testimony of eyewitnesses who report that they are very sure about their identifications even though most studies indicate that highly confident eyewitnesses are generally only slightly more accurate—and sometimes no more so—than those who are less confident.

In light of findings like these, courts in the United States have begun to modify their practices. To take one recent example, in 2011, the New Jersey Supreme Court ruled that if a defendant can show evidence that suggestive police procedures may have influenced the statements of an eyewitness, a pretrial hearing must be held to determine the
admissibility of that evidence (New Jersey Courts, 2012). If the judge ultimately decides that the eyewitness evidence is admissible, jurors must be provided with instructions regarding how to consider a variety of factors that can influence the reliability of eyewitness IDs. One paragraph of the new instructions (to be read by a judge to the jury) reads as follows:

Confidence and Accuracy: You heard testimony that (insert name of witness) expressed his or her level of certainty that the person he or she selected is in fact the person who committed the crime. As I explained earlier, a witness’s level of confidence, standing alone, may not be an indication of the reliability of the ID. Although some research has found that highly confident witnesses are more likely to make accurate identifications, eyewitness confidence is generally an unreliable indicator of accuracy. (New Jersey Model Criminal Jury Charges, 2012, pp. 4–5, emphasis added)

Notice that these instructions fail to draw any distinction between the confidence initially expressed by the eyewitness at the time of the ID versus the confidence expressed months or years later at trial or pretrial hearings (Roediger, Wixted, & DeSoto, 2012). We will argue that the combined weight of theory, empirical evidence, and revelations from DNA exoneration cases converge on the conclusion that initial eyewitness IDs are far more reliable than they have been portrayed in the literature. Moreover—and this is the key point—the failure to appreciate this fact may be impeding a straightforward refinement of current practices that could do more to protect innocent defendants from being wrongly convicted than any other eyewitness ID reform that has been proposed to date (for a review of these other reforms, see Gronlund, Mickes, Wixted, & Clark, 2015). Our recommendation, which can be implemented without changing a single police practice, is as follows:

Jurors should consider the level of certainty expressed by an eyewitness during the initial identifications (at which time confidence is likely to be a reliable indicator of accuracy) while disregarding the level of certainty expressed at trial (because, by then, confidence may no longer be a reliable indicator of accuracy).

This recommendation is by no means unique to us (e.g., Brewer & Palmer, 2010), but its importance seems to have been largely overlooked by the legal system. A recent analysis of DNA exoneration cases conducted by Brandon Garrett as reported in his 2011 book Convicting the Innocent: Where Criminal Prosecutions Go Wrong suggests that a shift in focus from courtroom confidence to initial confidence could avert many wrongful convictions. Garrett (2011a) analyzed trial materials for 161 DNA exonerations who had been misidentified by one or more eyewitnesses in a court of law. His key finding (from our point of view) was described as follows:

I expected to read that these eyewitnesses were certain at trial that they had identified the right person. They were. I did not expect, however, to read testimony by witnesses at trial indicating that they earlier had trouble identifying the defendants. . . . Yet in 57% of these trial transcripts (92 of 161 cases), the witnesses reported that they had not been certain at the time of their earlier identifications. (p. 49, emphasis in original)

If, as we will argue, eyewitness certainty at the time of the initial ID is diagnostic of guilt (i.e., if high confidence = high accuracy and low confidence = low accuracy), then the implication of Garrett’s discovery is that in at least 57% of these cases, expressions of uncertainty on the initial memory test clearly signaled that these eyewitnesses were at high risk of having made an error. This was true even in one of the most famous cases of eyewitness misidentification, one often used to illustrate the unreliability of eyewitness memory. During a trial that was held in 1985, Jennifer Thompson confidently identified Ronald Cotton as the man who had raped her. Cotton was convicted largely on the basis of her testimony, but he was later exonerated by DNA evidence after spending more than 10 years in prison. Long before the trial, however, Thompson’s initial ID of Cotton from a photo lineup was characterized by a prolonged period of hesitation and indecision that lasted for nearly 5 min and that ended with a low-confidence verbal ID consisting of the words “I think this is the guy” (Thompson-Cannino, Cotton, & Torneor, 2009, p. 33; Garrett, 2011b). However, after confirmatory feedback from the police, Thompson became increasingly confident that Cotton was the rapist. From this perspective, the mistake was to rely on confidence expressed at the time of the trial (after it had become improperly inflated) instead of relying on confidence expressed at the time of the initial ID (before memory contamination had a chance to play a significant role).

Our recommendation that jurors should be informed about and rely upon the expression of eyewitness confidence made at the time of the initial ID (because doing so can protect innocent suspects from being wrongly convicted) contrasts with the field’s almost universal recommendation that expressions of confidence should, even under the best of conditions, be given limited consideration. As we describe next, the idea that initial confidence is not strongly related to accuracy conflicts with virtually all empirical evidence—both in the lab and in the real world—that has accumulated over the last 15 to 20 years.

Empirical Research on the Confidence–Accuracy Relationship

The notion that scientific research has established the unreliability of eyewitness confidence was largely set in stone in the 1980s and early 1990s when researchers routinely measured the relationship between confidence and accuracy using a potentially misleading statistic—one that is capable of masking (and, as it turns out, actually did mask) the strong relationship that we now know to exist. Reinforcing the original message about the apparently weak relationship between confidence and accuracy was the fact that, at about the same time that these studies were being reported, news about DNA exonerations of individuals who had been wrongly convicted on the basis of high-confidence eyewitness testimony was just coming to light. These simultaneous developments provided a convincing indict-
ment of the reliability of eyewitness memory. However, although one would not know it from reading psychology textbooks or newspaper articles, nor from surveying the experts (Kassin, Tubh, Hosch, & Memon, 2001; Simons & Chabris, 2011), the story of the relationship between eyewitness confidence and accuracy has changed dramatically since then.

**Early Research Based on the Point-Biserial Correlation Coefficient**

A large body of laboratory-based research using forensically relevant experimental designs (e.g., each eyewitness watches a mock crime video and then attempts to identify the perpetrator from a target-present or target-absent lineup) initially measured the CA relationship using a statistic called the point-biserial correlation coefficient. This approach involves computing a standard Pearson product-moment correlation between the correctness of a response (coded as 0 or 1) and the corresponding confidence rating. In an early review of the literature, Wells and Murray (1984) reported that the point-biserial correlation between confidence and accuracy was only .07, and on that basis they concluded that “...the eyewitness accuracy-confidence relationship is weak under good laboratory conditions and functionally useless in forensically representative settings” (p. 165).

In a later meta-analysis of the literature, Sporer, Penrod, Read, and Cutler (1995) found that the correlation is noticeably higher—approximately .41 on average—when the analysis is limited to only those who make an ID from a lineup (i.e., when the analysis is limited to “choosers”). The restriction to choosers is reasonable because nonchoosers do not end up testifying against the suspect from the lineup they rejected—except perhaps when they are repeatedly tested until they finally do choose the suspect (see later section on the malleability of memory). Nevertheless, even in light of this higher correlation, Penrod and Cutler (1995) still concluded that eyewitness confidence “...is a weak indicator of eyewitness accuracy even when measured at the time an ID is made and under relatively ‘pristine’ laboratory conditions” (p. 830). More recently, Wells and Quinlivan (2009) offered the following assessment of the literature based on the point-biserial correlation coefficient of .41 for choosers: “Suffice to say that psychological scientists have generally concluded that eyewitness certainty, although of limited utility, can have some diagnostic value” (p. 12).

Although the point-biserial correlation is now known to be quite a bit higher than it was once thought to be (see Lindsay, Read, & Sharma, 1998, for evidence that it is higher still when encoding conditions vary across participants), comments like these indicate that prominent figures in the field still consider eyewitness confidence statements to be of limited value. In our view, a correlation of .41 indicates a robust relationship between confidence and accuracy. The point-biserial correlation coefficient is a standard effect-size statistic (e.g., Rosnow, Rosenthal, & Rubin, 2000), and a value of .41 falls between the conventional definitions for medium (.30) and large (.50) effects (Cohen, 1988). A medium-to-large effect size is vastly larger than the effect sizes for other variables thought to be important in the eyewitness ID literature, such as biased versus unbiased lineup instructions \((r = .03)\) and more-versus-less-similar fillers in lineups \((r = .09;\) Clark, 2012). However, instead of debating the meaning of a .41 correlation coefficient, one can measure the relationship of interest in a more informative way.

Addressing this issue from a different perspective, Juslin, Olsson, and Winman (1996) showed that the magnitude of the point-biserial correlation can be very low even when the relationship between confidence and accuracy exhibits perfect calibration. Perfect calibration exists when the level of confidence expressed by an eyewitness corresponds exactly to the percentage of eyewitnesses who are correct when they express that level of confidence. For example, under perfect calibration, witnesses who express 60% confidence in an ID are correct 60% of the time, and witnesses who express 80% confidence in an ID are correct 80% of the time. Even though the relationship between confidence and accuracy could not possibly be stronger than that, Juslin et al. showed that the point-biserial correlation could be low or high, depending on how responses are distributed across the confidence categories. Thus, a low point-biserial correlation coefficient (even one that corresponds to a small effect size) does not necessarily indicate a weak relationship.

**Later Research Based on Calibration**

Juslin et al. (1996) argued that the relationship between confidence and accuracy is best assessed by plotting proportion correct as a function of subjective confidence obtained using a rating scale that ranges from 0 to 100. In a similar CA characteristic analysis, proportion correct is plotted as a function of confidence using any monotonic scale (Mickes, 2015). When plotted in that manner, no special expertise is needed to determine whether the relationship is weak or strong. Moreover, such information corresponds directly to the information that judges and juries most want to know: What is the average accuracy associated with an ID made with a particular level of confidence? A correlation coefficient (even when interpreted as an effect-size measure) does not answer that question, but a calibration plot does.

What is the empirical relationship between confidence and accuracy when data are plotted in the manner suggested by Juslin et al. (1996)? To answer that question, these researchers conducted a mock crime study in which participants watched a videotaped theft and later attempted to identify the guilty suspect from a lineup that was constructed by experienced police officers. As shown in Figure 1, an impressively strong relationship between confidence and accuracy was observed (the diagonal line shows perfect calibration). Despite the near perfect relationship between confidence and accuracy in this study, the point-biserial correlation coefficient was not close to 1.0 (as intuition suggests it should be) but was instead about .45. Again, this value corresponds to a medium-to-large effect size, but if one is inclined to overlook that fact (as many have been in
Observed Relationship Between Proportion Correct and Confidence From Juslin et al. (1996)

Figure 1

Note. The estimated proportion correct data were averaged across their suspect-similarity and culprit-description conditions. Error bars are standard errors of a proportion.

In what is perhaps the most informative study on this issue to date, Palmer et al. (2013) conducted a large-scale, experimentally controlled field study to investigate the relationship between confidence and accuracy under naturalistic conditions. In their study, a pair of confederates approached individuals in various public places, and one confederate asked these potential eyewitnesses whether they would be interested in participating in a psychology experiment. If so, the other confederate (the target) moved into view at a distance of 10 m to be examined by the eyewitness. The participating eyewitnesses (n = 908) were later tested using an eight-person photo lineup, and confidence ratings were taken using a 0-to-100 scale. Their Experiment 1 included four conditions formed by the factorial combination of two study durations (5 s vs. 90 s) and two retention intervals (immediate vs. 1 week). The differences between these conditions were small (particularly for high-confidence IDs), so we combined the data across conditions to examine the relationship between confidence and accuracy.1 Figure 2 presents their results for the 538 witnesses who made an ID from the lineup (i.e., for choosers). In computing the proportion correct measure for this plot, we followed Palmer et al. by counting as errors filler IDs from target-absent lineups while excluding filler IDs from target-present lineups.2 In addition, IDs from the low end of the confidence scale (0–60) were aggregated because relatively few correct suspect IDs from target-present lineups were made in that range. Figure 2 leaves little doubt that an expression of confidence is an extremely informative indicator of the reliability of an ID made from a lineup. The story for choosers becomes even more interesting—and even more relevant to the judges and jurors—when filler errors are excluded from the calculations. To a judge or juror contemplating the innocence or guilt of a suspect who has been identified by an eyewitness, the relevant scientific evidence about the relationship between confidence and accuracy involves choosers who make suspect IDs, not choosers who make filler IDs because filler

1 Combining the data in this manner also facilitates our later comparison of these results with results from police department field studies, which also involved eyewitnesses who were exposed to a wide range of encoding conditions.

2 The accuracy scores would be slightly lower (range = .29 to .75) if filler IDs from target-present lineups were also counted as errors, as Juslin et al. (1996) did.
target-absent lineups. The relationship between suspect ID always and incorrect IDs consists only of suspect IDs from target-present lineups (as in Palmer et al., 2013). 4 Figure 2 shows the proportion of all IDs that were suspect IDs, that is, suspect IDs/(suspect IDs + filler IDs). Clearly, the probability that a suspect was identified increased dramatically with confidence, $\chi^2(1, N = 43) = 11.55, p < .001$. Indeed, almost all of the filler ID errors were made with confidence.

Confidence and Accuracy in Actual Criminal Cases

The advantage of a mock-crime study such as the ones considered earlier is that the experimenter knows if a suspect ID is correct or incorrect, thereby allowing a direct computation of suspect ID accuracy (as in Figure 3). In a police department field study, by contrast, it is not known if a suspect ID is correct or incorrect. Thus, although one can measure how often high-confidence and low-confidence IDs are made to suspects and fillers, a direct calculation of suspect ID accuracy as a function of confidence is not possible. Nevertheless, indirect information about suspect ID accuracy as a function of confidence can be obtained. Specifically, if low-confidence IDs are primarily made to fillers, whereas high-confidence IDs are primarily made to suspects, the results would suggest (albeit not prove) that confidence is predictive of suspect ID accuracy in the real world.

Behrman and Davey (2001) conducted an archival analysis of files from the Sacramento (California) City Police Department for crimes committed between 1987 and 1998. For 58 six-person (live) lineups, eyewitnesses who made an ID were asked to rate their confidence using what is essentially a 2-point scale, with verbal descriptors indicating either high confidence (I am sure . . .) or low confidence (I am not sure, but I think . . .). A suspect ID was made in 29 of these lineups and a filler ID was made in 14. The high number of known errors (i.e., 14 filler IDs, or 33% of all IDs) hardly seems like a ringing endorsement of the reliability of eyewitness memory, but the story changes when the results are broken down by confidence. Figure 4A shows the proportion of all IDs that were suspect IDs, that is, suspect IDs/(suspect IDs + filler IDs). Clearly, the probability that a suspect was identified increased dramatically with confidence, $\chi^2(1, N = 43) = 11.55, p < .001$.

Figure 2 does not present suspect ID accuracy because the dependent measure included all filler IDs from target-absent lineups as errors. We now consider what the relationship between confidence and accuracy is when correct IDs consist of suspect IDs from target-absent lineups (as always) and incorrect IDs consist only of suspect IDs from target-absent lineups. The relationship between suspect ID accuracy and confidence is shown in Figure 3. The results are notable in that even low-confidence suspect IDs are fairly likely to be correct (about 83% correct), though most would probably agree that the 17% error rate is too high to justify a conviction based on a low-confidence ID alone. Some might regard a high-confidence ID, at nearly 98% correct, as being accurate enough to do so, but it is important to keep in mind that this study did not measure the CA relationship for memories formed during a real crime. To what extent are these findings representative of the performance of real eyewitnesses involved in the emotionally arousing and sometimes life-threatening conditions of an actual crime?

\[ \text{Proportion Correct} = \frac{\text{correct suspect IDs}}{\text{correct suspect IDs} + \text{incorrect suspect IDs}}, \]

where the sum of suspect IDs from target-present lineups and filler IDs from target-absent lineups. The data are from Palmer et al. (2013); error bars are standard errors.
low confidence; for high-confidence IDs, 18 out of 19 (95%) were suspect IDs.

Then again, there are a number of reasons to be cautious about the Behrman and Davey (2001) results. For example, the study involved a small number of witnesses, some of the witnesses in that study may have had preexisting familiarity with the suspect, the fairness of the lineups was not assessed, and the lineups were administered by an investigator who knew the identity of the suspect (and who may therefore have influenced the decision made by the eyewitness). Fortunately, a new large-scale police department field trial addressed all of these methodological issues.

W. Wells (2014) recorded eyewitness decisions from six-person photo lineups administered as part of criminal investigations in the Robbery Division of the Houston (Texas) Police Department between January 22 and December 5, 2013. We focus here on their “blind” condition (which is more commonly referred to as the “double-blind” condition). This condition involved a large number of simultaneous and sequential lineups, the investigators were unaware of the identity of the suspect in the lineup, and the lineups involved suspects who were unknown to the eyewitnesses.5 Eyewitnesses who made a suspect ID or a filler ID were asked to supply a confidence rating on a 3-point scale (positive, strong tentative, or weak tentative), but the data were reported with the two tentative confidence levels combined. Thus, the reported data correspond to a 2-point confidence scale (high confidence = positive vs. low confidence = the two tentative categories combined). Out of the 349 lineups administered, an ID was made by 219 witnesses. Of these, 115 were suspect IDs and 104 filler IDs. Once again, on the surface, these results appear to attest to the unreliability of eyewitness memory because a filler ID (which is an ID of a known innocent) was made almost as often as a suspect ID. However, as with the Sacramento field study, when the data are plotted as a function of confidence, a different interpretation emerges (Figure 4B). Evidently, low-confidence IDs involve many erroneous filler IDs, but this is much less true for high-confidence IDs, $\chi^2(1, N = 219) = 48.45, p < .001$.

Again, the data shown in Figures 4A and 4B do not directly represent the accuracy of suspect IDs (i.e., they do not represent the proportion of suspect IDs that were correct—which is the critical information) but instead represent the proportion of filler IDs + suspect IDs that were suspect IDs (correct and incorrect alike, because it is not known which is which). However, we can use the experimentally controlled field data from Palmer et al. (2013) to gain further insight into what these data may imply about suspect ID accuracy. Figure 4C shows the Palmer et al. (2013) data collapsed across target-present and target-absent lineups—as if these data were collected as a part of a police department field study with unknown lineup type. The values shown in Figure 4C differ slightly from the values shown earlier in Figure 2 in order to make them compatibly scaled with the two police department field studies. Specifically, the confidence scale has been reduced to a 2-point scale, low (0–80) versus high (90–100), and the dependent measure consists of all suspect IDs (i.e., correct suspect IDs + estimated incorrect suspect IDs)—not just correct suspect IDs—divided by all lineup IDs (suspect and filler IDs alike). A highly significant effect of confidence is evident in Figure 4C, $\chi^2(1, N = 538) = 45.05, p < .001$, just as was true of the Sacramento and Houston field data (Figure 4A and Figure 4B, respectively).

In all three data sets, the proportion of IDs that are suspect IDs increased dramatically with confidence. Unlike the police department field data shown in Figures 4A and 4B, we can unpack the collapsed Palmer et al. (2013) data shown in Figure 4C to determine how the pattern shown there relates to the relationship of primary interest to the legal system, namely, the relationship between confidence and suspect ID accuracy. We can do so by comparing the collapsed (police-department-like) data from Palmer et al. (2013) shown in Figure 4C to the uncollapsed suspect ID accuracy data from Palmer et al. (2013) shown earlier in Figure 3. Importantly, and perhaps counterintuitively, this comparison reveals that the observed proportion of high-confidence suspect IDs in Figure 4C (approximately 80%) translates into a very high percentage of correct high-confidence suspect IDs, as shown earlier in Figure 3 (98% of the high-confidence suspect IDs were correct). The observed proportion of low-confidence suspect IDs in Figure 4C (approximately 40%) translates into a reasonably high, but noticeably more error prone, percentage of correct suspect IDs (83% of the low-confidence suspect IDs were correct).

---

5 A random subsample of 30 of these lineups was assessed for lineup fairness using mock witnesses supplied only with the description of the perpetrator. These lineups were found to be fair in that, on average, the suspect was identified about one sixth of the time (about what would be expected in a fair six-person lineup).
To the extent that the Palmer et al. (2013) data generalize to real police department field data, it seems reasonable to infer that high-confidence suspect IDs in the police department field studies are also highly accurate and that low-confidence suspect IDs are less accurate (though perhaps still informative). However, the base rate of target-present lineups in the Palmer et al. (2013) study was 50%, whereas the corresponding base rates for target-present lineups in the Sacramento and Houston field studies are unknown. If the base rate of target-present lineups in either or both of the police department field studies is less than 50% (e.g., 25%), then base-rate analyses reported by Brewer and Wells (2006) suggest that the accuracy of low-confidence suspect IDs would be substantially lower than that depicted in Figure 3, and the accuracy of high-confidence suspect IDs would be slightly lower as well. Nevertheless, it seems fair to infer from these data that the relationship between confidence and accuracy in the real world mirrors the impressive relationship observed in experimentally controlled research in the sense that low-confidence IDs (the kind of IDs that have often led to wrongful convictions) are relatively error prone, whereas high-confidence IDs are much less so.

Theoretical Considerations (Signal Detection Theory)

With a few exceptions, eyewitness memory researchers have tried to make sense of the relationship between confidence and accuracy without any guidance from prominent theories of recognition memory. By contrast, in the basic cognitive psychology literature, signal detection theory (SDT) is often used to conceptualize the level of confidence associated with recognition memory decisions (Bernbach, 1967; Egan, 1958). Under typical experimental conditions, this theory has long been known to predict a strong relationship between confidence and accuracy, and many studies have shown that the relationship is in fact strong when an old/new recognition memory task is used (see Mickes, Hwe, Wais, & Wixted, 2011, for a recent example). In the typical situation, students study a set of unrelated words then are tested on new words, half previously studied (old) and half not (new), with instructions to make old/new judgments and to provide confidence ratings for each judgment. Under these conditions, the relationship between confidence and accuracy is almost always strong in that high confidence implies high accuracy (often in the range of 90%–100% correct) and low confidence implies low accuracy (often close to 50% correct—no better than chance). When special steps are taken by the experimenter to induce false memories, SDT is also useful for conceptualizing why the typically strong relationship between confidence and accuracy breaks down and can even be reversed (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014).

Because a lineup is a special type of recognition test, one in which a witness views a variety of alternatives and then makes a decision to either identify one person or to reject the lineup, SDT is directly applicable to the eyewitness domain. In fact, the major theoretical model of lineup memory—Clark’s WITNESS model (Clark, 2003; Clark, Erickson, & Breneman, 2011)—makes use of signal-detection logic. That being the case, under standard conditions used in studies of eyewitness memory (e.g., reasonably good memory conditions, fair lineups, no undue influence on the witness, etc.), the model naturally predicts a strong relationship between confidence and accuracy for initial lineup-based decisions. Thus, the recent empirical revelations about the strong relationship between confidence and accuracy in studies of (initial) eyewitness ID resolves what had previously seemed like a curious discrepancy between...
Why Focus on Initial Versus Later Confidence Statements? Because Memory Is Malleable

The fact that memory is malleable—that is, that people can be induced to confidently remember events that never happened—has been firmly established by decades of research (see Roediger, 1996). According to our reading of the literature, there is no controversy about this. In her now classic work on misinformation effects, Elizabeth Loftus repeatedly showed that people who witness an event (e.g., a simulated car accident) can be led to believe that they saw an object they did not actually see if an interviewer simply mentions the object during questioning. For example, witnesses can be led to believe that they saw a yield sign when in fact they saw a stop sign (Loftus, Miller, & Burns, 1978). Other research by Loftus and her colleagues showed that eyewitnesses may misremember critical details of a simulated car accident following a subtly suggestive question regarding the car’s speed (Loftus & Palmer, 1974).

Subsequent research produced even more dramatic examples of false memories. Hyman, Husband, and Billings (1995) showed that some people could be led to believe that they had been hospitalized when in fact they had not, and Loftus and Pickrell (1995) led about 25% of their subjects to falsely believe, for example, that they had been lost and frightened in a shopping mall as a child. Often, these false memories are remembered with high confidence (e.g., Porter, Yuille, & Lehman, 1999).

Another problem, one that is closely related to the malleability of memory, has also been found to plague eyewitness IDs from a lineup. Wells and Bradfield (1998, 1999) found that postidentification feedback to eyewitnesses (i.e., a comment from the lineup administrator suggesting that the ID was accurate) increased their later recollections of how certain they had been at the time of their initial ID (i.e., before receiving any feedback). Moreover, Bradfield, Wells, and Olson (2002) found that confirming feedback (“you are right!”) had a larger influence on inaccurate witnesses than it did on accurate witnesses, “... thereby reducing the usefulness of retrospective certainty reports as cues to identification accuracy” (p. 117). It is worth emphasizing that confirming feedback can immediately inflate confidence associated with an initial ID, which is what happens to appear in the Jennifer Thompson/Ronald Cotton case (Garrett, 2011b). When her nearly 5 min of apparent indecision ended with the statement “I think this is the guy,” the detective, dissatisfied with that tentative ID, said, “You ‘think’ that’s the guy?” She replied by saying, “It’s him.” The detective then suggested a specific level of certainty by saying, “You’re sure?” Thompson replied that she was positive, but she still seemed plagued by doubts, seeking reassurance that she had correctly identified the perpetrator. “Did I do OK?” she asked, and the detectives answered, “You did great, Ms. Thompson.” By the time of the trial, Thompson’s initial doubts were gone. In our view, those initial doubts spoke volumes.

In addition to distortions caused by confirming feedback, repeated testing of face memory can have the effect of increasing confidence that an innocent suspect is the perpetrator (e.g., the mugshot exposure effect, Deffenbacher, Bornstein, & Penrod, 2006). In fact, studies using a variety of paradigms have shown that false memories can arise during repeated testing of memory (Bartlett, 1932; McDermott, 2006), and this is especially so if suggestive questioning is involved (Roediger, Jacoby, & McDermott, 1996; see Roediger, McDermott, & Goff, 1997, for a review and discussion).

The malleability of memory appears to account for why so many of the DNA exoneration cases started out with a low-confidence ID of the innocent suspect and ended up with a high-confidence ID of the same innocent suspect during the trial (Garrett, 2011a). The initial low-confidence ID presumably occurred because the match between the memory of the perpetrator and the suspect was weak (which makes sense given that the suspect was not the perpetrator). The later high-confidence ID presumably occurred because the eyewitness’s memory of the innocent defendant was artificially strengthened (or confidence was artificially inflated) by events that occurred between the initial ID and the trial for the reasons described above (e.g., repeated testing, feedback, etc.). Those mistaken courtroom IDs have often been taken to unambiguously establish the inherent unreliability of eyewitness memory, but they are perhaps better interpreted as reflecting the well-intentioned but unjustified handling of eyewitnesses by the criminal justice system. It was, after all, the criminal justice system (from the police to the courts) that (a) ignored the initial reliable (low) confidence statement made by the eyewitness, (b) unintentionally contaminated the eyewitness’s memory, and (c) took under advisement only the unreliable ID made by the eyewitness during a criminal trial.

How Should Eyewitness Confidence Be Treated in the Courtroom?

In partial agreement with what we are advocating here, Garrett (2012) recently made the following recommendation: “Directing my observations to criminal procedure reformers, I argue that courtroom identifications following prior identifications should be per se excluded” (p. 457). We understand these sentiments but emphasize that the key factor that should be excluded is the courtroom expression of confidence. In fact, our only real departure from Garrett’s view concerns the role of certainty statements in the legal system. In its Manson v. Brathwaite (1977) ruling, the Supreme Court developed a two-pronged test for the admissibility of eyewitness evidence. The trial court would first determine if eyewitness ID procedures were suggestive and, if so, would weigh various reliability factors—some of which were borrowed from its earlier decision in U.S. v. Wade (1967)—to determine if the eyewitness evidence was...
An initial low-confidence ID implies low accuracy, but it does not necessarily imply chance accuracy. Even when the ID is made with low confidence, more guilty suspects than innocent suspects may be identified (e.g., see Figure 3). Moreover, the subsequent inflation of confidence that occurs for innocent suspects who end up being wrongfully convicted also presumably occurs for guilty suspects who are rightfully convicted. Indeed, serial killer Ted Bundy represents an example. In November of 1974, Bundy attempted, and failed, to kidnap Carol DaRonch. Nearly a year later, police presented her with a photograph of Bundy, who she identified as her attacker. However, it was a very tentative ID that was clearly made with low confidence (Loftus & Ketcham, 1991, pp. 61–91). DaRonch’s confidence increased when she was subsequently shown another photograph of Bundy (perhaps because of repeated testing), and then she picked him out of a live lineup. Bundy waived his right to a jury trial, and the judge convicted him largely on the basis of DaRonch’s eyewitness testimony. Had the judge instead acted in accordance with what we are suggesting here (i.e., by taking under advisement confidence in the initial ID), Bundy may have been released to continue his killing spree. Instead, he was executed in 1989 after confessing to over 30 murders.

These considerations suggest that if the focus were shifted to certainty expressed at the time of the initial ID, and if jurors behaved as they do now (exhibiting reluctance to convict based solely on a low-confidence ID), then fewer innocent defendants would be wrongfully convicted than before (a clear benefit) but more guilty defendants would be wrongfully acquitted than before (a clear cost). If that happened, would society judge the cost to be worth the benefit? As noted by Clark (2012), the trade-off between false convictions and false acquittals presents a complex problem of social science and social values and requires careful consideration by policymakers and the criminal justice system. If the cost in terms of lost convictions of the guilty were ultimately deemed to be too high when measured against the benefit in terms of fewer false convictions, then jurors might become inclined to convict even if the initial ID was not made with the highest level of confidence. But jurors—and, more generally, the legal system—should weigh these issues using the reliable confidence statement that they have available to them, the initial statement of confidence, not the less reliable confidence statement that is more typically presented to them at the time of the trial.

Recording Confidence Associated With the Initial ID

Because eyewitness confidence on an initial test of memory is diagnostic of guilt (whereas eyewitness confidence on a later memory test may not be), it is important that the police record the initial level of confidence expressed by an eyewitness. Eyewitness memory researchers have long made the same recommendation not only because the initial ID is more reliable (e.g., Brewer & Palmer, 2010) but also because it provides a way to detect the possible inflation of...
agencies implement other recommendations from the Na-

goal may become easier in the future as law enforcement

focus is exactly what needs to happen. Accomplishing this

ously. Nonetheless, if our analysis is correct, this shift in

an ID from a line-up that occurred months or years previ-

have a hard time ignoring that expression of confidence and

jurors see a highly confident witness on the stand, they may

predictor of accuracy.

results considered earlier indicate that, under those condi-

tions, initial eyewitness confidence appears to be a strong

predictor of accuracy.

A new report from the National Academy of Sciences

on the science of eyewitness identification makes several

recommendations that correspond closely to our central

message. One recommendation is that investigators should

document witness confidence judgments at the time of the

initial ID because “Evidence indicates that self-reported

confidence at the time of trial is not a reliable predictor of

accuracy. The relationship between the witness’ stated con-

fidence and accuracy of identifications may be greater at

the moment of initial identification than at the time of trial”

(National Research Council, 2014, p. 74). Another closely

related recommendation is that judges should “... take all

necessary steps to make juries aware of prior identifica-

tions, the manner and time frame in which they were

conducted, and the confidence level expressed by the eye-

witness at the time” (National Research Council, 2014, p.

76). Also, because initial confidence reliably predicts accu-

racy only if the ID and the confidence statement made by

the eyewitness are not influenced by the investigating of-

ficer, the National Academy report recommended using a

blind lineup administrator (p. 73). The Houston field study

results considered earlier indicate that, under those condi-

tions, initial eyewitness confidence appears to be a strong

predictor of accuracy.

Although we recommend that only confidence in the

initial ID be taken under advisement, we recognize that if

jurors see a highly confident witness on the stand, they may

have a hard time ignoring that expression of confidence and

focusing instead on the expression of initial confidence in

an ID from a line-up that occurred months or years previ-

ously. Nonetheless, if our analysis is correct, this shift in

focus is exactly what needs to happen. Accomplishing this

goal may become easier in the future as law enforcement

agencies implement other recommendations from the Na-

goal. For example, embracing another idea long espoused by eyewitness ID researchers (e.g.,

Sporer, 1992; Kassin, 1998), the National Academy report

recommended videotaping the witness ID process (Na-

tional Research Council, 2014, p. 74). With videotaped

evidence of the initial ID, jurors would be able to see for

themselves just how confident the eyewitness was at the

outset, which may help them take that confidence judgment

under advisement while disregarding the less reliable con-

fidence judgment made at the time of the trial.

What kind of confidence scale should the police use to

assess initial confidence? The usual recommendation is that

the witness should provide a confidence statement in his or

her own words and that the statement be recorded verbatim

(e.g., Technical Working Group for Eyewitness Evidence,

1999). This longstanding recommendation was also en-
dorsed by the National Academy report (National Research

Council, 2014, p. 74). However, this approach seems ins-
ufficiently precise to us, in part because it makes it diffi-
cult to bring research findings to bear on how accurate a

particular expression of confidence might be, on average. A

better approach, following Behrman and Davey (2001),

would be to routinely ask the eyewitness to indicate which

statement applies to their memory-based assessment of

the people in the lineup, such as (a) “I am positive that number ___ is the person who committed the crime”; (b) “Although I am not positive, I think that number ___ is the person who committed the crime”; (c) “I do not recog-
nize anyone in the lineup as being the person who com-
mitted the crime.” As noted earlier, this scale amounts to a
2-point confidence scale for choosers. Alternatively, a
3-point confidence scale for choosers along the lines of the scale used by W. Wells (2014) could be used. Both of these scales have the distinct advantage of having been tested (and found to be useful) in actual criminal cases involving real eyewitnesses. An important goal for future research

will be to determine if a more fine-grained confidence scale—such as the 100-point scale that is often used in calibration studies—is feasible in the complex environ-
ments in which the police have to operate (e.g., dealing with witnesses who have a wide range of educational backgrounds, who vary in their ability to understand Eng-
lish, who may be highly stressed, etc.). For the time being,

the only empirical tests of the utility of a confidence scale

in a police setting used by real eyewitnesses involved the

simpler scales used by Behrman and Davey (2001) and W.

Wells (2014). Given how easily a 2- or 3-point scale could

be implemented, we see little reason not to do so.

Conclusion

The relationship between eyewitness confidence and accu-

racy for an initial eyewitness ID is higher than many

psychologists have claimed it to be and is certainly higher

than many in the media and in the legal system understand

it to be. Other psychologists (e.g., Brewer & Palmer, 2010;

Lindsay et al., 1998) have argued—correctly, in our view—that the best scientific evidence suggests that low

confidence implies low accuracy, and high confidence im-

plies high accuracy so long as confidence is assessed at the
time of the initial ID (not later, in court). Still, the message

seems not to have been received by the legal system, as

indicated, for example, by the newly adopted jury instruc-

tions in New Jersey that flatly declare that “eyewitness

confidence is generally an unreliable indicator of accu-

racy,” with no distinction drawn between an initial ID and a

later ID.

Our goal in writing this article was to bring together the

results of experimentally controlled research, police
department field studies, case-outcome analyses of DNA exonerees, and theoretical considerations to drive home a point that we believe needs to be understood better than it is now: On an initial memory test, eyewitnesses memory is reliable in the sense that confidence and accuracy are strongly related. Moreover, and critically, the strong relationship between initial confidence and accuracy means that initial IDs made with low confidence imply a high error rate. This, in turn, implies a high probability of convicting an innocent suspect when an initial low-confidence ID is overlooked in favor of a high-confidence ID at trial. Instead of discounting expressions of confidence altogether, a much better approach would be to treat the initial expression of confidence as a reliable indicator of accuracy. This is the take-home message of our article. Had testimony been focused (as we advocate) on confidence in the initial ID, many of the eyewitnesses involved in the DNA exoneration cases may not have persuaded jurors that guilt was established beyond a reasonable doubt. Thus, far from being the problem, confidence statements at the time of an initial lineup may be a big part of the solution to false convictions based on eyewitness misidentifications.

REFERENCES


