Check for updates

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science in the Public Interest
2017, Vol. 18(1) 10–65
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1529100616686966
www.psychologicalscience.org/PSPI
SAGE

The Relationship Between Eyewitness Confidence and Identification Accuracy: A New Synthesis

John T. Wixted[1] and Gary L. Wells[2]
[1]Department of Psychology, University of California, San Diego, and [2]Department of Psychology, Iowa State University

Summary
The U.S. legal system increasingly accepts the idea that the confidence expressed by an eyewitness who identified a suspect from a lineup provides little information as to the accuracy of that identification. There was a time when this pessimistic assessment was entirely reasonable because of the questionable eyewitness-identification procedures that police commonly employed. However, after more than 30 years of eyewitness-identification research, our understanding of how to properly conduct a lineup has evolved considerably, and the time seems ripe to ask how eyewitness confidence informs accuracy under more pristine testing conditions (e.g., initial, uncontaminated memory tests using fair lineups, with no lineup administrator influence, and with an immediate confidence statement). Under those conditions, mock-crime studies and police department field studies have consistently shown that, for adults, (a) confidence and accuracy are strongly related and (b) high-confidence suspect identifications are remarkably accurate. However, when certain non-pristine testing conditions prevail (e.g., when unfair lineups are used), the accuracy of even a high-confidence suspect ID is seriously compromised. Unfortunately, some jurisdictions have not yet made reforms that would create pristine testing conditions and, hence, our conclusions about the reliability of high-confidence identifications cannot yet be applied to those jurisdictions. However, understanding the information value of eyewitness confidence under pristine testing conditions can help the criminal justice system to simultaneously achieve both of its main objectives: to exonerate the innocent (by better appreciating that initial, low-confidence suspect identifications are error prone) and to convict the guilty (by better appreciating that initial, high-confidence suspect identifications are surprisingly accurate under proper testing conditions).

Keywords
calibration, confidence and accuracy, eyewitness identification, eyewitness memory, lineups, wrongful convictions

Introduction

In his book *On the Witness Stand: Essays on Psychology and Crime*, Hugo Münsterberg (1908) warned about the unreliability of eyewitness memory. As it turns out, he was prescient. Since 1989, 349 wrongful convictions have been overturned through DNA testing, and eyewitness misidentification played a role in over 70% of those cases—far more than any other contributing cause (Innocence Project, 2016). No one doubts that the large majority of these misidentifications were made in good faith. Somehow, these eyewitnesses came to honestly but mistakenly believe that the innocent defendant was the person who committed the crime. How did that happen? The short explanation is that the procedures used for testing eyewitness identification were not developed and validated in the scientific laboratory before being implemented in the field. Instead, they were developed

Corresponding Authors:
John T. Wixted, Department of Psychology, University of California, San Diego. La Jolla, CA 92093
E-mail: jwixted@ucsd.edu
Gary L. Wells, Psychology Department, West 112 Lagomarcino Hall, Iowa State University, Ames, IA 50021
E-mail: glwells@iastate.edu

within the criminal justice system and implemented under the mistaken assumption that they accurately identified the guilty without unduly jeopardizing the innocent.

When experimental psychologists began to empirically investigate the validity of these identification procedures in the 1970s and 1980s, they soon discovered that many seemed tailor-made for eliciting high-confidence misidentifications. For example, nowadays, a typical photo-lineup identification procedure consists of the simultaneous or sequential presentation of one photo of the suspect (the person the police believe may have committed the crime) and five or more fillers (photos of people who are known to be innocent but who physically resemble the suspect). Such a lineup offers protection to an innocent suspect because a witness who chooses randomly is far more likely to land on a filler than the suspect. However, before the dangers of eyewitness misidentification were understood, an investigating officer might present a lineup consisting of only suspects (with no fillers) and tell a witness who had just identified one of the suspects with low confidence that it was clearly the right decision, resulting in a higher expression of confidence the next time the witness was asked about it. By the time of the trial, the jury would see the witness honestly misidentify the suspect with high confidence and convict on that basis alone, often sending an innocent person to prison. Practices like these help to explain why, in every one of the DNA exoneration cases involving eyewitness misidentification examined by Garrett (2011), witnesses who mistakenly identified innocent defendants did so with high confidence when the case was tried in a court of law.

But what about the confidence expressed by an eyewitness tested using the scientifically validated procedures that have been developed over the years by eyewitness-identification researchers? That is the question we focus on here, and the answer will undoubtedly come as a surprise to many. Understandably, the disproportionate role played by eyewitness misidentification in the DNA exoneration cases has helped to create a widespread impression that eyewitness memory is unreliable even under the best of circumstances (i.e., that it is *inherently* unreliable). But over the last 20 years, eyewitness-identification researchers have discovered that when eyewitnesses are tested using appropriate identification procedures, the confidence they express can be, and usually is, a highly reliable indicator of accuracy (Brewer & Wells, 2006; Juslin, Olsson, & Winman, 1996). However, over that same period of time, the legal system has increasingly come to interpret the scientific literature as indicating no meaningful relation between confidence and accuracy. As a result, some courts now advise juries to disregard eyewitness expressions of confidence and to focus instead on a variety of other factors when trying to assess the reliability of an ID. The purpose of our article is to explain why a blanket disregard for eyewitness confidence not only is at odds with what has been learned in recent years but also can contribute both to the wrongful conviction of innocent suspects and to the unwarranted removal from suspicion of a guilty suspect.

Our article is organized as follows: We first document a growing trend within the legal system to disregard eyewitness confidence, with no distinction drawn as to whether the eyewitness-identification procedures were appropriate or not and with no distinction drawn between witness confidence at the time of the initial identification versus witness confidence at a later time. Next, we review a recommended set of appropriate ("pristine") identification procedures that have been developed in eyewitness-identification laboratory studies and how these pristine procedures can operate to prevent other factors from contaminating eyewitness confidence. The general idea is that a strong relation between confidence and accuracy is the natural state of affairs, but there are various things that can contaminate that relation. We then consider the nontrivial issue of how best to measure the confidence-accuracy relationship, followed by a detailed review and reanalysis of the empirical literature on the confidence-accuracy relation. The results will show that when pristine identification procedures are used, eyewitness confidence is a highly informative indicator of accuracy, and high-confidence suspect identifications are highly accurate. We go on to demonstrate that the confidence-accuracy relationship can be compromised when certain non-pristine identification procedures are used, and we enumerate priorities for future research on the confidence-accuracy relationship.

How Eyewitness Confidence Is Understood in the Legal System

In the legal system, eyewitness confidence is increasingly distrusted. For example, the state of New Jersey recently adopted jury instructions declaring that "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 Courts, 2012a; New Jersey Courts, 2012b). The report upon which the New Jersey instructions were based (Report of the Special Master, *State v. Henderson*, 2011) categorically asserted that "studies uniformly show, and the experts unanimously agree, that confidence is not closely correlated to accuracy" (p. 79). When discussing confidence, no distinction was drawn between identification procedures that are pristine and

those that are not. These jury instructions are, of course, accurate when applied to problematic eyewitness-identification procedures, but our question concerns the confidence-accuracy relationship when pristine procedures are used early in the investigation and prior to any memory contamination. As our review will demonstrate, there are known conditions under which confidence clearly informs accuracy and other known conditions under which it clearly does not.

A bleak view of eyewitness confidence is not in any way limited to New Jersey. Other jurisdictions have revised their jury instructions so as to encourage juries to place little faith in eyewitness confidence. In Massachusetts, for example, the relevant instructions stipulate that "a witness's expressed certainty in an identification, standing alone, may not be a reliable indicator of the accuracy of the identification, especially where the witness did not describe that level of certainty when the witness first made the identification" (Massachusetts Court System, 2015, pp. 5–6). These instructions appropriately focus on the importance on the initial identification, but they do not appropriately communicate the high information value of an initial statement of confidence obtained from a pristine identification procedure.

Next, consider this recent statement made by the Connecticut Supreme Court in *State v. Guilbert* (2012): "Courts across the country now accept that there is at best a weak correlation between a witness's confidence in his or her identification and its accuracy." In a subsequent case, the Connecticut Psychological Association filed an amicus brief with the state supreme court arguing that eyewitness confidence is so loosely correlated with accuracy that it should no longer serve as a criterion for evaluating the reliability of eyewitness identification (Berard, 2014). No distinction was made between the confidence of the witness at the time of identification and the confidence of the witness at trial. Similarly, in *Brodes v. State* (2005), the Georgia Supreme Court held that jury instructions should not encourage jurors to consider a witness's confidence when trying to determine the reliability of an ID, specifically citing scientific research on the correlation between confidence and accuracy:

> In light of the scientifically-documented lack of correlation between a witness's certainty in his or her identification of someone as the perpetrator of a crime and the accuracy of that identification, and the critical importance of accurate jury instructions as "the lamp to guide the jury's feet in journeying through the testimony in search of a legal verdict," we can no longer endorse an instruction authorizing jurors to consider the witness's certainty in his/her identification as a factor to be used in deciding the reliability of that identification. Accordingly, we advise trial courts to refrain from informing jurors they may consider a witness's level of certainty when instructing them on the factors that may be considered in deciding the reliability of that identification.

Again, no distinction was made by the court between the confidence of the witness at the time of identification and the confidence of the witness at trial. Along the same lines, in *State v. Mitchell* (2012), the Utah Supreme Court recently stated,

> In the end, we agree with the Connecticut Supreme Court that the available studies are not definitive on the question whether there is a significant correlation between certainty and accuracy. But we are also mindful that the literature suggests certainty may not always be as reliable an indicator of accuracy. . . Therefore, we hold it is error to instruct the jury on the degree of certainty factor, and we discourage its future use.

Undeniably, eyewitness certainty at pretrial hearings or at trial should be highly suspect for reasons we will discuss. But when a lineup is conducted under pristine testing conditions and the confidence statement of the witness is taken at the time of identification, the data indicate that confidence is a reliable indicator of accuracy.

The fact that courts increasingly distrust eyewitness confidence is not altogether surprising, given that expert witnesses and concerned organizations routinely paint a gloomy picture of the confidence-accuracy relationship. For example, a 2013 amicus brief filed by the Innocence Project said, "A witness's confidence bears, at best, a weak relationship to accuracy" (Innocence Project, 2013, p. 11). However, the evidence we will review suggests that "at best" (i.e., under pristine conditions), a witness's confidence bears a strong relationship to accuracy.

It is not just the Innocence Project that has a generally pessimistic view of the confidence-accuracy relationship. A recent amicus brief filed by the American Psychological Association painted a similarly bleak picture of the situation:

> . . . as one study explained, "[t]he outcomes of empirical studies, reviews, and meta-analyses have converged on the conclusion that the confidence-accuracy relationship for eyewitness identification is weak, with average confidence-accuracy correlations generally estimated between little more than 0 and .29." . . . Another slightly older analysis. . . has suggested a confidence-accuracy correlation of

> only 0.41 for certain types of identifications. . . Importantly, error rates can be high even among the most confident witnesses. Researchers have performed studies that track, in addition to identification accuracy, the subjects' estimates of their confidence in their identifications. In one article reporting results from an empirical study, researchers found that among witnesses who made positive identifications, as many as 40 percent were mistaken, yet they declared themselves to be 90 percent to 100 percent confident in the accuracy of their identifications. . . . This confirms that many witnesses are overconfident in their identification decisions. (American Psychological Association, 2014, pp. 17–18)

Claims like this do not accurately inform the legal system. If these claims of an untrustworthy confidence-accuracy relation had been restricted to specific non-pristine testing conditions that have been shown to compromise the information value of eyewitness confidence or to confidence statements taken later rather than at the time of the initial identification, then they would be defensible claims.

One of the key points we will emphasize is that the only time that confidence is known to be a highly reliable predictor of accuracy is when memory is first tested, before there is much opportunity for memory contamination to occur. An expression of low confidence on that first test is a glaring red flag because it is almost always an indication that the risk of error is high. Instead of being ignored, an initial expression of low confidence should take center stage—overshadowing all other considerations—when a jury's goal is to evaluate the reliability of a suspect ID. If the witness is assumed to be honest, and if the ID was made with low confidence, then it is an unreliable ID. In fact, most of the DNA exonerees who were misidentified by an eyewitness were, at the outset of the investigation, identified with low confidence (Garrett, 2011). It was only later, in court and in front of the jury, that the initial low-confidence ID somehow morphed into a high-confidence ID. If it had been understood that confidence is indicative of accuracy only on an initial memory test (i.e., that on an initial test, low confidence implies low accuracy and high confidence implies higher accuracy), then many of these wrongfully convicted individuals may never have been found guilty in the first place. Or, if prosecutors had understood that low confidence at the initial identification is indicative of a high risk of error, then the innocent suspects in these cases might not have been indicted in the first place. Thus, far from being a problem, *initial* eyewitness confidence is part of the solution to eyewitness-based wrongful convictions (Box 1).

Box 1. Jennifer Thompson's Misidentification of Ronald Cotton

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* identification of Cotton from a photo lineup was characterized by a prolonged period of hesitation and indecision that lasted for nearly 5 minutes and ended with a low-confidence verbal identification consisting of the words "I think this is the guy" (Thompson-Cannino, Cotton, & Torneo, 2009, p. 33; Garrett, 2011). 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). Indeed, in a very real way, it was the legal system—not Jennifer Thompson—that made the key mistake by ignoring her initial (low) confidence. From this perspective, the time has come to exonerate her, too.

To appreciate how important it is to take into account (not ignore) an initial expression of low confidence by an eyewitness, imagine an eyewitness-identification case involving an innocent suspect that is adjudicated using an approach in which eyewitness confidence is ignored but various factors known to affect eyewitness memory are taken into consideration by a jury. Many of these factors are *estimator variables*—that is, variables that affect memory but are outside of the control of the legal system (Wells, 1978). Some common estimator variables include:

1. Race (cross-race IDs are less accurate than same-race IDs)
2. Exposure duration (brief exposure results in worse memory for the perpetrator than longer exposure)
3. Lighting (poor lighting during the crime results in worse memory for the perpetrator than good lighting)
4. Retention interval (a longer duration between the witnessed crime and the first lineup test results in worse memory for the perpetrator than a shorter duration)
5. Stress (high stress can lead to worse memory for the perpetrator than low stress)
6. Weapon focus (memory for the perpetrator is worse when a weapon is present than when no weapon is present)

For this hypothetical case involving an innocent suspect identified by an eyewitness, assume that all of these factors were favorable. For example, assume it was a same-race ID, exposure duration was long, the lighting was good, the retention interval was short, the witness was not particularly stressed, and no weapon was present. Under such conditions, the jury might reasonably conclude that the eyewitness-identification evidence is reliable and find the innocent suspect guilty. Now imagine that, unbeknownst to the jury, the witness expressed low confidence when the ID was initially made—because the innocent suspect was not a particularly good match to the witness's memory. The evidence we will review shows that such an ID is highly error prone despite the fact that all of the estimator variables are such that one might reasonably conclude otherwise. A low-confidence initial ID *trumps these good witnessing conditions* when evaluating the reliability of eyewitness-identification evidence. For that reason, ignoring initial confidence can place innocent suspects at risk.

Whereas low-confidence initial IDs always signal low accuracy—whether the identification procedure was pristine or not—high-confidence IDs on an initial test generally signal high accuracy when pristine testing conditions were used. Thus, initial confidence can serve the cause of justice by protecting the innocent (because initial IDs made with low confidence are untrustworthy) and imperiling the guilty (because initial IDs made with high confidence are trustworthy given appropriate testing conditions). That being the case, it is important to consider what has been learned about the proper way to conduct an eyewitness-identification test.

What Are the Pristine Eyewitness-Identification Procedures?

The error-prone nature of high-confidence eyewitness identifications made in a court of law—after memory has been contaminated—should no longer come as a surprise. All forensic tests—even DNA tests—have the potential to be unreliable if improper testing procedures are used. Proper procedures for obtaining reliable DNA test results were worked out by scientists in the laboratory before they were ever implemented in the forensic domain. As noted earlier, the same cannot be said of eyewitness-identification procedures. Since the 1970s, however, eyewitness-identification researchers have made considerable progress in working out more effective ways of testing eyewitness memory.

A general framework for improving eyewitness-identification procedures was described by Wells and Luus (1990), who proposed the "lineups-as-experiments" analogy. In this analogy, the officer conducting the lineup is like an experimenter; the eyewitnesses are the subjects; instructions to the eyewitnesses can be likened to an experimenter's protocol; the suspect is a stimulus; and the selection of lineup members and the positioning of the suspect in the lineup are part of the design. In addition, police have a hypothesis (e.g., that #4 is the guilty party) and have created a design and procedure to test the hypothesis. The eyewitnesses' choices or identification behaviors constitute the data from which the validity of that hypothesis will be evaluated by police and possibly a prosecutor, judge, and jury.

The idea behind the lineups-as-experiments analogy is that steps that have been taken to enhance the validity of scientific experiments can be applied to police lineup procedures to achieve the same goal. As one example, according to standard laboratory practice, the experimenter is blind to the experimental condition to avoid unconscious biases that might otherwise skew the results in favor of the experimenter's hypothesis. In a police lineup, the "experimenter's" hypothesis is that the suspect is the perpetrator; it therefore stands to reason that the officer administering the lineup should be blind to who the suspect is to avoid unintentionally steering the witness to the suspect. This practice is known as a *double-blind* lineup procedure because neither the lineup administrator nor the witness is told in advance who the suspect in the lineup is. Thus, if the suspect is identified by the witness, one can be more confident that the ID was based on the memory of the witness compared to when a non-blind test is administered.

Another important conceptual distinction to keep in mind is the difference between *system variables* and estimator variables (Wells, 1978). Most eyewitness-identification research has focused on system variables, which are factors affecting the reliability of eyewitness identifications that the criminal justice system can control. As noted earlier, estimator variables are factors that can affect the reliability of an identification but are outside the control of the criminal justice system (e.g., duration of exposure to the perpetrator, the retention interval

between the witnessed crime and the first memory test, the presence or absence of a weapon). The main system variable we focus on here concerns how a lineup is administered. Research on lineups has led to a number of recommendations for enhancing the reliability of eyewitness IDs and, critically, for creating the conditions under which confidence is a reliable indicator of accuracy. We review those system-variable recommendations below. Although estimator variables cannot be controlled by the time a crime comes to the attention of the police and thus do not bear on the issue of how to conduct a pristine identification procedure, eyewitness confidence may be an important consideration with respect to those variables as well. Later, following our review of the empirical evidence on the confidence-accuracy relationship, we briefly consider the issue of eyewitness confidence and estimator variables. Here, we consider five system-variable recommendations for the pristine conduct of an eyewitness-identification procedure that were put forward in a white paper of the American Psychology-Law Society and Division 41 of the American Psychological Association (Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998).

Include only one suspect per lineup

A lineup should contain only one suspect, with the remaining persons being known-innocent fillers. The typical recommendation is that a lineup should contain at least five known-innocent fillers (National Institute of Justice, 1999). In the parlance of the lineups-as-experiments analogy, the use of known-innocent fillers can be construed as a method of controlling for guessing. Using an all-suspect lineup, a witness who is prone to simply pick someone will always manage to land on a suspect, and charges might be brought against that person. The dangers of all-suspect lineups have long been documented in the eyewitness-identification literature (Wells & Turtle, 1986). In effect, a lineup that contains only suspects (no fillers) is like a multiple-choice test with no wrong answer. A proper lineup should be constructed in such a way that the witness can "fail" by selecting a filler.

Although fundamental and seemingly elementary, this safeguard against mistaken identification was once commonly violated and is still too often violated today. In fact, in the case of Ronald Cotton and Jennifer Thompson (see Box 1), the photo lineup shown to Thompson was an all-suspect lineup from which she tentatively identified Cotton. This was followed later by a live lineup in which Cotton was the only suspect and the remaining members were fillers. Of course, the actual perpetrator, Bobby Poole, was not in either lineup. The one-suspect recommendation applies under all circumstances. For instance, if there are multiple suspects even though there was only one offender, each suspect should appear in his or her own lineup along with fillers selected for that lineup. If there were multiple offenders, each suspect should still appear in his or her own lineup.

The suspect should not stand out in the lineup

Merely having fillers in a lineup is not in itself a guarantee that they will serve their function of helping to prevent mistaken identifications. Consider, for instance, a case in which the eyewitness described the offender as being a tall, thin male with dark hair and a moustache. Suppose now that the suspect fits this description but some fillers in the lineup are short, others do not have moustaches, and others have light hair. In this case, the suspect will stand out to the witness as being the person who looks most like the offender relative to the other lineup members, regardless of whether the suspect is the actual offender or not. This is the classic idea of a biased lineup. Research shows that placing an innocent suspect who fits the description of the offender in a lineup in which the fillers do not fit the description results in a high rate of mistaken identifications of that person, even when absolute similarity between the innocent person and the offender is only moderate. Moreover, there is evidence for what has been called the *dud effect*, in which adding fillers who look nothing like the perpetrator ("duds") to a lineup increases the confidence with which witnesses choose an innocent person who resembles the perpetrator (Charman, Wells, & Joy, 2011). One way to test whether the fillers are serving their purpose of helping to protect against mistaken identification is to ask whether a non-witness could pick the suspect out from the lineup by merely knowing the description that the eyewitness gave of the offender or by identifying who stands out in the lineup. If the answer is "yes," the fillers are not serving their purpose in the lineup. Indeed, this is the foundation of the "mock witness test" that was developed in the early days of eyewitness-identification research for analyzing the fairness of lineups (Wells, Leippe, & Ostrom, 1979).

Biased lineups are such a severe threat to our ability to rely on the confidence of the witness to infer accuracy that it is important that we give this issue a bit more treatment. One kind of situation that can place an innocent suspect at risk of being mistakenly identified with high confidence is coincidental resemblance between the innocent suspect and the actual perpetrator. Even if all the lineup fillers fit the witness's verbal description of the perpetrator, coincidental resemblance will make an innocent suspect stand out, and empirical studies have shown

that unusual resemblance of this type leads to mistaken identifications and high confidence (R. C. L. Lindsay, 1986). We cannot rule out the possibility of coincidental resemblance. But the fact that it is coincidental suggests that it is likely to be extremely rare. In fact, we have found no DNA exoneration case thus far that seems to qualify as having been an example of coincidental resemblance (if by coincidental resemblance we mean that the resemblance was due merely to chance).

On the other hand, unusual resemblance can occur (and has occurred) between an innocent suspect and the perpetrator for reasons other than coincidence. For example, police sometimes use sketch artists or software programs with which witnesses attempt to create a likeness of the perpetrator's face for the purpose of finding possible suspects. In general, if the witness makes a good composite and the composite is then used to find a suspect, the suspect is going to show a strong resemblance to the perpetrator even if the suspect is not the perpetrator. Hence, if the composite is used to find a suspect but the fillers are selected based on the broad verbal description given by the witness, the suspect will stand out (see Box 2).

Box 2. A Striking Resemblance: The Mistaken Identification of Michael McAlister

After spending 29 years in prison for a sexual assault that he did not commit, Michael McAlister was exonerated in 2015. The real perpetrator (on the left) was a serial rapist who bore a striking resemblance to McAlister and the only trial evidence linking McAlister to the attack was the victim's eyewitness identification and testimony. The McAlister case is an example of unusual similarity that, we argue, is not simply a coincidence. McAlister became the suspect in the case based on a facial composite sketch developed with the assistance of the victim witness. If an innocent person becomes a suspect based on their resemblance to a composite sketch (or a surveillance image), there is a heightened risk that the innocent person will have unusual resemblance to the eyewitness's memory of the actual perpetrator. In these cases, the lineup fillers need to be selected based on the fact that they also resemble the composite so as to make sure that the suspect does not stand out in the lineup.

Another way in which an innocent suspect might have unusual similarity to the perpetrator is when surveillance images (e.g., from a convenience store camera) are used to produce a suspect. With the increasing prevalence of electronic surveillance devices in public places, this path to becoming a suspect is likely to be increasingly common. Interestingly, people are quite poor at being able to accurately match a stranger to a surveillance image, even for high-quality images (e.g., see Davis & Valentine, 2009). But the process of using a surveillance image to decide who might be a suspect is rather certain to lead to an individual who is highly similar to the perpetrator, even if the person is innocent. Hence, if an innocent person becomes a suspect and is placed in a lineup based on his or her similarity to a surveillance image, then there is likely to be unusual similarity between the innocent suspect and the eyewitness's memory of the perpetrator, which could lead to a high-confidence (but mistaken) identification.

Notice that this surveillance-image path to high similarity is like the composite example; the high similarity did not occur purely by chance, and therefore it is not coincidental resemblance. And that point is key to solving the problem of unusual similarity when similarity arises from composites or from surveillance images. The solution here is contained in the strategy for selecting fillers for a lineup. Recall that the overall idea for selecting good fillers for a lineup is to make sure that the suspect does not stand out based on what is already known about the perpetrator. For example, if the witness described the perpetrator as being a White male, mid-20s in age, slim build, clean shaven, with short dark hair, and investigators find a suspect with those characteristics, then all of the lineup fillers also need to fit that description. If a composite or surveillance image of the perpetrator was used to find a suspect, however, the composite or surveillance image should trump the verbal description as the criterion for selecting fillers. In other words, if an individual became the suspect based on his or her similarity to a composite or a surveillance image, then the fillers need to also be selected based on their similarity to the composite or surveillance image. Yes, the suspect will still have unusual similarity to the perpetrator even if the suspect is innocent, but so will the fillers. As a result of this strategy for selecting lineup fillers, an innocent suspect should not stand out, thereby controlling the chances of a mistaken identification and false confidence.

Caution that the offender might not be in the lineup

Eyewitnesses often approach lineups with the goal of finding the offender. They should be cautioned that the offender might not be in the lineup because they need to understand that they are not "failing" if they do not choose

someone; after all, the correct answer might be "none of the above." In fact, "none of the above" was the correct answer not only in the case of Ronald Cotton and Jennifer Thompson, but also in all the other mistaken-identification cases that have been overturned by DNA testing. The instruction that the perpetrator might not be in the lineup is commonly called the *pre-lineup admonition*.

One concern about the pre-lineup admonition is that it might be undermined by suggestions that occur well before the lineup procedure commences. Quinlivan et al. (2012) found that suggestions to eyewitnesses leading them to believe that the perpetrator would be in the lineup prior to the commencement of the lineup instructions largely canceled the effect of the pre-lineup admonition. This, in turn, increased mistaken identifications in perpetrator-absent lineups and increased the confidence that witnesses had in those mistaken identifications. Consider, for example, an investigator contacting an eyewitness and saying, "We got the guy. We just need for you to come pick him out of a lineup." It seems quite likely that, as Quinlivan et al. found, this suggestion would largely cancel the pre-lineup admonition that would be given later when the formal lineup procedure begins.

Suggestions that occur prior to the commencement of a lineup procedure are concerning because they might be difficult to control. When jurisdictions have adopted pristine eyewitness-identification procedures, those procedures have typically covered only the official commencement of the pre-lineup instructions. Workable solutions to the potential problem of suggestions occurring prior to the initial identification have not been developed; we mention it here simply to raise awareness of it. Although the degree to which it is an actual problem is unknown, it seems reasonable to suppose that it could become more of a problem once the information value of initial eyewitness confidence becomes more widely appreciated. Thus, for the time being, we simply encourage vigilance against this possible contaminating factor.

Use double-blind testing

As noted above, the person who administers a lineup should not know which person in the lineup is the suspect. The use of such double-blind procedures is common in the social and medical sciences. Consider, for instance, the use of placebo control conditions in testing new drugs. Not only is the patient unaware of whether he or she received the drug or a placebo (single-blind), but so are any medical personnel who examine the patients (hence, the term *double-blind*). In this context, "blind" is figurative, not literal. Although the reason for keeping the patient blind as to whether he or she received the drug or a placebo is obvious, the need to keep the tester blind is less obvious.

The reason for keeping the tester blind is to prevent the tester from unintentionally influencing the outcome of the results. The double-blind testing recommendation for lineups does not assume that the tester intends to influence the eyewitness or is even aware of any such influence. This is not an integrity issue. Instead, it is merely an acknowledgment that people in law enforcement, like people in behavioral and medical research, are influenced by their own beliefs and may unknowingly "leak" this information, both verbally and nonverbally, in ways that can influence the person being tested. A vast scientific literature shows that the need for double-blind testing procedures is particularly crucial when there is close face-to-face interaction between the tester and the person being tested (e.g., see Rosenthal & Rubin, 1978).

It should be noted that using a lineup administrator who is blind to the suspect's identity is not the only way to prevent the lineup administrator from influencing the eyewitness. There are other methods, which have been called "blinded" procedures, that prevent the lineup administrator from knowing the position of the suspect in a photo lineup. The U.S. Department of Justice's (1999) guide on eyewitness evidence, for example, describes a folder or envelope "shuffle" method to prevent the officer from knowing which photo the witness is viewing. The shuffle method can be used for both simultaneous and sequential lineups, as it was in the blinded condition of a recent police department field study (Wixted et al., 2016). Alternatively, photo lineups can be administered using laptop computers that shuffle the order of the array, with the screen kept out of view of the lineup administrator.

Collect a confidence statement at the time of the identification

At the time an eyewitness makes an identification, a statement should be obtained from the eyewitness indicating how confident he or she is that the person identified is the offender. Of course, this assumes double-blind testing: The statement should be obtained by a lineup administrator who does not know which lineup member is the suspect. It is this initial confidence statement—and only this confidence statement—that is known to be a reliable indicator of accuracy. As we note in the next section, later statements of confidence by the eyewitness may not be reliable indicators of accuracy because confidence is malleable as a function of later events.

Additional Notes on Concerns About Contamination of Confidence

Before we discuss the issue of measuring the confidence-accuracy relation, we offer a deeper discussion of factors that can contaminate witness confidence and threaten its

relation to accuracy. This discussion can help produce a better understanding of the five recommendations for pristine procedures that were discussed above as they relate to witness confidence.

The confidence that people have in a memory is malleable. Studies show that simply imagining that some childhood event happened (when in fact it did not) can lead people to develop false confidence that they remember the fictitious event actually happening (Garry, Manning, Loftus, & Sherman, 1996). In the case of eyewitness identification, both the anticipation by eyewitnesses that they will later be cross-examined about their identification and the encouragement to prepare themselves for cross-examination have been shown to inflate witnesses' confidence (e.g., Wells, Ferguson, & Lindsay, 1981). Presumably, this confidence inflation occurs because witnesses rehearse the event in preparation for cross-examination, which makes the memory more vivid and fluently retrieved and thereby makes it seem more true, even if it is a false memory. Again, however, our thesis about the diagnosticity of confidence applies only to the initial confidence of the witness at the time of identification, not to later feelings of confidence that might be the product of post-identification contamination.

Perhaps the biggest threat to our ability to rely on confidence in eyewitness identification occurs when witnesses receive post-identification feedback that suggests they made an accurate identification (Wells & Bradfield, 1998). There is now a large body of eyewitness-identification studies showing that a simple comment to an eyewitness who has made a mistaken identification (e.g., "Good, you identified the suspect") can lead to immediate strong inflation of the witness's confidence. The effect of post-identification feedback is large. A recent meta-analysis of post-identification feedback studies showed that the eyewitnesses' confidence in their mistaken identifications was inflated by approximately a full standard deviation following such a comment (Steblay, Wells, & Douglass, 2014). The post-identification feedback effect is more muted for accurate eyewitness identifications, which means that confirmatory post-identification feedback actually harms the relation between accuracy and confidence (Charman & Wells, 2012).

There is a provocative and important twist to the post-identification feedback effect. Specifically, in post-identification feedback experiments, the question asked of witnesses is "How confident were you *at the time of the identification?*" Whereas few might be surprised that witnesses' post-identification confidence is inflated by confirmatory post-identification feedback, these studies measure the witnesses' retrospective confidence (not current confidence) by asking them to report how confident they recall having been at the time of the identification (before they received the feedback). So, post-identification feedback not only affects current confidence but also distorts eyewitnesses' recall for how confident they were at an earlier time. In fact, in a *New York Times* op-ed in 2000, Jennifer Thompson had this to say about her initial mistaken ID of Ronald Cotton: "Several days later, looking at a series of police photos, I identified my attacker. I knew this was the man. I was completely confident. I was sure" (Thompson, 2000). In truth, Jennifer Thompson was not completely confident at the time: Her initial ID was made with low confidence. However, feedback that she received at a later time led her to erroneously recall having been sure from the outset.

Interestingly, when witnesses were asked if post-identification feedback might have influenced how they answered the confidence question, most said "no," yet those who said "no" were no less influenced than were those who said "yes" (Wells & Bradfield, 1998). Moreover, post-identification feedback produces this same type of distortion not just for retrospective confidence but also for other testimony-relevant self-reports of eyewitnesses, such as reports of how much attention they paid at the time of witnessing and how good their view was of the perpetrator (see Steblay et al., 2014, for a meta-analysis of all these measures).

Experimental evidence indicates that lineup administrators' own expectations are likely to influence the confidence of the witness even when the lineup administrators are given an objective script to follow and are instructed to not deviate from that script. Garrioch and Brimacombe (2001) randomly assigned people to the role of witness or lineup administrator and then randomly assigned lineup administrators to a condition in which they were led to believe that the perpetrator was in a particular position of the lineup or a condition in which the lineup administrators were told nothing about the perpetrator position in the lineup. In reality, the perpetrator was never in the lineup. But when witnesses chose the lineup member who the lineup administrator had been led to believe was the perpetrator, the witnesses reported being much more confident than when the lineup administrator had no expectations about which person was the perpetrator. Videotapes of the lineup administrators' behaviors showed different patterns of post-identification nonverbal or paralinguistic behaviors as a function of lineup administrators' expectations about which lineup member was the perpetrator. These influences of the lineup administrators' expectations on the confidence of the witnesses occurred despite the fact that there were no incentives or other motivations on the part of the lineup administrators. Furthermore, 100% of the lineup administrators indicated that they believed they did not provide any post-identification feedback, and 95% of the witnesses believed they did not receive any post-identification feedback.

So, post-identification feedback appears to be a pernicious problem. Fortunately, we have long known the solution for preventing the contamination of post-identification feedback, namely the double-blind lineup procedure that eyewitness researchers have been proposing for over 25 years (Wells & Luus, 1990). In fact, one of the primary reasons for double-blind lineup testing is to prevent the lineup administrator from giving inadvertent feedback that could distort the confidence of the witness. The simple beauty of the double-blind lineup procedure is that the lineup administrator does not know if the witness picked a known-innocent filler or picked the suspect in the case. That same double-blind lineup administrator can then secure a confidence statement from the witness prior to any opportunity for the witness to be given feedback about whether the identified person was the suspect or was a lineup filler (Wells & Bradfield, 1999).

In addition to conducting the lineup using a double-blind procedure, eyewitness-identification researchers have long advocated videotaping the entire eyewitness-identification procedure (e.g., Kassin, 1998). And the idea of videotaping all identification procedures was recently endorsed by a committee of the National Academy of Sciences (National Research Council, 2014). The initial confidence statement is then a matter of record, and it is that initial confidence statement, not later confidence statements, that prosecutors and courts should rely upon. If the case reaches trial, juries should use only this initial confidence statement for assessing the reliability of the identification.

Of course, having a pristine assessment of witness confidence at the time of the identification does not prevent witnesses from undergoing confidence inflation later and perhaps being quite positive at trial. But that is why we emphasize so strongly that the reliability of confidence statements must be based on the eyewitness's initial confidence, not later claims of confidence. And this is where courts have commonly made a serious mistake. Courts routinely permit witnesses to state their confidence at pretrial hearings or at trial, well after they might have undergone serious confidence inflation from repeated identifications, coaching, confirmatory feedback, and so on. The confidence of the witness at the time of a preliminary hearing or at trial is not a pristine assessment of confidence.

Interestingly, the U.S. Supreme Court's guiding ruling on eyewitness identification, which is now nearly 40 years old, urged lower courts to consider the confidence that the eyewitness had *at the time of the identification* in evaluating the reliability of an eyewitness identification (*Manson v. Braithwaite*, 1977). What can be done if the lineup administrator failed to secure a confidence statement from the witness at the time of the identification? Some courts might be tempted to simply ask the witness to cast his or her mind back to the lineup and recall how confident he or she was at the time of the identification. But, as the literature on the post-identification feedback effect shows, witnesses do not accurately recall their initial uncertainty if confidence inflation has occurred as a result of contaminating influences, and instead recall having been confident all along. There is no substitute for taking a confidence statement at the time of the identification.

It is also important to keep in mind that our claims about the reliability of confidence as an indicator of accuracy in eyewitness identification apply only to cases in which the eyewitness-identification test procedures were pristine (Box 3). Unfortunately, at this point, not all jurisdictions in the United States collect a confidence statement at the time of the identification, and when they do, they often do not use a double-blind procedure. Indeed, as recently as 2001, there was no jurisdiction in the United States that used double-blind lineup procedures. Fortunately, efforts by eyewitness-identification researchers, in partnership with the Innocence Project, local and state-level reform commissions, and other policymakers, have managed to facilitate reforms on eyewitness-identification procedures in a growing number of jurisdictions in the United States. As of the time of this writing, for example, state laws have been passed by legislators that require double-blind lineup administration in Connecticut, Colorado, Kansas, Illinois, Maryland, North Carolina, Ohio, and Vermont. Additional states have used other mechanisms to force the use of double-blind lineup administration. New Jersey, for example, requires double-blind lineup administration via a plenary mandate from the attorney general of New Jersey. Oregon's Supreme Court issued a decision (*State v. Lawson*, 2012) that largely makes double-blind lineup procedures necessary in Oregon. In addition, the states of Texas, Rhode Island, Wisconsin, and West Virginia have achieved substantial compliance for using double-blind lineup procedures through a combination of laws and influential task-force recommendations. In addition, individual jurisdictions such as Suffolk County, Massachusetts (Boston and surrounding areas), Santa Clara County, California (including San Jose and Palo Alto), Minneapolis, Minnesota, and many other large and small jurisdictions have made eyewitness-identification reforms that include the requirement of double-blind lineup administration. At the time of this writing, numerous other states are considering requiring double-blind lineup administration.

Box 3. Pristine Lineup Conditions

1. Include only one suspect per lineup
2. The suspect should not stand out in the lineup
3. Caution that the offender might not be in the lineup
4. Use double-blind testing
5. Collect a confidence statement at the time of the identification

Because there remain many jurisdictions that have not yet adopted pristine eyewitness-identification testing procedures, it is important that we emphasize a caveat to our primary thesis. Specifically, our claim regarding the high diagnosticity of eyewitness identifications made with high confidence does not extend without qualification to those jurisdictions that have not yet made reforms to ensure pristine procedures. For example, as we will show later, a high-confidence ID made from an unfair lineup is considerably more error prone than a high-confidence ID made from a pristine lineup. A similar risk of error may occur when eyewitness-identification procedures depart from the other recommended procedures as well, though detailed investigations into their effect on high-confidence accuracy have not been performed. Nevertheless, it seems safe to say that prosecutors and defense attorneys are likely to debate the reliability of a suspect ID whenever the procedures summarized in Box 3 have not been followed, and for good reason. For example, if a lineup was administered in a non-blind fashion, the question will inevitably arise as to whether the lineup administrator unintentionally influenced the identification made by the witness and the confidence of the witness, which research shows is a real possibility (e.g., Garrioch & Brimacombe, 2001). As noted in the National Research Council (2014) report, "The use of double-blind procedures will eliminate a line of cross-examination of officers in court" (p. 107). The same argument can be made for each of the practices listed in Box 3.

Whereas the information value of a high-confidence ID may be called into question whenever non-pristine testing conditions are used, the information value of a low-confidence ID is never open to question. No matter how good or how bad the eyewitness-identification procedure is, a low-confidence ID implies that the ID is error prone. As noted above, if an identification was made in a jurisdiction that has not adopted pristine testing conditions, the defense and the prosecution may end up debating in court about whether or not the testing procedure was good enough. However, that debate is rendered moot if it is known that the eyewitness made an initial good-faith ID with low confidence. Such an ID is error prone, even under pristine testing conditions.

Returning to the main point, if the pristine conditions listed here (summarized in Box 3) are followed, then a low-confidence ID implies low accuracy, and a high-confidence ID implies high accuracy. Although eyewitness-identification research conducted over the last 20 years has shown this to be true, our understanding of this issue emerged rather gradually, which may help to explain why it is not more widely understood within the legal system. We turn now to a consideration of the eyewitness confidence-accuracy literature, beginning with a review of the methods used to measure the confidence-accuracy relationship (a key part of the story). We emphasize that except where noted, the studies we consider were carried out using the pristine testing conditions summarized above. How reliable is an ID made under those conditions, according to what we have learned over the last 20 years? To answer that question, we first have to consider which approach to measuring the confidence-accuracy relation most accurately conveys the information sought by judges and juries.

Measuring the Eyewitness Confidence-Accuracy Relationship

The data used to investigate the confidence-accuracy relationship for eyewitness identification come mostly from forensically relevant lab studies in which participants become witnesses to a mock crime (e.g., by watching a live enactment or a video of someone committing a crime, such as snatching a purse, planting a bomb, or robbing someone at an ATM) and are later shown a lineup in which the perpetrator (the target) is either present or absent. A target-present lineup includes the perpetrator along with (usually five or seven) similar fillers; a target-absent lineup is the same except that the perpetrator is replaced by another similar filler, as illustrated in Figure 1. In some studies, the individual depicted in the replacement photo serves the role of the designated "innocent suspect." In other studies, no one in the target-absent lineup is designated to serve the role of an innocent suspect, so the risk to an innocent suspect is calculated by dividing the number of identifications in the target-absent lineup by the number of fillers (thereby assuming a perfectly fair lineup). When presented with a target-present or target-absent lineup, a witness in a mock-crime study first makes a decision—which consists of identifying the suspect, identifying a filler, or rejecting the lineup (i.e., saying that the perpetrator is not there)—and then supplies a confidence rating associated with that decision. A correct response consists of (a) a suspect

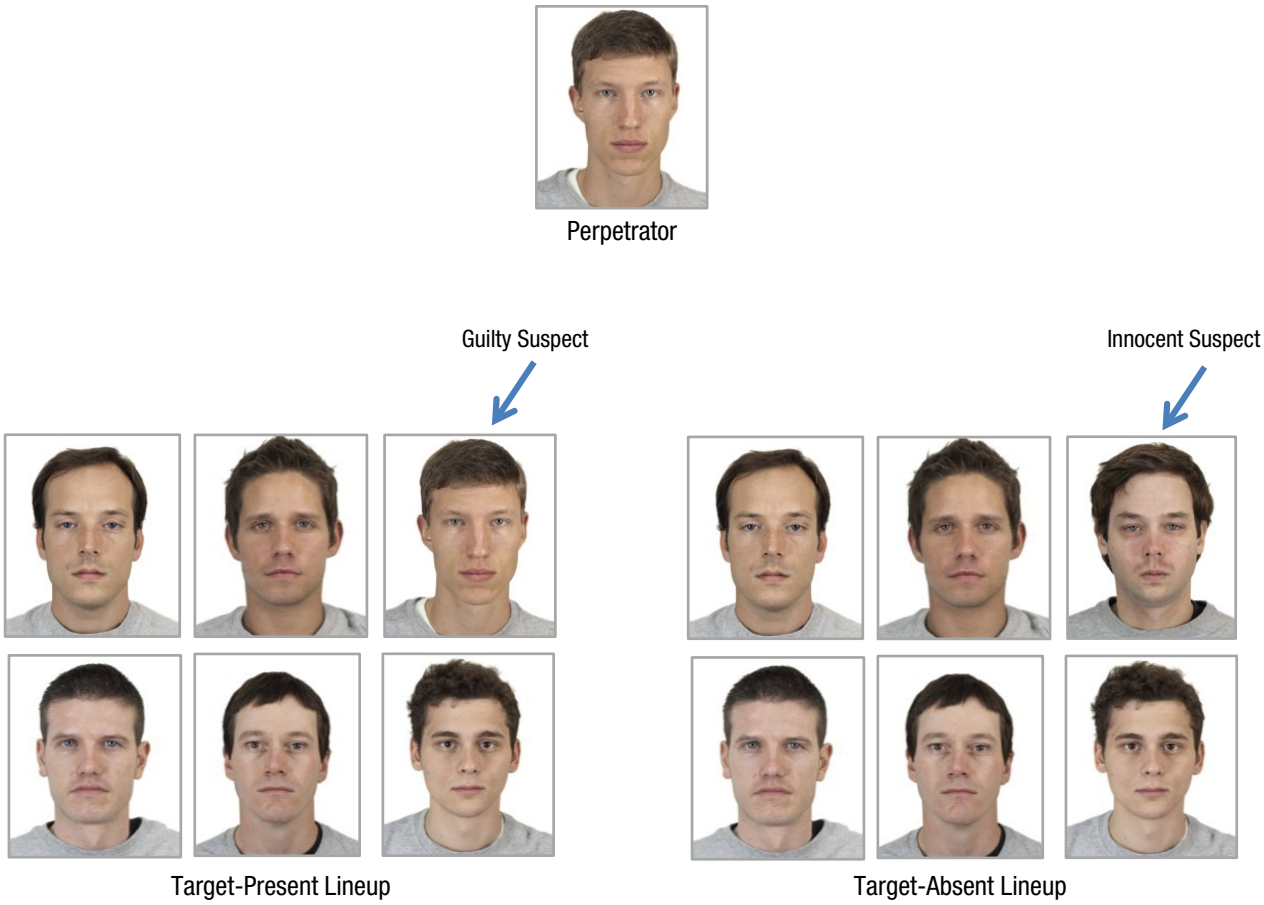

Fig. 1. Example lineups used in mock-crime studies. (Images drawn from the Chicago Face Database; Ma, Correll, & Wittenbrink, 2015.)

ID from a target-present lineup or (b) the rejection of a target-absent lineup, whereas an incorrect response consists of (a) a suspect ID from a target-absent lineup (if there is a designated innocent suspect), (b) a filler ID from either type of lineup, or (c) the rejection of a target-present lineup.

Our appreciation of the information value of confidence has grown considerably in recent years, partly as a result of methodological changes in the way that researchers measure the confidence-accuracy relationship. Prior research on the issue can be divided into three phases according to the measure that was used. In Phase 1, the point-biserial correlation coefficient was the preferred measure. In Phase 2, calibration curves were more commonly used. In Phase 3, confidence-accuracy characteristic (CAC) analysis (Mickes, 2015) and closely related but more complete Bayesian analyses (Wells, Yang, & Smalarz, 2015) have been used. An argument we will advance is that only the measures used in Phase 3 directly address questions of interest to the legal system.

The point-biserial correlation coefficient

In Phase 1, the relationship between confidence and accuracy was measured by computing the standard Pearson r correlation coefficient between the accuracy of a response (e.g., coded as 0 or 1) and the corresponding confidence rating (e.g., measured using a 5-point scale from *just guessing* to *very sure that is the person*). Because accuracy is coded as a dichotomous variable, the Pearson r in this case is known as a point-biserial correlation coefficient. Using this measure, much of the initial research examining eyewitness certainty suggested that certainty was largely unrelated to identification accuracy (e.g., Clifford & Scott, 1978; Deffenbacher, Brown, & Sturgill, 1978; Leippe, Wells, & Ostrom, 1978), with correlation coefficients generally falling into the .00 to .20 range.

In these studies, all of the data were bundled together for the analysis, whether the eyewitness made a suspect ID, a filler ID, or a non-ID. In a later meta-analysis,

Sporer, Penrod, Read, and Cutler (1995) found that the relationship was noticeably stronger—about .41—when the analysis was limited to only those who made an ID from a lineup (i.e., when the analysis was limited to "choosers" who identified a suspect or a filler). Limiting the analysis to choosers is reasonable because only witnesses who choose someone end up testifying in court against the person they identified. This move—separating choosers (those who make suspect IDs and filler IDs) from non-choosers (those who make a non-ID)—foreshadowed a later move that we will argue is critical, namely separating choosers who make suspect IDs from choosers who make filler IDs. All three decision outcomes (suspect IDs, filler IDs, and non-IDs) need to be assessed for the independent information they provide about whether or not the lineup contains a guilty suspect. Nevertheless, Sporer et al.'s separation of choosers from non-choosers led to an important advance in our understanding of the confidence-accuracy relationship.

The novel message from Sporer et al. (1995) was that confidence is a more reliable indicator of accuracy for choosers than had been previously assumed. At .41, the correlation for choosers was clearly too large to argue that eyewitness confidence should be disregarded. Nevertheless, over the years, the interpretation of the Sporer et al. (1995) meta-analysis has generally drifted in the negative direction, as if the message were actually the opposite. For example, Reinitz, Seguin, Peria, and Loftus (2012) said, "It is well known that confidence is often a relatively poor predictor of accuracy (e.g., Bothwell, Deffenbacher, & Brigham, 1987; Sporer et al., 1995)" (p. 1089). Buratti and Allwood (2012) noted that "although many witnesses may feel confident about their identification, the relation between identification confidence and the correctness of the identification is weak (Brewer & Wells, 2011; Sporer et al., 1995)" (p. 590). Neal, Christiansen, Bornstein, and Robicheaux (2012) pointed out that "contrary to jurors' beliefs, eyewitness confidence is not a strong indicator of accuracy (Penrod & Cutler, 1995; Sporer et al., 1995)" (p. 50). And Wilson, Hugenberg, and Bernstein (2013) recently maintained that "one surprising lesson that psychologists have learned about memory is that the confidence of an eyewitness is only weakly related to their recognition accuracy (p. 98; see Sporer et al., 1995, for a review)."

It seems fair to say that these characterizations do not accurately convey what the Sporer et al. (1995) meta-analysis actually found. What Sporer et al. actually found was that, for choosers, the confidence-accuracy relationship is surprisingly strong. They also emphasized the fact that later events can inflate an eyewitness's confidence, obviously without increasing the accuracy of the initial ID. Some of the post-ID factors that can inflate confidence include hearing that other witnesses have identified the same suspect (Luus & Wells, 1994), being exposed to the identified suspect again (Brown, Deffenbacher, & Sturgill, 1977), and receiving encouraging feedback from police about the accuracy of the ID (Wells & Bradfield, 1998). However, for an *initial* ID made from a pristine lineup, the Sporer et al. (1995) meta-analysis showed that initial confidence is a reasonably good indicator of accuracy. In fact, 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).

Shortly after Sporer et al.'s (1995) meta-analysis was published, the argument was made that even their upgraded assessment of the confidence-accuracy relationship was, if anything, an understatement. Juslin et al. (1996) showed that the magnitude of the point-biserial correlation can be 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. In Appendix A, we provide a concrete example illustrating how this could be. The key point is that the .41 correlation coefficient for choosers is potentially compatible with a *very* strong confidence-accuracy relationship.

These considerations suggest that the point-biserial correlation coefficient is not the best statistic to use when trying to inform the legal system about the utility of eyewitness confidence. Note that this is not a criticism of the statistic itself. The point-biserial correlation coefficient is a perfectly valid effect-size statistic when used in conjunction with certain statistical tests, such as a *t* test (Rosnow et al., 2000). For example, in eyewitness-identification studies, one might ask whether the average level of confidence is higher for correct IDs than for incorrect IDs. This would be the appropriate way to analyze the data if you knew, for each eyewitness, whether his or her ID was correct or incorrect and you wanted to estimate his or her likely level of confidence. In fact, this is how the data were plotted in Figure 1 of Sporer et al.'s (1995) seminal article, and the corresponding point-biserial correlation coefficient of .41 indicates a moderate-to-large

average effect size. Yet this is not the question of interest to the legal system, because in actual practice, the situation is reversed: An eyewitness provides a confidence rating associated with an ID (this is the predictor variable, which is not averaged), and the legal system wants to make the best estimate as to the likely accuracy of that ID (this is the dependent variable, and it equals the average level of accuracy associated with each level of confidence that an eyewitness might express).

This logic suggests, as Juslin et al. (1996) pointed out, that plotting average accuracy (on the *y*-axis, as the dependent measure) versus different levels of confidence (on the *x*-axis, as the independent measure) is a more informative way to analyze the data. When plotted this way, the data come closer to providing an answer to the question asked by judges and juries trying to evaluate the reliability of an eyewitness. Their question is: Given that an eyewitness has a particular level of confidence in his or her ID of a suspect, how accurate is that ID likely to be? With regard to that question, a calibration curve provides much more relevant information than a correlation coefficient. Once this fact was understood, Phase 2 was ushered in as eyewitness-identification researchers began to measure the confidence-accuracy relationship by plotting calibration curves.

Calibration analysis

Following Sporer et al. (1995), calibration analyses are also typically performed separately on choosers (those who make a suspect ID or a filler ID) and non-choosers (those who make a non-ID decision). A calibration analysis can be performed whenever a confidence rating scale ranging from 0 to 100 is used. It is important to be clear about the exact computational formula used to compute calibration, so we consider the formula below. In the notation we use here, *nFID* stands for "number of filler IDs" and *nSID* stands for "number of suspect IDs." We also attach subscripts to these symbols, such as *TP*, which denotes target-present lineups, and *TA*, which denotes target-absent lineups. Thus, $nSID_{TP}$ means "number of suspect IDs from target-present lineups." Finally, we add the subscript *c*, which represents the confidence expressed by the witness. Thus, $nSID_{TP\text{-}c}$ means "number of suspect IDs from target-present lineups with confidence *c*," where *c* might be 90% to 100% confident.

Basically, in a calibration analysis of choosers, the percentage-correct accuracy score for a given level of confidence, *c*, is equal to 100 multiplied by the number of (correct) suspect IDs from target-present lineups made with confidence level *c* ($nSID_{TP\text{-}c}$) and divided by the total number of IDs (to suspects and fillers alike) made with confidence level *c*. Many calibration studies have used a target-absent lineup that does not have a designated innocent suspect, so the number of incorrect IDs consists of the number of filler IDs made from target-present lineups with confidence level *c* ($nFID_{TP\text{-}c}$) plus the number of filler IDs made from target-absent lineups with confidence level *c* ($nFID_{TA\text{-}c}$). Thus, for confidence level *c*, calibration equals $100 \times (nSID_{TP\text{-}c}) / (nSID_{TP\text{-}c} + nFID_{TP\text{-}c} + nFID_{TA\text{-}c})$. In practice, $nFID_{TP\text{-}c}$ is often excluded from the denominator, but the results tend to be similar either way.

Calibration studies typically find a strong relationship between confidence and accuracy when (a) the analysis is limited to choosers, (b) the witnesses are adults, (c) the lineups are fair, and (d) the confidence ratings are taken immediately after the ID is made (e.g., Brewer & Palmer, 2010). That is, they find a strong relationship between confidence and accuracy using pristine eyewitness-identification procedures that were also used in previous studies measuring the relationship using the point-biserial correlation coefficient. As an example, Figure 2a presents a calibration curve taken from Brewer and Wells (2006). As we will see, the results shown in Figure 2a are fairly typical of calibration studies, and they show that low-confidence IDs (*c* = 0%–20%) are associated with low accuracy (26.6% correct), whereas high-confidence IDs (*c* = 90%–100%) are associated with much higher accuracy (84.9% correct). It seems difficult to characterize the results shown in Figure 2a as indicating anything other than a very strong confidence-accuracy relationship for choosers. This is true even though the overall point-biserial correlation coefficient in this study was low (.32 for identifications of the thief in the video and .36 for identifications of the waiter in the video). These findings underscore the fact that the confidence-accuracy correlation can be low even when the confidence-accuracy relationship is strong. Note that the story for non-choosers is different. For them, the confidence-accuracy relationship is noticeably weaker (Fig. 2b), which is a conclusion that also corresponds to work using the point-biserial correlation coefficient (Sporer et al., 1995).

These results, like the point-biserial results discussed above, correspond to the 50% base rate of target-present lineups used in that study. As we discuss in more detail later, real police lineups may contain a guilty suspect less than 50% of the time. In such cases, the accuracy rates for choosers would be correspondingly lower than the values shown in Figure 2a, and the accuracy rates for non-choosers would be correspondingly higher than the values shown in Figure 2b. Nevertheless, the basic story would not change: For choosers, the confidence-accuracy relationship is strong, and for non-choosers it is considerably weaker.

Although the results in Figure 2a reflect a strong confidence-accuracy relationship, it also seems fair to say—and it often is said—that witnesses who express high

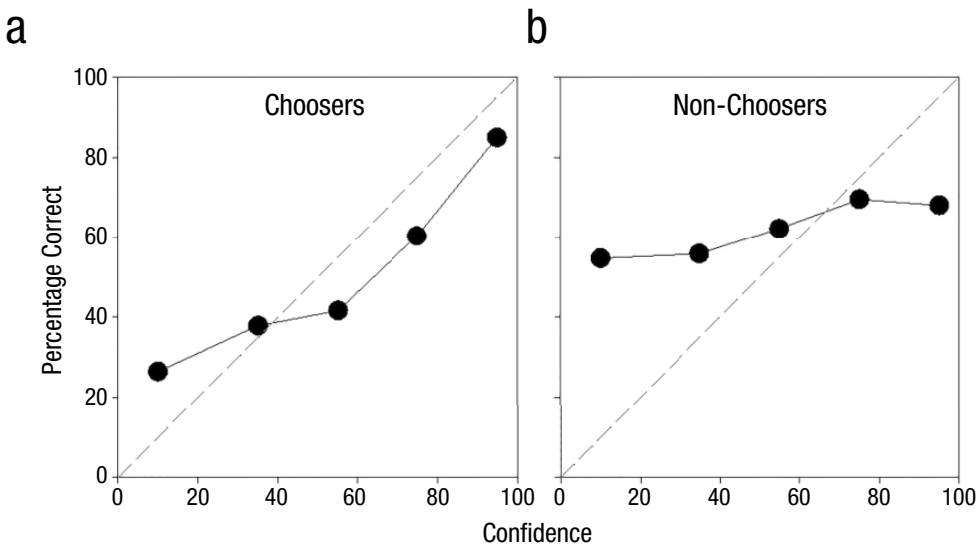

Fig. 2. Calibration data from Brewer and Wells (2006) for choosers (a) and non-choosers (b). The data were pooled across witnesses' identifications of either of the two targets who appeared in a mock-crime video (a thief and a waiter). The dashed line represents perfect calibration.

confidence, such as 90% to 100% confidence, are overconfident because their corresponding accuracy is typically lower than 90% (e.g., Lampinen, Neuschatz, & Cling, 2012; Leach, Cutler, & Van Wallendael, 2009; Valentine & Davis, 2015). Leach et al. (2009) put it this way:

> Given the modest correlation between confidence and identification accuracy, the tendency for witnesses to be overconfident in their decisions (Brewer & Wells, 2006), and the factors that further suppress the confidence-accuracy relation, confidence is of questionable utility in the assessment of eyewitness identification accuracy. (p. 161)

However, this pessimistic assessment seems premature, because the data have still not yet been analyzed in a way that most directly addresses the question of interest to judges and juries tasked with assessing the reliability of an initial eyewitness ID made with a particular level of confidence. In the courtroom, the question of interest is as follows: What does confidence tell us about the reliability of an eyewitness who has identified a *suspect*? The answer to this question is provided by limiting the analysis not just to choosers but to choosers who identify a suspect—just as the legal system limits its consideration to choosers who identify a suspect by referring identified suspects (but not identified fillers) for prosecution.

We refer to the dependent variable in an analysis that excludes filler IDs as *suspect-ID accuracy*, and we refer to a plot of suspect-ID accuracy versus confidence as a CAC analysis to distinguish it from the closely related calibration plot (Mickes, 2015). Unlike a calibration curve, a CAC plot provides the information that judges and juries want to know when they are trying to assess the reliability of an eyewitness who identified a suspect from a lineup.

Once again, our argument should not be construed as a criticism of the calibration statistic. A calibration curve is a perfectly appropriate way to represent the relevant data when the question concerns the confidence-accuracy relationship from the witness's perspective. In a calibration study, witnesses are instructed to choose a confidence rating of 80% (for example) when they believe they would be correct 80% of the time. From the witness's perspective, a correct ID consists of choosing a suspect from a target-present lineup, whereas an error consists of choosing a suspect from a target-absent lineup or choosing a filler from either type of lineup. Thus, a witness presumably interprets the instruction to mean that a confidence rating of 80% should consist of correct responses (suspect IDs from target-present lineups) 80% of the time and errors (suspect IDs from target-absent lineups and filler IDs) 20% of the time. A calibration curve appropriately shows the relationship between what a witness believes about his or her performance and what that performance is actually like.

However, the legal system is concerned with a different issue, because if the eyewitness picked a filler, we already know that the witness did not pick the perpetrator. So, the forensically relevant question is this: Given that the eyewitness picked the suspect with a particular level of confidence, how likely is it that the suspect is guilty? The answer to that question is provided by a CAC plot in

which the dependent measure is suspect-ID accuracy. We next describe how to compute suspect-ID accuracy, and then we reanalyze and plot the published data in terms of CAC analysis. We then use representative calibration data and reanalyze those results using the more detailed Bayesian analysis described by Wells et al. (2015). This analysis shows suspect-ID accuracy across the full range of base rates of target-present lineups (instead of limiting the analysis to the 50% base rate typically used in studies, as CAC analysis does). Using the same basic approach, we also consider a topic that is only rarely considered: What is the information value of a filler ID or a non-ID? These decision outcomes also bear on the likelihood that the suspect in the lineup is guilty, but the information they provide points in the opposite direction than that provided by a suspect ID, in that they are both probative of innocence (Wells et al., 2015). That fact is another reason suspect IDs and filler IDs should not be bundled together when the goal is to inform the legal system. They should not be bundled together because they provide independent (and opposing) information about the likelihood that the suspect in the lineup is guilty.

Confidence-accuracy characteristic analysis

Suspect-ID accuracy is based on the number of suspect IDs from target-present lineups (guilty-suspect IDs) made with confidence level *c*, $nSID_{TP\text{-}c}$, and the number of suspect IDs from target-absent lineups (innocent-suspect IDs) made with the same confidence level, $nSID_{TA\text{-}c}$. More specifically, suspect-ID accuracy is equal to 100% × $nSID_{TP\text{-}c}$ / ($nSID_{TP\text{-}c}$ + $nSID_{TA\text{-}c}$). Unlike in a real police lineup involving an innocent suspect, in a lab study there is no obvious person to use as the innocent suspect. In other words, there is obviously no one in the target-absent lineup who is suspected of having committed the crime depicted in the mock-crime video (because the experimenter selected the perpetrator in the video and so already knows who he is). How, then, does one compute the number of innocent-suspect IDs? Using one reasonable approach, the innocent suspect in a target-absent lineup is simply a designated filler, usually the filler that was used to replace the perpetrator's photo (as in Fig. 1). This approach is arguably the most logical approach because only the suspect differs across target-present and target-absent lineups (the other fillers are held constant). It also has the advantage of making it easy to count not only the number of guilty-suspect IDs that are made from target-present lineups but also the number of innocent-suspect IDs that are made from target-absent lineups. In that case, computing suspect-ID accuracy for each level of confidence is entirely straightforward.

If the replacement photo is not designated as the innocent suspect in target-absent lineups, then $nSID_{TA}$ can instead be estimated from the number of filler IDs from target-absent lineups divided by lineup size (*n*). In that case, suspect-ID accuracy would be given by 100% × $nSID_{TP\text{-}c}$ / ($nSID_{TP\text{-}c}$ + ~$nSID_{TA\text{-}c}$), where ~$nSID_{TA\text{-}c}$ = ($nFID_{TA\text{-}c}$ / *n*). For most of the studies we will review, the number of IDs of the replacement photo was not reported, so this approach to estimating the false-ID rate was the only option available. In the long run, these two approaches to computing the false-ID rate (namely, designating an innocent suspect vs. counting all target-absent filler IDs and dividing by lineup size) will yield the same average results so long as all of the fillers—including the one that replaces the perpetrator—are selected using the same decision rule, such as the rule that fillers must match the description of the perpetrator. For any particular study, however, the two approaches can yield different results. For example, by chance, half the time, the replacement photo (i.e., the natural choice to serve as the innocent suspect) will be a more attractive option than the average filler in the lineup, and 1/*n* of the time it will, by chance, be the most attractive option. In these studies, dividing the target-absent filler-ID rate by *n* will underestimate the false-ID rate, making suspect-ID accuracy at each level of confidence look better than it would look if the replacement photo had been designated as the innocent suspect. On the other hand, the other half of the time, the replacement photo will be a less attractive option than the average filler in the lineup, and 1/*n* of the time it will be the least attractive option. In these studies, dividing the target-absent filler-ID rate by *n* will overestimate the false-ID rate, making accuracy at each level of confidence look worse than it would if the replacement photo had been designated as the innocent suspect. Thus, our conclusions about the confidence-accuracy relationship will be based on what studies suggest in the aggregate, not on what any particular study suggests. Nevertheless, in light of these considerations, we believe that researchers should report the frequency with which each target-absent lineup member was identified. In any given study, it might be the case that, by chance, the replacement filler was chosen more often than the other fillers. If so, a conclusion derived from that study alone would apply more to unfair lineups than to fair lineups. When target-absent filler-ID rates for each lineup member are not reported, there is no way to tell if this is a problem or not.

A third approach to designating an innocent suspect in a target-absent lineup is problematic if the goal is to measure the confidence-accuracy relationship under pristine testing conditions. Using this third approach, the innocent suspect is defined to be the filler in the

target-absent lineup who most resembles the perpetrator. The innocent suspect stands out in that sense and will therefore be chosen more often than the other fillers. For example, in a recent study, Sučić, Tokić, and Ivešić (2015) first selected a set of six fillers who matched the description of the target in the target-present lineup and then selected the one who would serve as the designated innocent suspect in the following way: "The six photographs that were selected were top ranked for photograph similarity and the match to modal description, and the highest ranked photograph was used as the designated innocent suspect (suspect replacement) in a [target-absent] lineup" (p. 802). In other words, the designated innocent suspect in this study was chosen in such a way as to ensure that it would stand out in the target-absent lineup (i.e., by design, this was an unfair lineup). As noted earlier, in an ideal lineup, the suspect does *not* stand out, and if the police made lineups following this approach, those lineups would be unfair. We will separately review studies that used this approach, and we will see that it has a profoundly negative effect on the confidence-accuracy relationship even for an otherwise pristine identification procedure. However, the bulk of our review will consist of a reanalysis of studies that used lineups in which the replacement photo in the target-absent lineup was not selected using a different decision rule than the other fillers. For these studies, the number of innocent-suspect IDs was estimated by counting all filler IDs from target-absent lineups and dividing by lineup size. Again, for any single study, this approach to estimating the false-suspect-ID rate could mask the fact that the target-absent lineup was, by chance, biased toward or away from the replacement photo (i.e., toward or away from the photo that would most logically serve as the innocent suspect). Our conclusions about the relationship between confidence and accuracy are not based on any single study but are instead based on the aggregate results from many studies.

It is important to emphasize that the suspect-ID accuracy measure in CAC analysis is not another measure of calibration. As described in more detail below, random chance accuracy for suspect IDs is typically 50% correct, not 0% correct. Thus, if a 0-to-100 confidence scale is used, one would not expect to see suspect-ID accuracy match the level of confidence at the low end of the scale. In a CAC analysis, the question is not how well confidence and accuracy match; instead, the question is simply this: How does suspect-ID accuracy vary as confidence ranges from low to high? Because that is the question, CAC analysis can be carried out using any monotonic confidence rating scale (unlike a calibration analysis, which requires a 0-to-100 scale). In point of fact, very few police departments use 0-to-100 scales to assess initial confidence, so calibration is not often at issue in the legal system (although it is of interest in laboratory studies).

Suspect-ID accuracy is what judges and juries want to know when trying to evaluate the reliability of an eyewitness identification: Given that the suspect in this trial has been identified by an eyewitness with a particular level of confidence, what is the probability that the ID is correct? This is a question about the subset of eyewitnesses who identify a suspect. No other way of plotting the data (and no numerical summary of the data—not the point-biserial correlation coefficient nor any other statistic) provides the answer to that question more directly and more understandably than a visual plot relating suspect-ID accuracy to confidence. At a glance, it not only reveals how much suspect-ID accuracy changes as a function of confidence, it also shows how reliable eyewitness IDs are for each level of confidence. Therefore, we use this approach in our review of actual experiments, to which we now turn.

A Review and Reanalysis of Eyewitness Confidence and Accuracy Data

We begin with a reanalysis of some of the studies that were included in the Sporer et al. (1995) review, which used the correlation coefficient to quantify the confidence-accuracy relationship, and then we reanalyze subsequent data that were originally published as calibration curves or as receiver operating characteristic (ROC) curves.

A reanalysis of three of the original Sporer et al. (1995) meta-analysis experiments

What would the data that were reviewed by Sporer et al. (1995) suggest about the confidence-accuracy relationship if, instead of computing the point-biserial correlation for choosers, one simply plotted the data as a CAC plot? To find out, we contacted each of the authors of that article and requested the original data from the 30 experiments analyzed in their Table 1 so that the data could be replotted as CAC curves. Quite understandably, most of the data are no longer available. However, data from three of those 30 experiments, which are representative of the larger data set in terms of the point-biserial correlation, are still available and were kindly provided to us by J. Don Read. The three experiments are Experiments 1 and 2 from Read, Yuille, and Tollestrup (1992) and Experiment 3 from Read, Tollestrup, Hammersley, McFadzen, and Christensen (1990). The data from Experiment 2 of Read et al. (1992) were originally analyzed separately for two targets (a central suspect and a peripheral suspect)

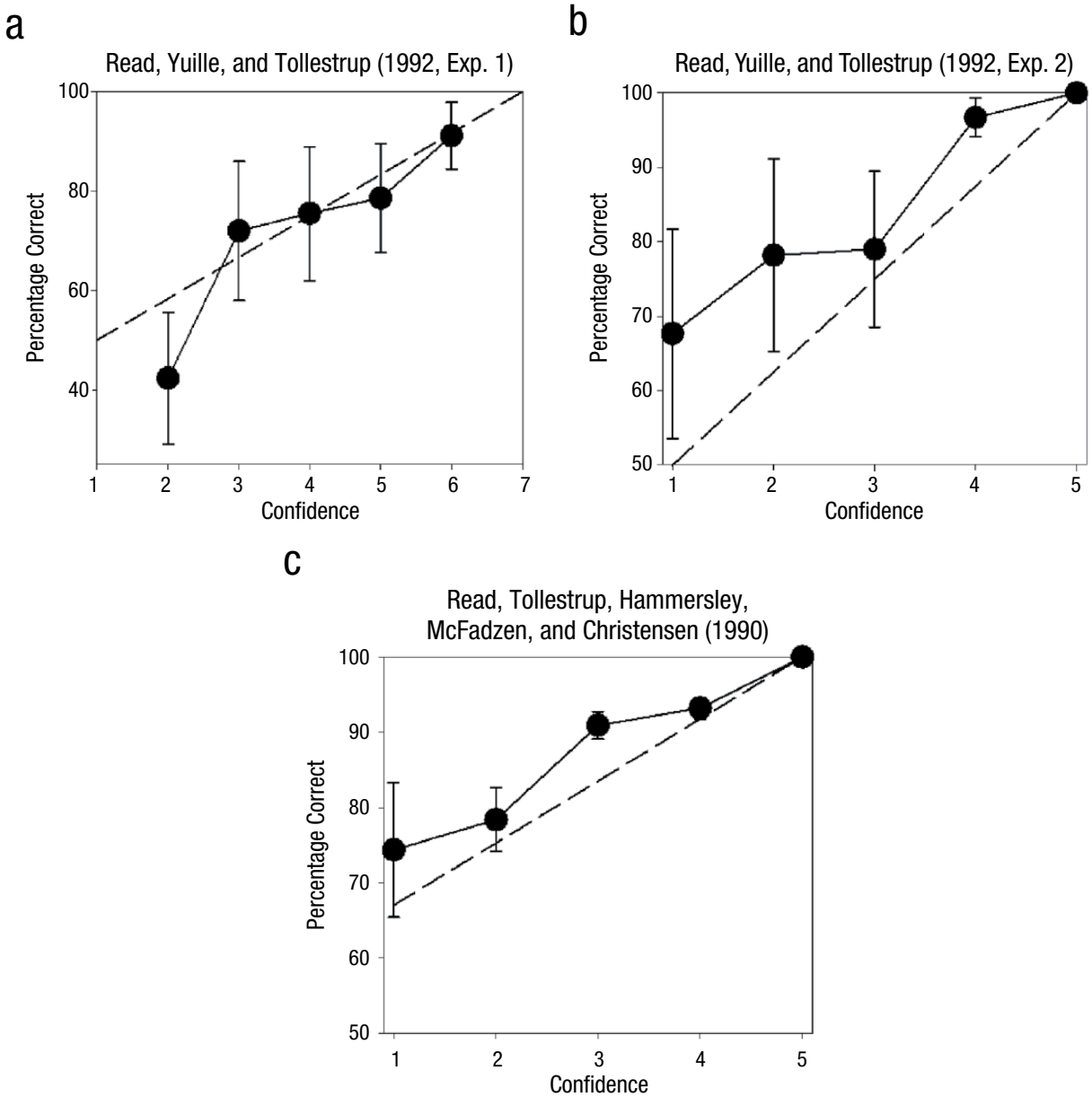

Fig. 3. Confidence-accuracy characteristic curves for three studies from the original Sporer et al. (1995) meta-analysis. The dashed lines indicate chance accuracy.

who were tested using different lineups, but we combined these data in our analysis because the results were quite variable when plotted separately because of the relatively small number of participants tested.

The point-biserial correlation coefficients for choosers in these three experiments were, respectively, .246, .511, and .359. The mean and standard deviation of these three values (M = .37, SD = .13) are similar to the corresponding values in the full data set (M = .41, SD = .16). In other words, these data appear to be reasonably representative of the larger set of studies analyzed by Sporer et al. (1995). What do these same data look like when plotted as a CAC curve? Figure 3 shows the CAC curves for the three obtainable data sets from the experiments analyzed in the original Sporer et al. (1995) meta-analysis. In each figure, a dashed diagonal line has been drawn for reference purposes. The line is drawn in such a way that the lowest confidence rating corresponds to chance suspect-ID accuracy and the highest confidence rating corresponds to 100% correct (perfect accuracy).

Note that chance accuracy corresponds to the level of performance that would be obtained if choosers randomly identified individuals from lineups. If an equal number of target-present and target-absent lineups were used, then chance accuracy for suspect IDs would be 50%. For example, if 1,000 "choosers" randomly sampled

from six-person target-present lineups, about 1/6 of them would land on the guilty suspect. Thus, this group would, on average, identify (1 / 6) × 1,000 = 167 guilty suspects. Similarly, if another 1,000 choosers randomly sampled from six-person target-absent lineups, about 1/6 of them would land on the innocent suspect. Thus, this group would identify 1,000 fillers and an estimated (1 / 6) × 1,000 = 167 innocent suspects. In other words, each group would identify the same number of suspects. Half of the suspect IDs would be randomly made to guilty suspects and half would be randomly made to innocent suspects. Thus, suspect-ID accuracy would be 167 / (167 + 167) = .50. Generally speaking, when using CAC analysis, random chance accuracy is equal to the base rate of target-present lineups used in a study.

For the two experiments from Read et al. (1992) shown in Figure 3a and 3b, the base rate of target-present lineups was approximately 50%. Thus, random chance suspect-ID accuracy in these two studies was 50% correct (and, of course, perfect accuracy is 100% correct). For the experiment from Read et al. (1990) shown in Figure 3c, the base rate of target-present lineups (and, therefore, chance accuracy) was approximately 67%. In every plot we show (both here and for all subsequent calibration and ROC experiments we consider), the diagonal line represents the full range of performance from chance accuracy (usually 50%) to perfect accuracy (100%).

By any reckoning, the data shown in Figure 3 exhibit a strong relationship between confidence and accuracy.[1] This is true despite the fact that the very same data are associated with a mean point-biserial correlation between confidence and accuracy of .37, meaning that only 14% of the variance was accounted for (i.e., $.37^2 = .14$). The data in Figure 3a range from 42% correct for the lowest level of confidence (a rating of 2 on the 7-point scale) to 91% correct for the highest level of confidence (a rating of 6 on the 7-point scale). Participants in that particular study used confidence ratings of 1 and 7 too rarely to estimate performance associated with the lowest and highest possible confidence levels. A similar range of performance is evident for the data shown in Figure 3b (low-confidence accuracy = 62% correct; high-confidence accuracy = 100% correct). In Figure 3c, a smaller range is evident but only because low-confidence accuracy was fairly high because of the high target-present base rate in that study.

Note that the highest levels of confidence in Figure 3b and 3c are both associated with 100% accuracy. Averaged across the three experiments, accuracy associated with the lowest level of confidence was 61.4% correct (this score would be slightly lower had all three studies involved a 50% target-present base rate), whereas accuracy associated with the highest level of confidence was 97.0% correct. That is, according to these studies, which are associated with an average point-biserial correlation of only .37, low confidence implies low accuracy, whereas high confidence implies very high accuracy. These are some of the same studies that have helped to convince the legal system to disregard eyewitness confidence because the correlation between confidence and accuracy is low.

A reanalysis of later research using the calibration approach

Next, we review studies that reported calibration curves for choosers and were designed to create pristine testing conditions. Most of these studies did not report choosing rates for each filler ID in target-absent lineups, so we cannot be sure that a fair lineup was used in every case. Nevertheless, in these studies, the same decision rule was used to select the fillers and the replacement photo, so the results, considered in the aggregate, represent what would likely be observed when pristine testing conditions are used. Every one of these studies, many of which come from the Neil Brewer lab, has shown a strong relationship between confidence and accuracy, as the authors of these studies have repeatedly emphasized. The calibration plots in the published literature have generally counted only suspect IDs from target-present lineups while counting all filler IDs (not just estimated suspect IDs) from target-absent lineups. Here, we replot those same data in the form of CAC plots, which means we focused on correct suspect IDs from target-present lineups and (usually estimated) incorrect suspect IDs from target-absent lineups.

Recently, a number of studies using ROC analysis in the context of eyewitness identification have been published. These studies have used confidence ratings to construct the ROC, but most were not specifically concerned with the confidence-accuracy relationship. Still, these studies also provide the data needed to construct a CAC plot, so we included ROC studies as well. More specifically, we included in our analysis the calibration and ROC studies that met the following criteria:

1. The studies investigated recognition memory for faces (not recall of details).
2. The participants were adults.
3. The lineups were designed to be fair in that the replacement photo for target-absent lineups was chosen using the same decision rule that was used to choose the other fillers (we consider unfair lineups in a later section).
4. Confidence ratings were taken soon after the ID (5 minutes or less post-ID).
5. Memory was tested using a lineup.

Table 1. Confidence-Accuracy and ROC Studies Included in Our Review

Study	Experiment	Original figure or table	Notes
1*	Brewer, Keast, and Rishworth (2002)	Figure 1	
2*	Brewer and Wells (2006)	Figure 1	
3	Carlson and Carlson (2014)	Figure 1	Reanalysis of raw data that were supplied by first author; the 7-point confidence scale was reduced to a 3-point scale to reduce error variance; and we excluded a condition involving photos with an artificial feature (all faces in the lineup had a large black letter *N* sticker on one cheek) because it seemed far removed from the forensic situation.
4*	Carlson, Dias, Weatherford, and Carlson (2016)	Figure 2	
5*	Dobolyi and Dodson (2013)	Figure 2	
6*	Dodson and Dobolyi (2016)	Figure 1	Multiple numerical confidence scales were used; all were converted to 0%-to-100% scales.
7*	Horry, Palmer, and Brewer (2012)	Figure 3	
8	Juslin, Olsson, and Winman (1996)	Figure 4	This study used a 75% target-present base rate; accuracy scores were estimated as described in the appendix of Wixted, Read, and Lindsay (2016).
9*	Keast, Brewer, and Wells (2007)	Figure 1	This study reported data for adults only, which were a subset of the data in Brewer and Wells (2006).
10	Lindsay, Nilsen, and Read (2000)	Table 3	The 11-point confidence scale was reduced to a 3-point scale (*low*, *medium*, and *high*) to reduce error variance.
11*	Mickes (2015)	Figure 2; Figure 4	Data were collapsed across the recollection and no-recollection conditions of Experiment 1 because too few low-confidence IDs were obtained to yield stable accuracy estimates; simultaneous lineup data shown in Figure 4 (Experiment 2) are also included in our plot.
12*	Mickes, Flowe, and Wixted (2012), Experiment 1a and 1b combined	Figure 6a	We reanalyzed the raw data (ROC data were reported in the original article).
13*	Palmer, Brewer, Weber, and Nagesh (2013), Experiment 1	Figure 1	
14*	Palmer, Brewer, Weber, and Nagesh (2013), Experiment 2	Figure 3	
15	Read, Lindsay, and Nichols (1998), Experiment 3	Figure 6.4	Data were collapsed across the prewarned and nonwarned conditions to reduce random error; raw data provided by the first author were reanalyzed.
16*	Sauer, Brewer, and Wells (2008)	Table 3	
17*	Sauer, Brewer, Zweck, and Weber (2010)	Figure 1	
18*	Sauerland and Sporer (2009)	Figure 3	
19	Smith and Flowe (2014)	Figure 2	We reanalyzed the ROC data reported in the original article.
20*	Weber and Brewer (2004)	Figure 5	Mini-lineups (four members) were used in this experiment.

Note: The 15 studies marked with an asterisk all used a 100-point confidence scale.

Relevant studies were identified by searching the Web of Science database using the keywords "calibration," "confidence," and "eyewitness identification." In addition, we searched references cited by the identified studies, and we examined all studies that later cited the articles included in our review. We do not claim this to be an exhaustive review, but it is undoubtedly a large and representative sample of calibration studies. The studies that satisfied these criteria and that were included in our review are listed in Table 1. We also included Read, Lindsay, and Nichols (1998) and D. S. Lindsay, Nilsen, and Read (2000) even though those authors did not specifically present their data as a calibration curve or as an ROC curve. However, J. Don Read provided us with the raw data from Read et al. (1998), and D. S. Lindsay et al. (2000) presented their data in enough detail that a CAC plot could be constructed. We further included the adult sample from Keast, Brewer, and Wells (2007) even though

their study used a subsample of adult participants who were tested by Brewer and Wells (2006), which is also included in our review.

Panels (a) through (s) of Figure 4 present the CAC plots from the calibration studies and ROC studies that we identified. Some of the studies included their raw data, making it possible to directly compute suspect-ID accuracy. For studies that did not, we precisely estimated accuracy scores from their calibration plots using WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/) and converted the reported accuracy score that included filler IDs to one that included only suspect IDs. This was accomplished by taking the reported accuracy score for a given level of confidence, $a1$; converting it to an odds score, o, where $o = a1 / (1 - a1)$; and then computing suspect-ID accuracy, 100% × $a2$, using the formula $a2 = o / (o + 1 / n)$, where n = lineup size. An example showing how this works is presented in Appendix B. Figure 4 does not show error bars because it was not possible to compute them when the data were estimated. However, an aggregate plot presented later in Figure 5 provides an indication of the consistency across studies. Note that most of the studies on the confidence-accuracy relationship that reported only the point-biserial correlation could not be included in our review because there is no way to produce a CAC plot when all that is known is a correlation coefficient. Overall, four studies that originally reported a point-biserial correlation coefficient were considered here: the three studies shown in Figure 3 and the study by Read et al. (1998) shown in Figure 4n. As noted above, it was possible to include these studies because J. Don Read still had (and provided us with) the raw data.

Most of the studies we review reported data from multiple conditions, so for each study shown in Figure 4, we have plotted the results from the individual conditions on the left and the results aggregated across conditions on the right. The results are presented alphabetically by first author, except for the D. S. Lindsay et al. (2000) and Sauerland and Sporer (2009) studies, which are both shown in the final panel (Fig. 4s) because they had only one condition each. Figure 4b is based on the same data we used earlier to illustrate calibration curves for choosers (Fig. 2a). Generally speaking, the average plots on the right for the studies with multiple conditions in Figure 4 are representative of the individual-condition plots on the left, so the bottom-line story from those studies can be most easily appreciated by scanning the plots on the right. It is visually apparent that in most cases, high-confidence accuracy is very high (95%–100% correct), whereas low-confidence accuracy is obviously lower.

Fifteen of the relevant studies reported their data on a 100-point confidence scale. Most reported their results using the following scale: 0–20, 30–40, 50–60, 70–80, and 90–100. In a few studies, a 6-point scale was used consisting of 0%, 20%, 40%, 60%, 80% and 100% confidence. For those, we collapsed the 0% and 20% ratings together to create a 5-point scale so the data could be averaged with data from the other studies using a 100-point scale. Across those 15 experiments, the average accuracy of a low-confidence (0–20) ID was 63.7% correct (range = 37.5%–83.3%), whereas the average accuracy of a high-confidence (90–100) ID was 97.1% correct (range = 94.2%–99.7%). The resulting aggregate CAC curve is shown in Figure 5a.

Overall, the data from the calibration studies reviewed here tell the same story as the data from the experiments included in the original Sporer et al. (1995) meta-analysis that we were able to reanalyze (Fig. 3). Confidence is highly predictive of accuracy in the straightforward sense that low-confidence suspect IDs are error prone (though often well above 50% chance, so such IDs are somewhat probative of guilt) whereas high-confidence suspect IDs are largely, but not perfectly, accurate. Moreover, these data indicate that, with respect to suspect-ID accuracy, eyewitnesses are, if anything, *under*confident (not overconfident). Note that all of these studies were methodologically similar to the earlier studies that were reviewed by Sporer et al. (1995), which are the studies that have helped to convince the legal system to increasingly disregard eyewitness confidence. What differs is how the data are analyzed, and that difference changes the story of the relationship between eyewitness confidence and accuracy as it is currently understood by the legal system (based largely on the point-biserial approach).

Figure 5b shows the average calibration plot (counting filler IDs from target-present lineups as errors). Although the data shown in Figure 5a are of most interest to judges and juries, the data shown in Figure 5b are certainly of interest to scientists. This plot is relevant to the question of how well eyewitnesses can express confidence in a way that corresponds to their subjective impression of accuracy. Any viable theory of eyewitness confidence would have to accommodate these data as well. Remarkably, the data exhibit almost perfect calibration (cf. Juslin et al., 1996).

Unfair lineups

As indicated earlier, our conclusions about the relationship between confidence and accuracy apply to initial IDs made from fair lineups without undue influence from a lineup administrator. A fair lineup is one in which everyone in the lineup resembles the perpetrator to the same approximate degree, so the suspect would not be identified more often than chance by a group of mock witnesses provided with the perpetrator's description.

The situation is undoubtedly different when unfair lineups are used. An unfair lineup is one in which the suspect stands out from the fillers such that the suspect

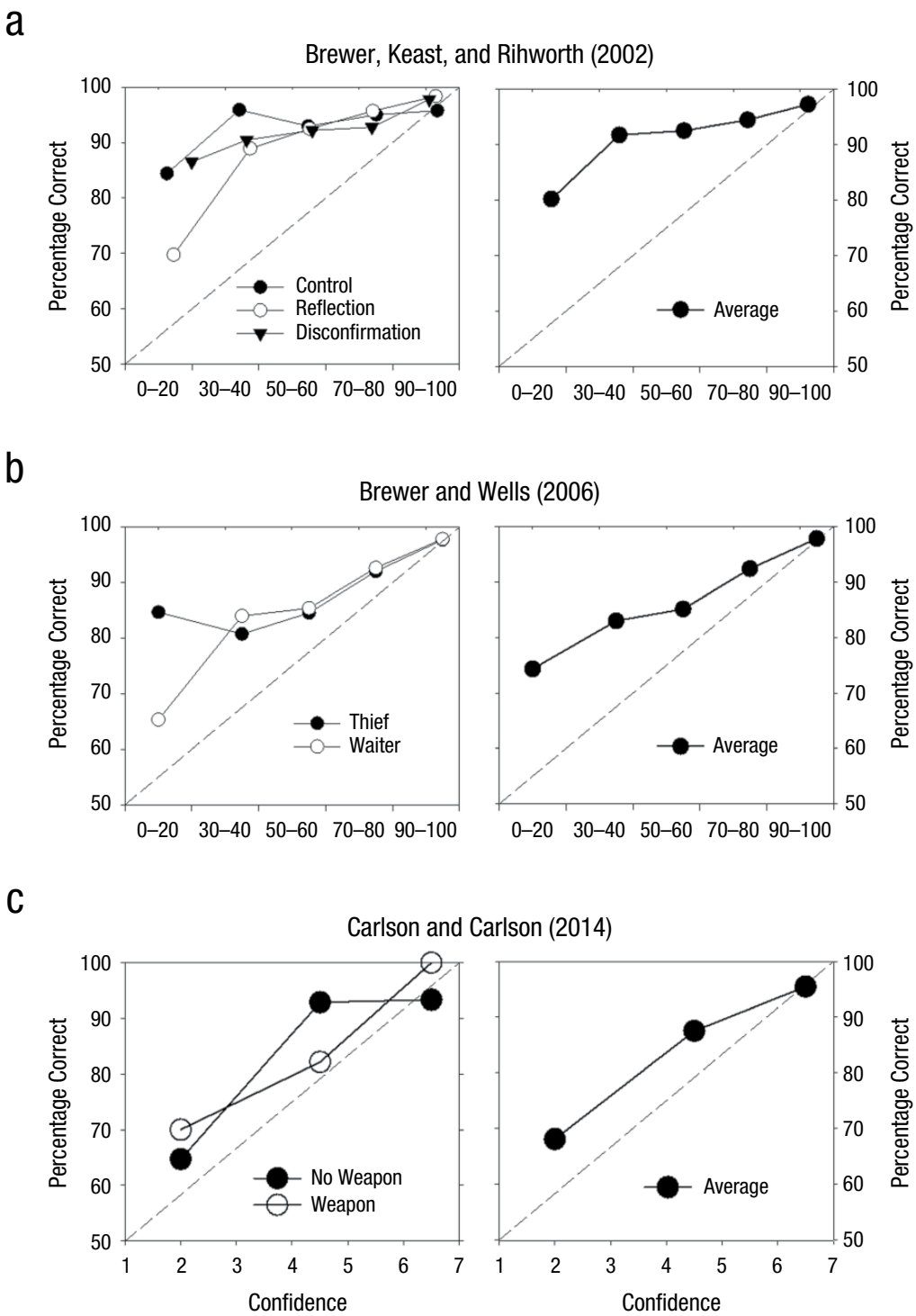

Fig. 4. (continued)

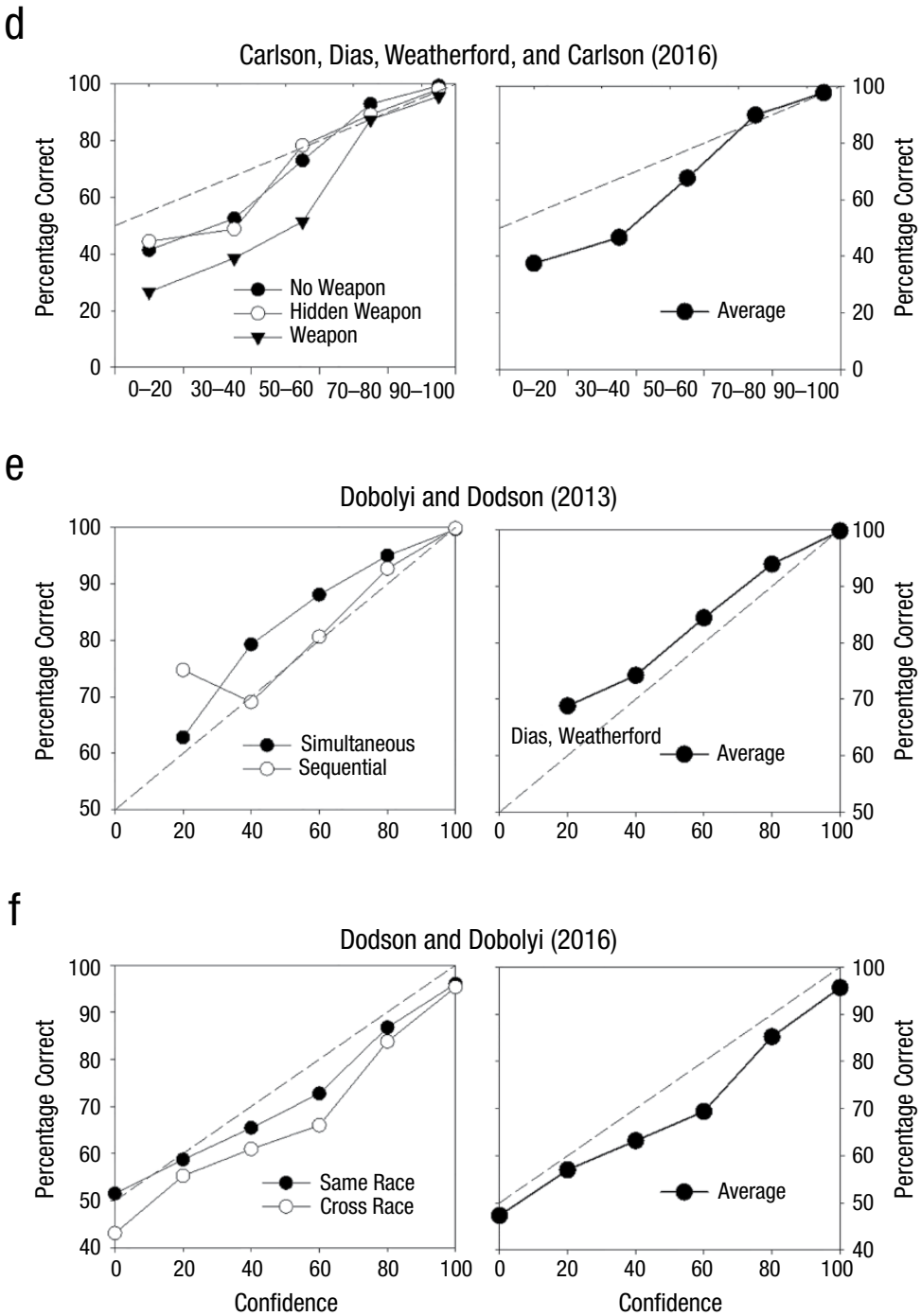

Fig. 4. (continued)

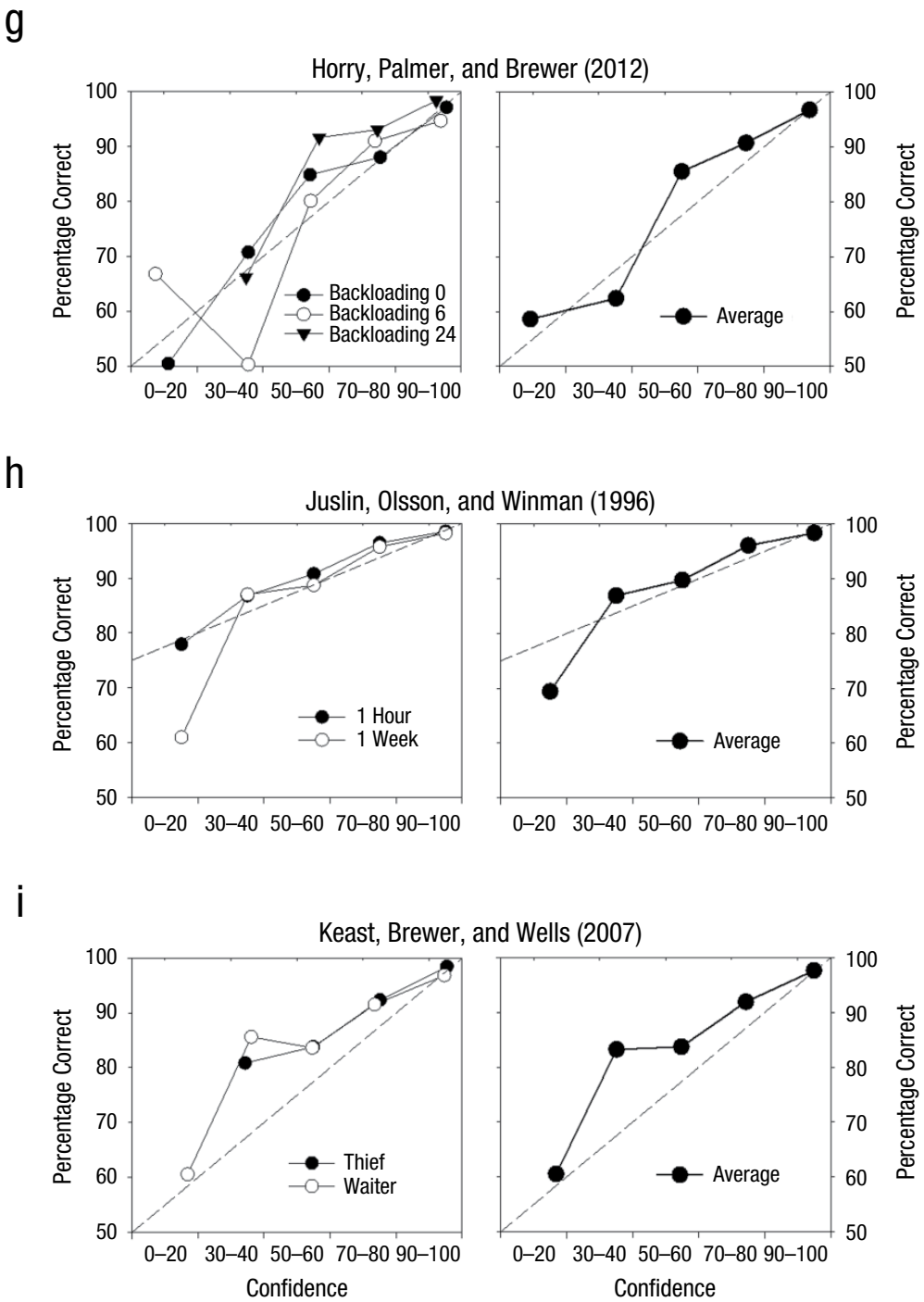

Fig. 4. (continued)

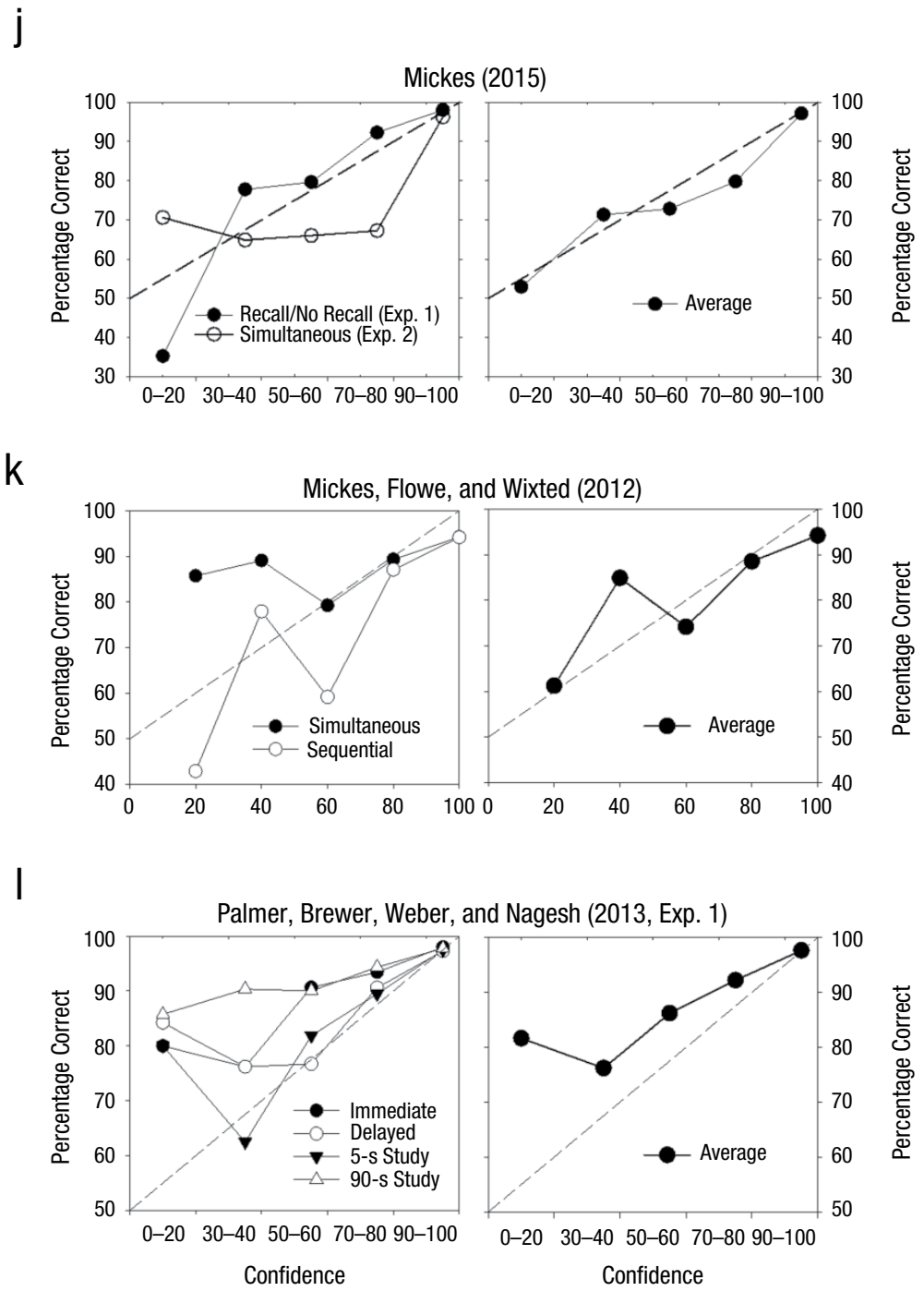

Fig. 4. (continued)

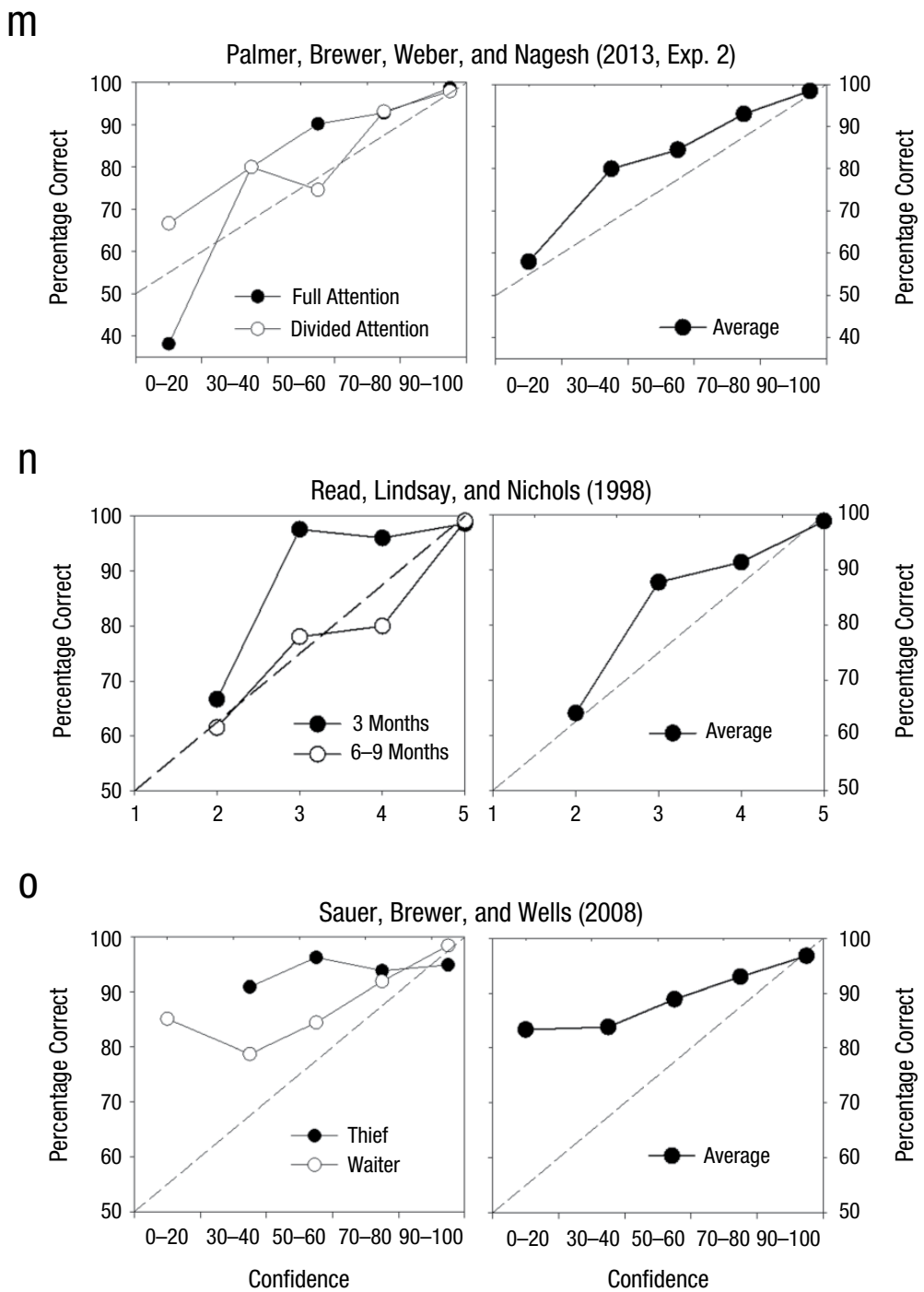

Fig. 4. (continued)

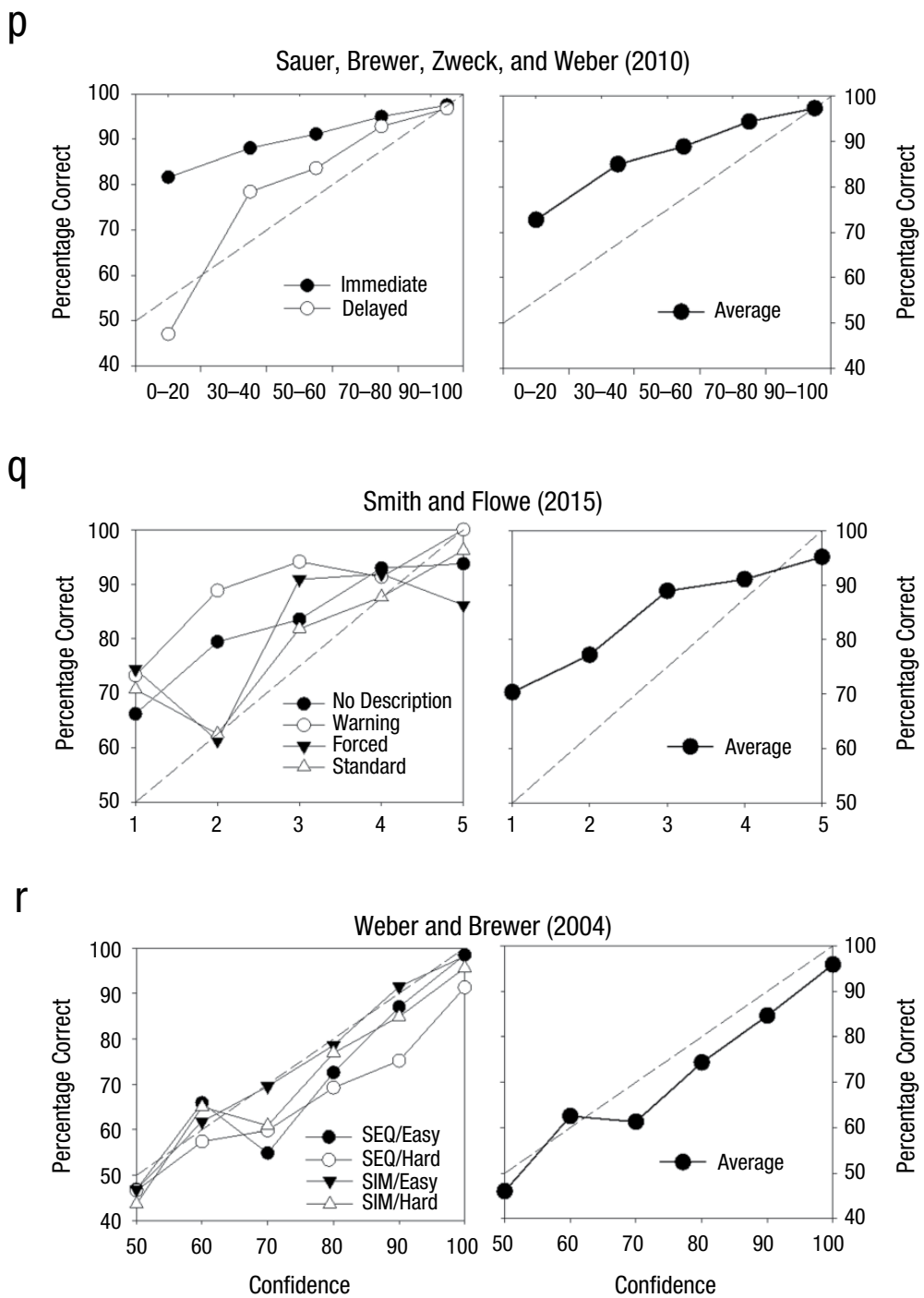

Fig. 4. (continued)

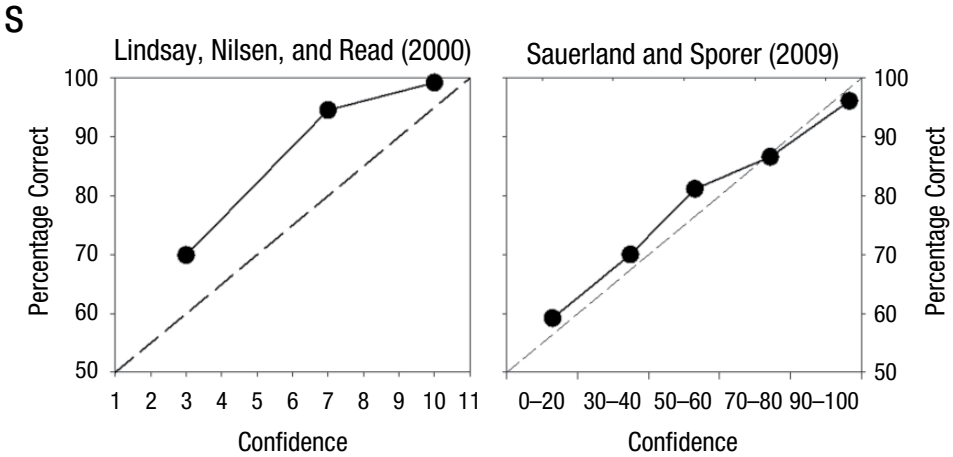

Fig. 4. Suspect-ID accuracy (percentage correct) as a function of confidence for the 20 studies listed in Table 1. SEQ = sequential; SIM = simultaneous.

(innocent or guilty) resembles the perpetrator to a noticeably greater extent than the fillers. It is well known that an unfair lineup leads to a higher rate of suspect identification and higher confidence in that identification, whether or not the suspect is the perpetrator (Fitzgerald, Price, Oriet, & Charman, 2013; R. C. L. Lindsay & Wells, 1980; Wells, Rydell, & Seelau, 1993). It stands to reason that unfair lineups would also reduce the utility of eyewitness confidence and would decrease the reliability of high-confidence IDs. In the extreme, placing the perpetrator's identical twin in a target-absent lineup would undoubtedly yield many incorrect high-confidence IDs of the innocent suspect, wreaking havoc on the accuracy of high-confidence suspect IDs.

Gronlund, Carlson, Dailey, and Goodsell (2009) conducted a large-scale investigation into the diagnostic accuracy of simultaneous and sequential lineups using target-absent lineups in which the designated innocent suspect resembled the perpetrator more than the fillers did. They also varied how much the picture of the guilty suspect resembled what the perpetrator looked like while committing the crime. In some conditions of that experiment, performance was near chance (e.g., when the perpetrator's appearance had substantially changed and the

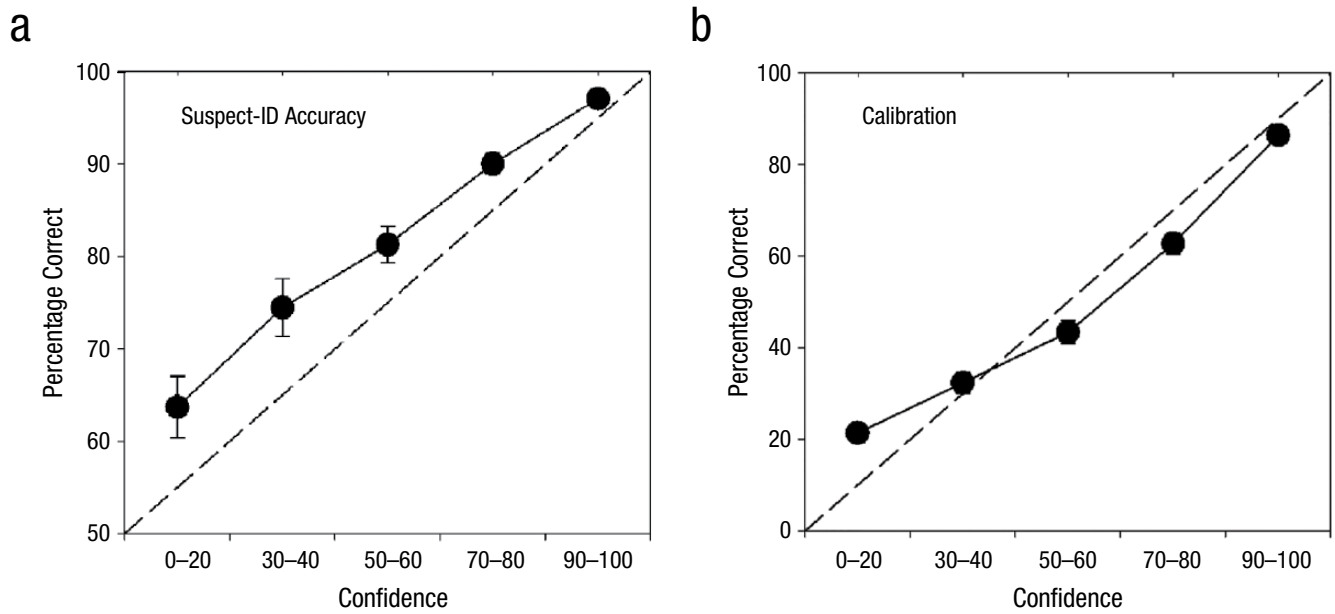

Fig. 5. Suspect-ID accuracy averaged across 15 studies with comparable scaling on the confidence (*x*-) axis (a), and the same data plotted as a calibration curve (b). The studies included in this analysis, which all used a 100-point confidence scale, are indicated in Table 1 with an asterisk.

innocent suspect looked a lot like the original view of the perpetrator). In other conditions, performance was above floor, and Gronlund et al. (2012) reported confidence-based ROC data for those conditions. Collapsed across simultaneous and sequential lineups, the target-absent lineups from the conditions they analyzed were still unfair in the sense that the innocent suspect more closely resembled the perpetrator than the fillers did (so the innocent suspect was identified with much higher probability than the individual fillers were). Thus, these data can be used to gain some insight into the confidence-accuracy relationship when unfair lineups are used. Figure 6a presents the CAC plots for simultaneous and sequential lineups from Gronlund et al. (2012). The data exhibit a strong relationship between confidence and accuracy, but high-confidence accuracy (88% correct) is noticeably lower than it has been for the fair lineups considered to this point. The lower accuracy score for high-confidence IDs presumably reflects the impact of lineup unfairness. Note that the other conditions in their experiment, which yielded chance performance (because in some cases the photo of the innocent suspect resembled the perpetrator more than the photo of the guilty suspect did) would clearly wreak havoc on the confidence-accuracy relationship.

A clear illustration of the effect of unfair lineups can be observed by analyzing some of the data reported by Mickes, Flowe, and Wixted (2012). In their Experiment 2, the innocent suspect in the target-absent lineups was an altered photo of the perpetrator himself. The perpetrator's photo was altered using Photoshop to change the hair color, skin tone, nose shape, and face shape. Because these changes were all relatively minor, this experiment approximated a situation in which target-absent lineups contained a near twin of the perpetrator. As might be expected, the researchers' ROC analyses indicated that overall performance was rather poor. For simultaneous lineups, the overall correct-ID rate was .50 and the false-ID rate was .26 ($d' = 0.63$). For sequential lineups, the correct-ID rate was .42 and the false-ID rate was .22 ($d' = 0.55$). Of more interest for present purposes are the CAC plots shown in Figure 6b. Obviously, the relationship between confidence and accuracy is weaker than what is observed for fair lineups. Perhaps even more importantly, high-confidence suspect-ID accuracy is quite low (near 70% correct for simultaneous lineups and sequential lineups). As noted earlier, Sučić et al. (2015) also arranged unfair lineups, and the data from that study (expressed as a CAC plot) are shown in Figure 6c. Once again, high-confidence accuracy (only 85% correct) falls well below what is typically observed when fair lineups are used.

Two recent studies are particularly informative because they directly compared fair versus unfair lineups. Wetmore et al. (2015) tested participants using six-person simultaneous lineups either immediately after watching a mock-crime video or following a 48-hour delay. In their conditions in which the innocent suspect had only moderate similarity to the perpetrator (what they referred to as the "InnocentWeak" condition), some lineups were fair and others were biased against the innocent suspect. Their Table 2 presented choosing rates for the designated innocent suspect in each condition (fair vs. biased), so it was possible to determine that their manipulation of lineup fairness was successful. That is, the innocent suspect was disproportionately chosen over the other fillers in the biased condition only. Figure 6d presents the CAC results from that study collapsed across the retention-interval manipulation. As would be expected, the data from the fair condition are similar to the data presented earlier in Figure 4. Specifically, confidence is a strong predictor of accuracy, and high-confidence accuracy is very high (100% correct for confidence ratings of 7; 96% correct for confidence ratings of 6). However, in the biased condition, high-confidence accuracy is far lower (80% correct for confidence ratings of 7; 75% correct for confidence ratings of 6).

A similar pattern was evident in a recent study by Colloff, Wade, and Strange (2016). Participants watched a video of a perpetrator who had a distinctive feature, such as a black eye. In the unfair condition, the distinctive feature appeared only on the suspect in both target-present and target-absent lineups, not on any filler. Thus, whether innocent or guilty, the suspect stood out. In their fair conditions, by contrast, the distinctive feature either was present on all lineup members or was covered up for all lineup members (the data from several conditions in which the distinctive feature was added to or eliminated from all lineup members were very similar and have been averaged together here). As is apparent in Figure 6e, for the unfair condition, high-confidence accuracy was very low (~66% correct) and was much lower than high-confidence accuracy in the fair conditions (~86% correct). Although the effect of lineup fairness on high-confidence accuracy was consistent with other findings, high-confidence accuracy in the fair condition was noticeably lower than the ~95% correct levels of accuracy typically observed in the other studies reviewed here (e.g., Fig. 5a). Because there is no obvious reason for the observed difference, this result serves as a reminder that the determinants of high-confidence accuracy are not fully understood and that more research is needed to identify the conditions under which high-confidence accuracy can be compromised even when fair lineups are used.

These findings underscore the critical point that our claims about the relationship between confidence and accuracy (and, in particular, the very high level of accuracy usually associated with high-confidence suspect IDs) apply to fair lineups, not to unfair lineups. As noted by

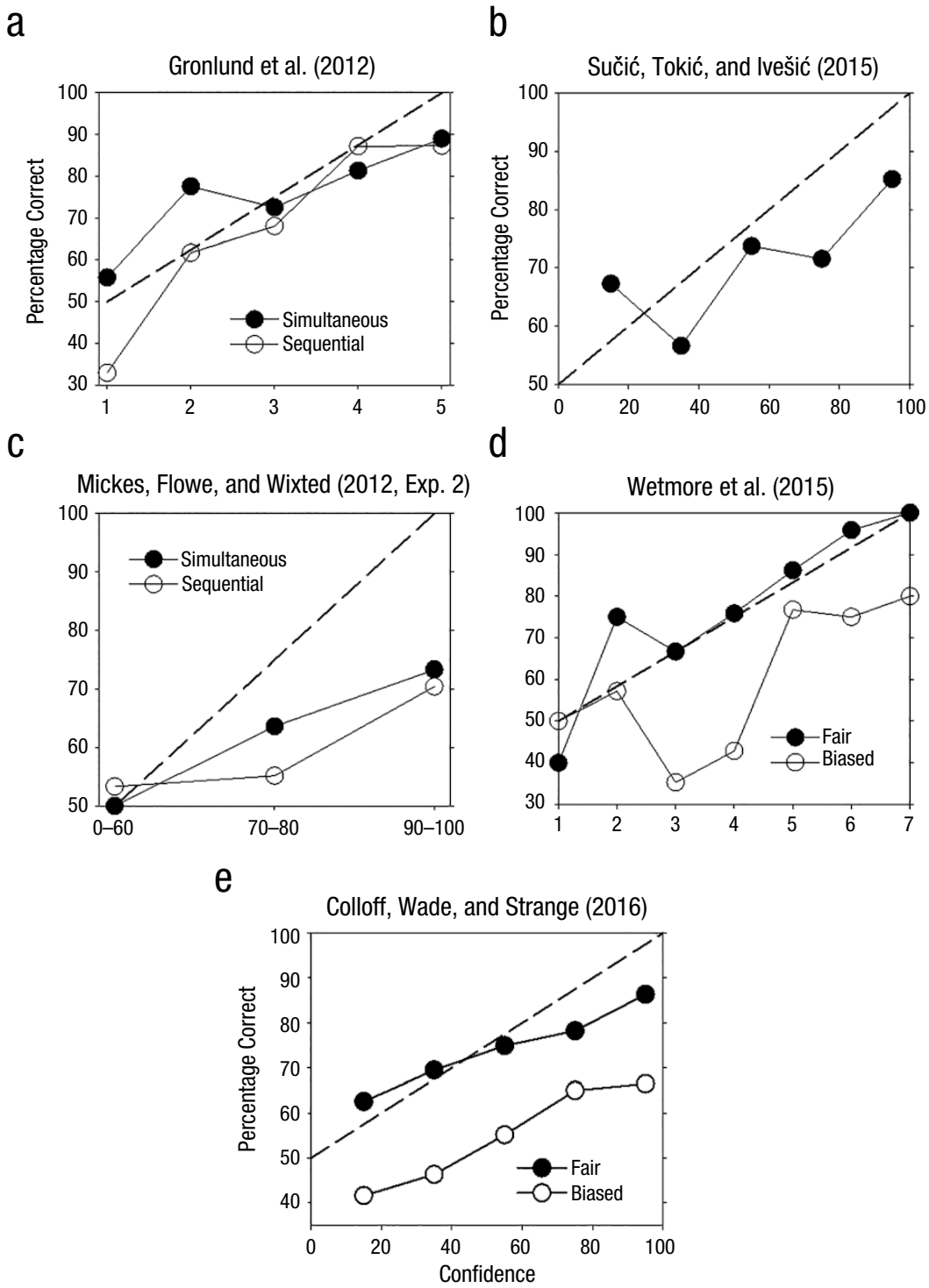

Fig. 6. Confidence-accuracy characteristic plots from studies that used unfair lineups.

Brewer and Palmer (2010), other circumstances in which the confidence-accuracy relationship may be degraded include (a) when the eyewitnesses are children (e.g., age 13 or younger), (b) when confidence ratings are not taken contemporaneously with the ID but are instead retrospective, and (c) when witnesses reject the lineup.

Police Department Field Studies

The advantage of a mock-crime study such as the ones considered above is that the experimenter knows if a suspect ID is correct or incorrect, thereby allowing a direct computation of suspect-ID accuracy. 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 if (a) the perpetrator is a stranger to the witness (so the suspect in the lineup is not chosen because of preexisting familiarity), (b) the lineup is fair (so the suspect is not chosen because he or she stands out), and (c) blind administration is used (so the suspect is not chosen by the witness because of administrator influence). Under those conditions, the only way that the witness can land on the suspect with a probability that exceeds $1/n$, where n is lineup size, is if the suspect matches the memory of the witness. Except in rare cases in which an innocent suspect bears an uncanny resemblance to the perpetrator despite the fact that the lineup procedure was pristine, a strong memory-match signal would usually happen because the suspect actually is the perpetrator. Thus, if confidence is predictive of accuracy in the real world, suspect IDs should occur with probability greater than $1/n$, and that probability should increase as a function of confidence. To our knowledge, only two police department field studies have used fair lineups that were blindly administered and also reported confidence data. Both of these studies yielded data suggesting that high-confidence IDs are highly reliable, whereas low-confidence IDs are much less reliable (just as the lab data summarized in Fig. 5a would suggest).

Hennepin County police department field study

Klobuchar, Steblay, and Caligiuri (2006) conducted a pilot study of 206 actual eyewitnesses who were tested using six-person sequential photo lineups in four municipal police departments in Hennepin County, Minnesota. The lineups were not specifically tested for fairness but were presumably fair because department policy required the use of photographs depicting individuals of similar age, skin color, complexion, hairstyle, and build. The lineups were administered by an officer who was blind to the suspect's identity, and confidence was recorded in the witness's own words. Some lineups contained a suspect previously known to the witness, whereas other lineups contained a suspect previously unknown to the witness. The key measure was the frequency of *jump-out IDs*, which are rapid IDs accompanied by expressions of absolute certainty. In other words, jump-out IDs are high-confidence IDs.

Of 175 choosers in this study, 96 (55%) made jump-out IDs. Remarkably, 99% of these IDs were made to suspects, not fillers, which is to say that only one of the 96 jump-out IDs was made to a filler. From their Table 5, it was possible to determine that 26 of the jump-out IDs were made to strangers, and 70 were made to suspects previously known to the eyewitness. The stranger data are of interest here. The one jump-out ID that landed on a filler occurred in a stranger lineup (Nancy Steblay, personal communication, April 25, 2016); thus, 25 out of 26 jump-out IDs in stranger lineups (96%) landed on the suspect.

Keep in mind that there were 5 times as many fillers as suspects in any given lineup, so random responding for jump-out IDs would result in $26 \times (5 / 6) \approx 22$ filler IDs (yet only one was actually observed) and only about $26 \times (1 / 6) \approx 4$ suspect IDs (yet 25 were actually observed). Thus, the number of suspect IDs made with high confidence in this study was far greater than would be expected by chance. It is possible that the lineups in the Hennepin County study were not fair lineups. But if they were fair lineups (as they were designed to be), it is hard to come up with a logical explanation for these results without assuming that high-confidence accuracy was close to perfect. IDs made with lower confidence in that study (non-jump-out IDs) landed on the suspect much less often, approximately 60% of the time. That is still much more often than would be expected by chance alone, so even these more error-prone suspect IDs appear to be somewhat probative of guilt. These results suggest a strong confidence-accuracy relationship that is not appreciably different from that revealed by the lab results depicted in Figure 5a.

Houston Police Department field study

Another recent police department field study was specifically designed, in part, to examine the information value of eyewitness confidence (Wixted et al., 2016). In this study, eyewitness decisions were recorded from six-person photo lineups administered as part of criminal investigations in the Robbery Division of the Houston Police Department between January 22 and December 5, 2013. This study involved the administration of 348 simultaneous and sequential lineups, the investigators were unaware of the identity of the suspect in each lineup (i.e., double-blind administration was used), and the lineups involved suspects who were unknown to the eyewitnesses prior to the crime. Lineup fairness was examined for a random sample of 30 photo lineups by providing

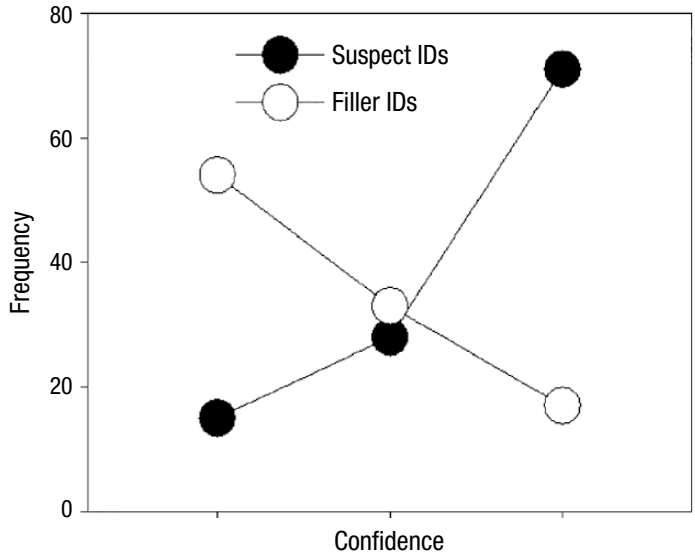

Fig. 7. Suspect IDs and filler IDs made with low, medium, and high confidence in the Houston Police Department field study (Wixted, Mickes, Clark, Dunn, & Wells, 2016).

the selected photo lineups to 49 mock witnesses and asking them to try to identify the suspect based only on the suspect's physical description. As noted above, in a fair six-person lineup, the suspect should be identified by a mock witness only 1/6 (.17) of the time. The mean proportion of suspect IDs made by the mock witnesses (.18) did not differ significantly from the expected value for a fair six-member lineup, $t(29) = 0.76$. Thus, according to this measure, the 30 lineups that were randomly selected were, on average, fair. For purposes of our analyses, we assumed that the remaining lineups were also fair. 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*).

The critical results are reproduced in Figure 7. Obviously, most suspect IDs were made with high confidence, whereas most filler IDs were made with low confidence. This pattern again immediately suggests a strong confidence-accuracy relationship. Moreover, as with Klobuchar et al. (2006) and in agreement with lab studies (Fig. 5a), high-confidence IDs appear to have been highly accurate. Even though there were 5 times as many fillers as suspects in the police lineups used in this study, high-confidence IDs landed on the suspect 72 times and landed on a filler 17 times. Using perfectly fair lineups, one would expect 5 times as many high-confidence filler IDs as high-confidence suspect IDs. Thus, as a crude approximation, the 17 high-confidence filler IDs translate to an estimated $17 / 5 \approx 3$ high-confidence innocent-suspect IDs. If three of the 72 high-confidence suspect IDs were made to innocent suspects, it means that 69 of the 72 suspect IDs made with high confidence (96%) were correct. A formal signal-detection model fit to these data estimated high-confidence suspect-ID accuracy to be approximately 97% correct, whereas low-confidence suspect-ID accuracy was estimated to be closer to 50% correct. Again, these results are not dramatically different from the lab results summarized in Figure 5a.

Base Rates of Target-Present Lineups in the Laboratory and in the Real World

In most of the lab studies that we have considered here, the base rate of target-present lineups was 50%. An issue in generalizing from the lab to the real world is that the base rate of target-present lineups is unknown, and it is quite likely that the base rate will vary from one police department to another, or even from one detective to another, as a function of how much evidence an investigator requires before placing a possible suspect in a lineup (Wells, 1993). In order to explore the effect of different base rates, we used the data from Wetmore et al.'s (2015) fair lineups. Figure 6d showed a CAC on these data based on a 50% base rate. We created Bayesian curves called *prior-by-posterior curves* that map the probability that an identification of the suspect was accurate (i.e., that the suspect is the perpetrator) across all possible values of the base rate from 0% (all lineups had an innocent suspect) to 100% (all lineups had a guilty suspect). See Appendix C for a short tutorial on this

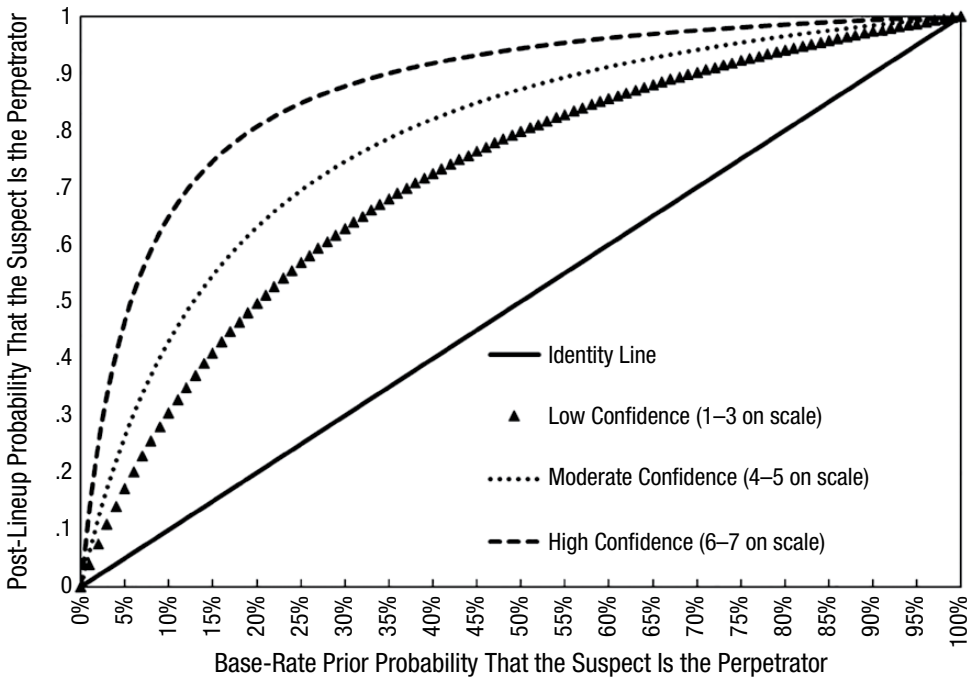

Fig. 8. Post-lineup probability that the suspect is the perpetrator as a function of the base rate of target present lineups and whether the ID was made with low, moderate, or high confidence. The data are from Wetmore et al.'s (2015) fair lineups.

Bayesian approach. With a sufficiently large sample size, we can create a curve for every level of confidence. But, because sample sizes can get small when every level of confidence is examined separately, we used three levels of confidence. The Wetmore et al. study used a 7-point confidence scale, so we collapsed confidence ratings of 1 through 3 into the category of low confidence, 4 and 5 into moderate confidence, and 6 and 7 into high confidence. These curves are shown in Figure 8.[2]

The solid line in Figure 8 is called an *identity line*, and it simply represents where the data would fall if the identification had no diagnostic utility. Clearly, all three curves are above the identity line and, as would be expected, the height of the curve for the high-confidence eyewitnesses is far above that of the curves for the moderate- and low-confidence witnesses.

Notice that the probability that the identified suspect is the perpetrator (which is the same as the probability that the witness is accurate) for high-confidence eyewitnesses remains relatively high (above 90%) until the base rate drops below 35%. Contrast that, however, with low-confidence witnesses, for whom the accuracy drops below 90% as soon as the base rate drops below 70%. In fact, whereas the high-confidence witnesses are still 90% accurate when the base rate is a mere 35%, the low-confidence witnesses drop all the way to a mere 63% accuracy if the base rate is 35%.

The data in Figure 8 underscore an important point made by Wells et al. (2015), namely that the base rate matters. Moreover, the base rate for lineups is a system variable. If a police department places a suspect who matches the perpetrator's description in a lineup on nothing more than a hunch, then the base rate of guilt in that jurisdiction is likely to be on the low side. Requiring at least some independent evidence of guilt (i.e., requiring more than just a hunch) will move a jurisdiction to the right on the base-rate dimension in Figure 8, thereby increasing the probability that an identification of a suspect is an accurate identification for all IDs made with any level of confidence.

What does law enforcement believe about the need to have evidence indicating that the suspect is likely to be the perpetrator before placing a suspect in the jeopardy of a lineup? A national survey indicated that more than one-third of U.S. crime investigators believed that they needed no evidence at all about the likely guilt of a person before placing that person in a lineup (Wise, Safer, & Maro, 2011). Behrman and Richards (2005) examined records from 306 lineups in Northern California in which a witness identified someone. They then coded how much evidence existed against the suspects before they were placed in a lineup. Behrman and Richards found that in 30% of the cases the evidence was "minimal," and in an additional 40% of the cases there was no pre-lineup

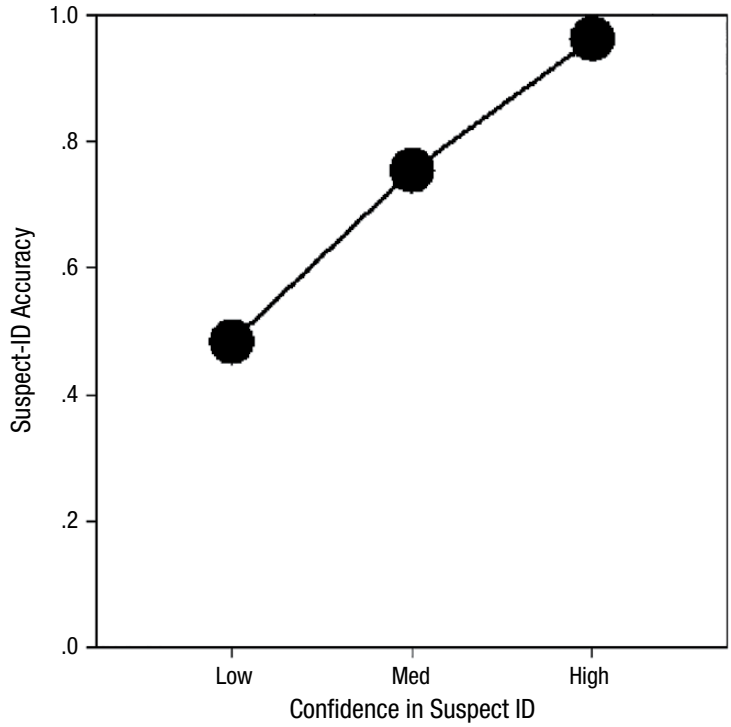

Fig. 9. Estimated suspect-ID accuracy as a function of confidence for the data from the Houston Police Department field study (Wixted Mickes, Clark, Dunn, & Wells, 2016). The estimates were based on a signal-detection model, which further estimated the target-present base rate of lineups to be 35%.

evidence at all. This does not tell us directly about what the base rates are in those jurisdictions, but it does not lend much confidence toward the idea that the base rate is high.

The base rate of guilt in lineups is generally assumed to be an unknowable variable in the real world. However, the signal-detection model used by Wixted et al. (2016) provided a principled estimate of the base rate in the Houston Police Department. The base-rate estimate that Wixted et al. reported—35%—is just that, an estimate, so it could be wrong. However, it is a principled estimate because it is based on a theory that has long guided thinking about recognition memory in other contexts. Moreover, it is a demonstration that, with the right theory, the base rate of guilt in a particular jurisdiction is not necessarily an unknowable value. The signal-detection model used by Wixted et al. (2016) may not be the right theory. And recall that the theory giving rise to this estimate assumed that the lineups were fair lineups based on an analysis of only a subset of the lineups. But the point is that base-rate information is not inherently unknowable, and the first principled estimate in a police jurisdiction came out surprisingly low. Fortunately, using pristine identification procedures, the laboratory data shown in Figure 8 suggest that, at a base rate of only 35%, confidence is highly predictive of suspect-ID accuracy, and high-confidence IDs are still quite accurate (about 90% in Fig. 8), whereas low-confidence IDs, despite having probative value, are highly error prone. Similarly high accuracy was obtained when the signal-detection model was used to estimate suspect-ID accuracy in the Houston field study assuming a 35% base rate of target-present lineups (Fig. 9). However, if the true base rate were much lower than that, then high-confidence IDs would begin to become highly error prone as well. Moreover, base rates likely differ from jurisdiction to jurisdiction, which means that some may fall well below the 35% estimate in Houston. Thus, conceptualizing the base rate as a system variable—and taking concrete steps to increase it—seems like a prudent strategy for law enforcement to consider.

Filler IDs and Non-IDs

In a police lineup, there are three possible decision outcomes: a suspect ID, a filler ID, and a non-ID (a rejection). To this point, we have focused on suspect IDs because those IDs are the ones that have often ended up putting an innocent person in prison, only to be exonerated by DNA evidence years later. Wells et al. (2015),

however, pointed out that the other two decision outcomes—filler IDs and rejections—also provide useful information about the chances that the suspect is the perpetrator. Unlike identifications of the suspect, however, the information value of rejections and of filler identifications is exculpatory rather than incriminatory.

The fact that rejections provide exculpatory information is somewhat obvious and stems from the simple observation that witnesses are more likely to reject the lineup (make a non-ID decision) if it is a target-absent lineup than if it is a target-present lineup. But it is somewhat more difficult to intuit that filler IDs also provide exculpatory information. Empirically, it has long been recognized that filler identifications are more likely to occur in response to target-absent lineups than in response to target-present lineups (Wells & Lindsay, 1980). Accordingly, it makes sense that filler identifications would have exculpatory value. In effect, a witness who identifies a filler is offering an opinion that there is a filler in the lineup who looks more like the perpetrator than does the suspect. And, of course, that means that filler IDs are more likely to happen when the suspect is not the perpetrator than when the suspect is the perpetrator (Wells et al., 2015).

The information that rejections and filler identifications provide can be expressed at different levels of witness confidence using prior-by-posterior curves just as we did with identifications of the suspect. Prior-by-posterior curves for rejections are shown in Figure 10a, and the curves for filler identifications are shown in Figure 10b using the data from Wetmore et al. (2015). The dependent measure in Figures 10a and 10b is the probability that the lineup is a target-present lineup (i.e., that the suspect is the perpetrator). Notice that, unlike identifications of the suspect (see Fig. 8), both rejections and filler identifications produce curves that fall below rather than above the identity (no information) line. That is because both rejections and filler identifications have exculpatory information value rather than incriminatory information value.

In the case of rejections, which are shown in Figure 10a, the vertical axis is equivalent to the proportion of witnesses who made a correct decision to not identify anyone from the lineup. As can be seen, high-confidence rejections produce a curve that is farther below the identity line than the lines produced by moderate- or low-confidence rejections. This reinforces an important point, namely that lineup administrators should be obtaining confidence statements from witnesses for rejection decisions at the time of identification in addition to collecting confidence statements for identifications of suspects.

In the case of filler identifications, which are shown in Figure 10b, the vertical axis does not represent the proportion of witnesses who made a correct decision. After all, all filler identifications are errors. Nevertheless, filler identifications have information value because a filler identification is more likely to occur when the suspect is not the perpetrator (target-absent lineup) than when the suspect is the perpetrator (target-present lineup). Notice that the exculpatory value of filler identifications can be as high, and sometimes more so, than the exculpatory value of rejections. In other data sets (e.g., Brewer & Wells, 2006), high-confidence filler identifications were more exculpatory than were lower levels of confidence, whereas in the Wetmore et al. (2015) data, it was moderate-confidence filler identifications that were most informative in the exculpatory direction (with low- and high-confidence filler identifications being equally informative in the exculpatory direction). However, this might be due to the fact that there were very few high-confidence filler identifications in the Wetmore et al. data, making the high-confidence filler-identification curves somewhat unstable. Wells et al. (2015) argued that high-confidence filler identifications should generally be more exculpatory than lower-confidence filler identifications because high confidence filler identifications indicate stronger confidence by the witness that the filler is a better match to the perpetrator than is the suspect.

The fact that filler identifications have exculpatory value is an important observation in light of evidence that law enforcement agencies often fail to make records of filler identifications. In their analyses of police files to score the outcomes of photo lineups in actual cases, researchers have found that lineup administrators failed to make records of filler identifications but always made records of suspect identifications (Behrman & Davey, 2011; Tollestrup, Turtle, & Yuille, 1994). Consistent with this, in a recent national survey, U.S. law enforcement agencies admitted that they do not even prepare a report of a lineup if the witness does not ID the suspect (Police Executive Research Forum, 2013). In a controlled experiment, Rodriguez and Berry (2014) assigned research participants to the role of lineup administrators who were either blind to which lineup member was the suspect or knew which lineup member was the suspect and which were fillers. When participant-administrators were blind, they made records of all of the witnesses' identifications (both suspect IDs and filler IDs). When the participant-administrators were not blind, however, they commonly failed to make records of filler IDs. Hence, this is yet another argument in favor of why eyewitness identifications should be conducted using double-blind procedures. In the absence of double-blind procedures, the results can be selectively reported.

Another important point is that the exculpatory value of filler identifications and rejections (pointing toward innocence) is generally less than the incriminating value

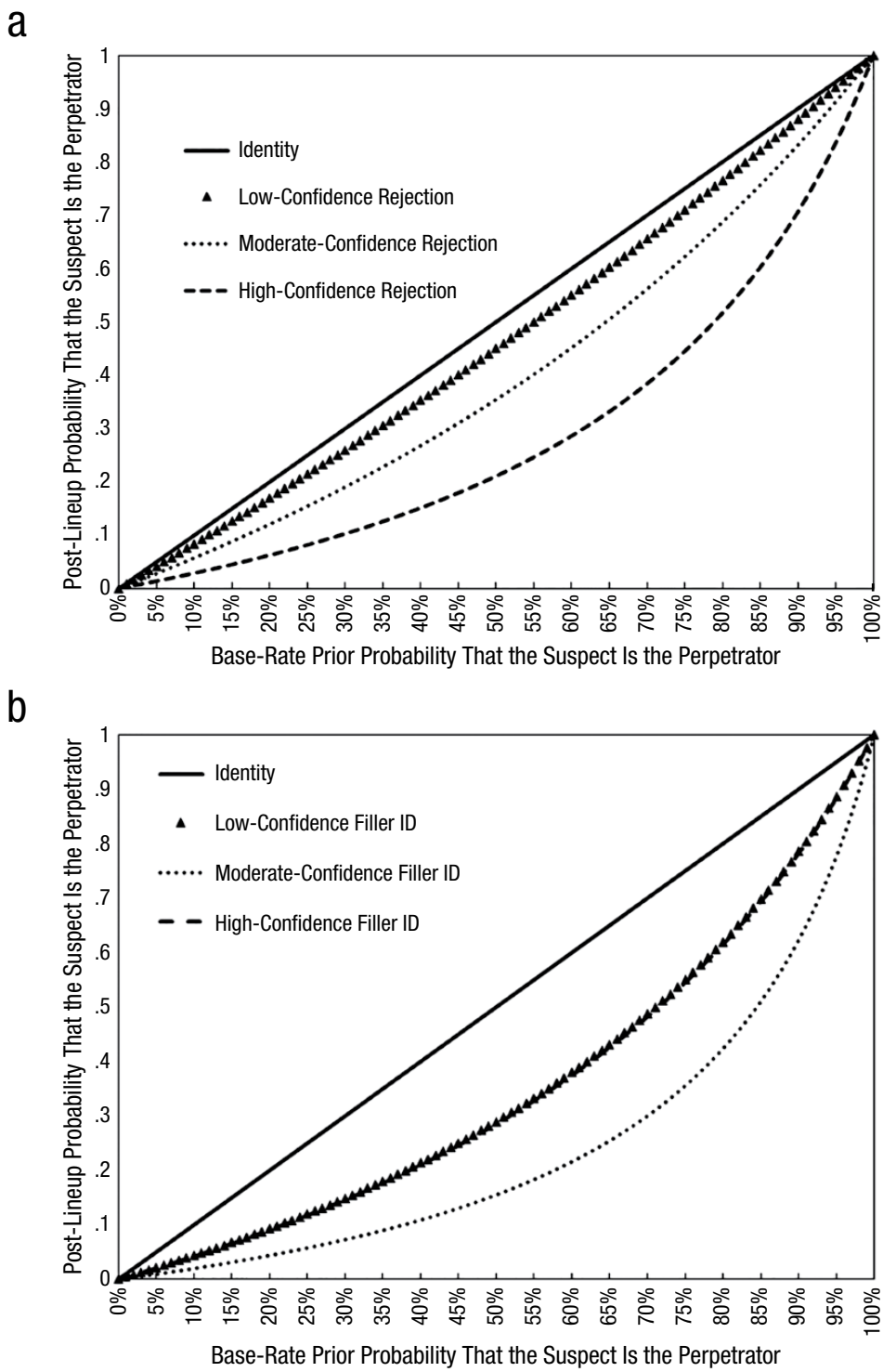

Fig. 10. Post-lineup probability that the suspect is the perpetrator as a function of the target-present base rate for lineups and the confidence of the witness who rejected the lineup (a) or identified a filler (b). The data are from Wetmore et al.'s (2015) fair lineups.

of identifications of the suspect (pointing toward guilt). This is apparent from noting that the area under the prior-by-posterior curves for identifications of the suspect (see Fig. 8) is greater than the area under the curve for rejections or identifications of fillers (Fig. 10). That type of pattern in the broader eyewitness-identification

literature led Wells et al. (2015) to conclude that lineups are more effective for incriminating suspects than they are for exculpating suspects.

Theoretical Considerations

What explains the fact that under appropriate testing conditions, eyewitness confidence is such a reliable indicator of accuracy, but under other testing conditions it is not? We begin by discussing why confidence and accuracy ought to be related in the first place. Then, we discuss the ways in which non-pristine testing conditions manage to confound this relation.

The signal-detection-theory account of a strong accuracy-confidence relation

In a fair lineup administered in double-blind fashion, it will usually be the case that the only face in the lineup that will generate a strong memory-match signal is the face of the perpetrator (i.e., the face that created the memory trace in the first place). Except in rare cases of chance resemblance between an innocent lineup member and the perpetrator, no other face in the lineup should generate a strong memory-match signal because these other faces were not the source of the witness's memory. This is true whether the operative memory signal is the absolute strength of the match between the memory of the perpetrator and a single face in the lineup (without regard for the other faces in the lineup) or is instead the relative strength of that match compared to the match generated by the other faces in the lineup. Either way, only a guilty perpetrator is likely to generate a strong memory signal.

Presumably, witnesses have learned through the course of daily life that a strong memory signal is an indicator of high recognition accuracy (and therefore warrants a high-confidence ID), whereas a weak memory signal is an indicator of low recognition accuracy (and therefore warrants either a low-confidence ID or a lineup rejection). Thus, under pristine testing conditions, simply relying on the strength of the absolute or relative memory signal ought to result in a strong confidence-accuracy relation (Mickes, Hwe, Wais, & Wixted, 2011). These ideas can be formalized in terms of a simple signal-detection model (Fig. 11), which has long been used to conceptualize the strong confidence-accuracy relationship observed in list-memory tasks used by basic memory researchers. The model in Figure 11 is usually applied to word-list memory tasks, but the basic concepts also apply to decisions made from a lineup. A version of the model applied to lineups would be somewhat more complicated, but its basic predictions about the confidence-accuracy relationship would remain unchanged. Thus, for the sake of simplicity, we use the standard (list memory) version of the model to illustrate what it predicts about the confidence-accuracy relationship.

In the context of eyewitness memory, signal-detection theory specifies how face-memory strength is distributed across guilty suspects (targets) and innocent suspects and fillers (lures) in a fair lineup. As depicted in Figure 11, the mean and standard deviation of the target distribution are both greater than the corresponding values for the lure distribution (a common but not necessary assumption). The model assumes that a decision criterion is placed somewhere on the memory-strength axis, such that a positive identification is made if the memory strength of a face (target or lure) exceeds it. Each level of confidence is associated with its own decision criterion. The overall correct-ID rate is represented by the proportion of the target distribution that falls to the right of the leftmost decision criterion, and the overall false-ID rate is represented by the proportion of the lure distribution that falls to the right of the leftmost decision criterion. Our concern here is not with the overall correct- and false-ID rates but is instead with the frequency of confidence-specific correct and false IDs. As illustrated in Figure 11, high-confidence IDs occur when a face generates a strong memory signal, one that exceeds the rightmost decision criterion. For the specific example shown in that figure, high-confidence IDs will often occur for target faces (37% of target-present trials result in a correct high-confidence ID) but will rarely occur for non-target faces (only 2% of target-absent trials result in an incorrect high-confidence ID). In other words, high-confidence IDs will be highly accurate. By contrast, weaker memory signals that surpass only the leftmost criterion for making an ID with low confidence are almost as likely to be incorrect as correct (13% of target-present trials result in a correct low-confidence ID; 9% of target-absent trials result in an incorrect low-confidence ID). Thus, low-confidence IDs will be inaccurate according to this account. Although this is just one specific example, it illustrates why it has long been understood that a strong confidence-accuracy relationship is an inherent feature of signal-detection theory.

How Non-Pristine Testing Conditions Harm the Confidence-Accuracy Relation

Although signal-detection theory's prediction of a good confidence-accuracy relation is well founded, it tends to be based on an assumption that the only source of information for confidence is the strength of the memory signal. And, in a typical memory experiment, signal strength is the only available informational cue on which to base one's confidence. But eyewitness confidence in an

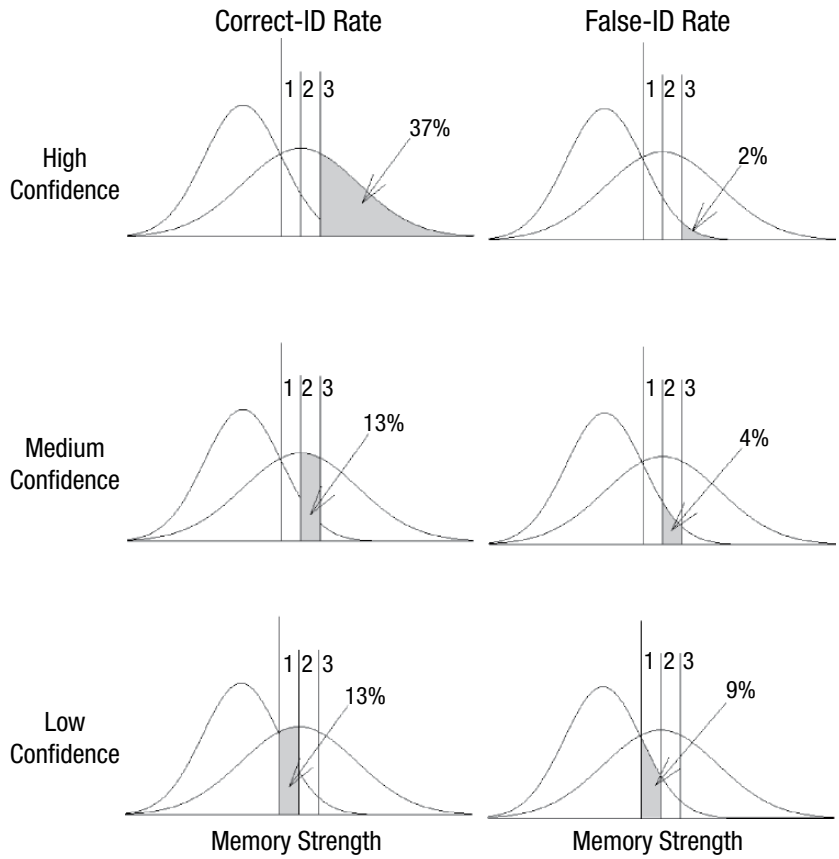

Fig. 11. Signal-detection-based interpretation of correct-ID rates (left panels) and false-ID rates (right panels) for high-confidence (top), medium-confidence (middle), and low-confidence (bottom) IDs. In each panel, the lure (innocent suspect) distribution is the narrow distribution on the left and the target (guilty suspect) distribution is the wider distribution on the right. Confidence criteria are shown as vertical lines, with the tallest vertical line representing the criterion for making an ID. The numbers 1, 2, and 3 represent low, medium, and high confidence, respectively.

identification is, in effect, the eyewitness's belief about the chances that the person he or she has identified is the perpetrator. And, as with other beliefs, eyewitnesses probably use whatever informational cues they have available to them when they state their confidence (Smalarz & Wells, 2015). If the only informational cue the eyewitness has at the time of making a confidence statement is a sense of the strength of the signal, then we would expect a good relation between accuracy and confidence because signal strength should be closely related to whether the target is the perpetrator or not. If, on the other hand, the eyewitness makes an identification and then overhears some seemingly confirmatory comment before making a confidence statement (e.g., "Your co-witness identified the same person"), then this confirmatory information is likely to be an additional cue driving his or her belief about the chances that the person identified is the perpetrator. In that case, the confidence of the witness is not based purely on the strength of the memory signal. If the confidence statement is based on considerations other than signal strength, then signal-detection theory's prediction of a confidence-accuracy relation no longer holds.

In our account, the requirement of pristine testing conditions applies not only to the composition of the lineup but also to the confidence statement, which should be assessed by the lineup administrator at the time of the initial identification (ideally, by a double-blind administrator whose behavior would not be influenced by knowledge of who the suspect is) before any other events can contaminate the confidence judgment. Consider, for example, the problem of assessing the confidence of an eyewitness who has been asked repeatedly to identify the same person (e.g., at the lineup, at a pretrial hearing, at trial). In such cases, the signal strength is likely to feel stronger to the eyewitness each time he or she encounters the person. Of course, the increase in signal strength is the result of repeated presentations of the suspect rather than the strength of the initial memory.

If, however, the witness fails to appreciate the effect of the intervening exposures on memory strength and relies on the (usually diagnostic) strong memory signal during subsequent tests, an error-prone high-confidence ID will be made. In the signal-detection model illustrated in Figure 11, this situation would be conceptualized as both distributions shifting to the right (as memory strength increases with repeated presentations) with the confidence criteria remaining fixed in place. In that case, high-confidence accuracy would plummet, because a much larger percentage of the lure distribution would now exceed the rightmost high-confidence criterion. Although a higher percentage of the target distribution would also now exceed the rightmost high-confidence criterion, the proportionate increase in false IDs would exceed the proportionate increase in correct IDs, so high-confidence accuracy would decrease. In effect, a source-monitoring failure will result in the witness relying on an internal memory cue—namely, strong memory—that is ordinarily diagnostic but no longer is (D. S. Lindsay, 2014; Roediger & DeSoto, 2015).

Why does a lineup that is composed of weak fillers (an unfair lineup) undermine our ability to infer high accuracy from high confidence? There are likely several reasons. First, one should not overlook the simple fact that unfair lineups increase the rate of mistaken identifications of innocent suspects at all levels of confidence. In a perfectly fair six-person lineup, for example, the maximum possible rate of mistaken identifications of an innocent suspect is 16.7%. And that maximum rate assumes that witnesses are performing at chance, that the perpetrator is never present in the lineup, and that all witnesses make an identification. But if just half of those witnesses do not make an identification and the perpetrator is in the lineup half of the time, the rate of mistaken identifications of innocent suspects from fair lineups would be less than 5% for low-confidence eyewitnesses and near floor for high-confidence eyewitnesses. An unfair lineup, in contrast, runs a much higher overall rate of mistaken identifications of the innocent suspect. This higher rate of mistaken suspect identifications from unfair lineups means that some are likely to end up in the high-confidence category.

In addition to raising the overall level of mistaken identifications of innocent suspects, there is also some evidence that unfair lineups can increase the confidence with which eyewitnesses make a mistaken identification. For example, as noted earlier, Charman et al. (2011) found that including highly dissimilar "dud" lineup members inflated witnesses' confidence in their mistaken identification of a non-dud. In a more recent study, Horry and Brewer (2016) manipulated the similarity between the suspect and the fillers in four-person simultaneous lineups and found that confidence judgments for positive identifications were predicted by the balance of evidence between the chosen item and the unchosen alternatives. In other words, as target-filler similarity decreased, confidence increased. This suggests that simultaneous lineup decisions are based at least in part on a relative memory-strength signal, which may be the reason why unfair lineups are so problematic. In an unfair lineup, the suspect (innocent or guilty) will generate a strong memory-match signal relative to those generated by the other lineup members (in Fig. 11, this would be conceptualized as both distributions being shifted to the right with the confidence criteria remaining fixed). The result would be a bias to choose that individual (Wells, 1984), even when making a high-confidence ID. As a bias to choose with high confidence increases, accuracy decreases. All of these problems are avoided (or at least minimized) if fair lineups are used.

An alternative but related theoretical interpretation is provided by fuzzy-trace theory's distinction between verbatim and gist memory traces (Brainerd & Reyna, 2005). According to fuzzy-trace theory, witnesses store both verbatim traces of the perpetrator plus more general (gist) traces of conceptually related information. Applied to eyewitness identification, the verbatim trace would be the perceptual representation of the perpetrator's face, whereas the gist trace might correspond to the general description of the perpetrator (e.g., an approximately 20-year-old White male with short dark hair and a scruffy beard). Depending on how retrieval is tested, a witness will rely on either the verbatim trace or the gist trace. When a participant relies on a verbatim trace, a strong memory-match signal (and attendant high confidence) will occur only when a face in the lineup matches that trace. As a general rule, such a match will occur only when the actual perpetrator is in the lineup. Thus, high-confidence accuracy will be high. The use of a pristine lineup seems well suited to promoting the retrieval of a verbatim trace because everyone in a fair lineup matches the gist (so the gist trace is of no help). However, in an unfair lineup, only the suspect corresponds to the gist trace, thereby promoting reliance on the gist trace instead of the verbatim trace. As noted by Brainerd and V. F. Reyna (2002):

> Retrieval of gist traces usually supports a more generic form of remembering, sometimes called familiarity, in which nonexperienced items are perceived to resemble experienced items but their occurrence is not explicitly recalled. However, when gist traces are especially strong, they can support high levels of phantom recollective experience for certain types of nonexperienced items—namely, items that are good cues for the gist of experience. (p. 166)

In other words, an unfair lineup might at times promote strong phantom recollection, leading to high-confidence errors. Accordingly, fuzzy-trace theory provides an additional theoretical rationale for recommending that the police use fair lineups.

General Conclusions

Our review of research concerned with the confidence-accuracy relationship in eyewitness identification is the first since Sporer et al. (1995) reviewed the literature more than 20 years ago. They found that when the analysis was limited to choosers, the correlation between confidence and accuracy was considerably higher than it was previously thought to be. That was their main message, even though their article is cited surprisingly often as suggesting the opposite. Nevertheless, the measure they used to assess that relationship—the point-biserial correlation coefficient—does not directly address the question of most interest to judges and juries. The point-biserial correlation coefficient is a perfectly reasonable effect-size statistic for a comparison between the average level of confidence associated with correct IDs versus the average level of confidence associated with incorrect IDs. However, the question asked by judges and juries concerns the average accuracy associated with suspect IDs made with a particular level of confidence. The correlation coefficient does not directly provide that information, but a calibration plot comes closer to doing so (Juslin et al., 1996). A calibration plot displays the proportion of correct IDs for choosers (or non-choosers) as a function of the level of confidence expressed, with confidence measured using a 100-point scale.

Calibration studies have consistently shown that for choosers, the confidence-accuracy relationship is strong (e.g., Brewer & Wells, 2006). The relationship is strong in the straightforward sense that high-confidence accuracy is much higher than low-confidence accuracy. Still, most calibration studies have found that highly confident witnesses are overconfident, and in one sense they are. Although CAC analysis treats only innocent-suspect IDs as relevant errors, from the eyewitness's point of view, filler IDs and innocent-suspect IDs are both relevant errors. Thus, when witnesses are asked to provide a confidence rating (e.g., 90%) that is commensurate with their accuracy (e.g., 90% of their IDs are guilty-suspect IDs, whereas 10% of their IDs are filler IDs or innocent-suspect IDs), their actual accuracy (e.g., 80% correct) can be said to reflect overconfidence. However, judges and juries in a case involving eyewitness-identification evidence are not interested in using an eyewitness's confidence to help them decide whether the witness picked a filler. Judges and juries already know that this particular witness did not pick a filler—the witness picked the suspect. Hence, judges and juries want to know how likely it is that the suspect is the perpetrator given that the witness identified the suspect with a particular level of confidence. The answer to their question, therefore, is provided by an analysis of the accuracy of suspect IDs per se without consideration of filler IDs. Of the suspect IDs that are made with a particular level of confidence, what proportion of those IDs were of guilty suspects and what proportion were instead of innocent suspects?

The answer to that key question is provided by CAC analysis, which is a measure of suspect-ID accuracy at each level of confidence for the base rate of target-present lineups used in the study (usually 50%). A more complete picture is provided by a Bayesian analysis that indicates what suspect-ID accuracy would be for the full range of possible base rates (0%–100%). Analyses of suspect-ID accuracy show that for a wide range of base rates, high confidence implies high accuracy (with no sign that witnesses are overconfident) and low confidence implies much lower accuracy. This is true of both lab studies and police department field studies, so long as pristine testing conditions are used. However, when the base rate is low enough (e.g., less than 25% of the lineups contain a guilty suspect), accuracy starts to become compromised across the board (even for high-confidence IDs). That fact provides a rationale for treating the base rate of guilty suspects as a system variable and for taking steps to ensure that the base rate is not unreasonably low. One way to do so is to require some objective evidence of guilt before placing a suspect in a lineup (Wells et al., 2015).

Importantly, a low-confidence ID on an initial test of memory from a lineup signals low accuracy whether or not pristine testing procedures are used. For this reason, low confidence should never be ignored and should instead always raise red flags about the reliability of the ID (Wixted et al., 2015). Although low-confidence IDs have some probative value when pristine procedures are used, under non-pristine testing conditions, they are even more error prone. As noted earlier, the majority of DNA exoneration cases in which eyewitness misidentification played a significant role were associated with, at best, a low-confidence ID on the initial memory test (Garrett, 2011). In some cases, the witness initially made a non-ID (i.e., confidence was so low that the witness identified no one) or a filler was identified. Thus, a low-confidence initial ID of a suspect from a lineup (or worse) corresponds to an uncomfortably high probability that the suspect is innocent. Had this simple fact been better understood by the legal system, many of the innocent defendants who were convicted based in part on a high-confidence ID that occurred in court may never have been convicted in the first place.

The news about the unreliability of a low-confidence initial ID will come as no surprise to most readers. Presumably, most readers are already under the impression that eyewitness memory is inherently unreliable, such that a suspect ID is error prone even under the best of conditions and even when confidence is high. Thus, the main news we have to offer is that eyewitness memory is not inherently unreliable. Under pristine testing conditions, a high-confidence suspect ID appears to be highly probative of guilt. Ignoring that fact—as the legal system is increasingly inclined to do—only serves to inappropriately exonerate the guilty. At the same time, ignoring low confidence at the time of an initial ID inappropriately imperils the innocent. The take-home message is that initial eyewitness confidence obtained from a pristine eyewitness-identification procedure serves both of the fundamental goals of the criminal justice system: to clear the innocent and to convict the guilty. By contrast, any later expression of confidence (including the confidence expressed by the eyewitness at trial in front of a jury) should be ignored, because doing otherwise works against the cause of justice.

Filler IDs and non-IDs are probative of innocence

Just as suspect-ID accuracy provides the information of interest to judges and juries tasked with evaluating the reliability of an eyewitness who has identified a suspect, analyses performed separately on filler IDs provide the information of interest to judges and juries tasked with evaluating the implications of the fact that an eyewitness picked a filler from a lineup instead of a suspect. Such an eyewitness would not testify against the defendant (because the eyewitness did not identify the defendant), but the fact that a filler ID occurred at an earlier stage of investigation nevertheless provides relevant information. The fact that a filler ID was made is somewhat probative of innocence. In other words, when filler IDs are examined separately, the data suggest that, given that a filler ID occurred, it is somewhat more likely that the lineup contained an innocent suspect than a guilty suspect.

In other cases, the eyewitness may have made a non-ID at the outset of the investigation. In a case like that, judges and jurors would be interested in the information value of a non-ID, and that information is provided by separately performed analyses of lab data for eyewitnesses who made non-IDs from target-present and target-absent lineups. When such an analysis is performed, the data indicate that non-IDs are also probative of innocence. The key point is that whether a case involves a suspect ID (the kind of ID that has helped to send innocent people to prison), a filler ID, or a non-ID, the information value of the ID in question is provided by analyzing the data separately, not by combining the data across suspect IDs, filler IDs, and non-IDs or by combining the data for choosers (suspect IDs and filler IDs) and analyzing them separately from data for non-choosers (non-IDs).

One of the relevant situations in which good records of rejections and filler IDs is important is in multiple-witness cases. Suppose, for example, that one witness identified the suspect and the other two rejected the lineup. What does that mean? Clark and Wells (2008) analyzed a large number of lab studies to estimate the probability that the suspect was the perpetrator under various combinations of suspect-ID, filler-ID, and lineup-rejection decisions in multiple-witness cases. In most cases, if one witness identified the suspect and the other two either rejected the lineup or picked a filler, the overall evidence pointed toward innocence rather than guilt of the suspect. Going forward, it will be important to address questions like this, taking into account IDs made with various levels of confidence (e.g., one high-confidence suspect ID and two low-confidence filler IDs).

Clark and Wells's (2008) analysis of the multiple-witness situation made it clear that one cannot ignore the witnesses who failed to pick the suspect. Nevertheless, in a 2012 national survey of U.S. law enforcement agencies, 37% of the agencies reported that they do not even write a report making a record of a lineup if the witness did not identify the suspect in the case (Police Executive Research Forum, 2013). Following on the lineups-as-experiments analogy described earlier in this article, this is akin to an experimenter ignoring data that are inconsistent with the hypothesis. Wells et al. (2015) argued that a failure to make a clear record of non-IDs and filler IDs could be construed as a "Brady violation"—that is, the violation of a constitutional requirement that the state reveal to the defense any evidence that might favor the defense (*Brady v. Maryland*, 1983).

Non-pristine testing conditions

How informative is confidence in a suspect ID that was made under non-pristine testing conditions? This is an important question to consider because, in the real world, pristine testing conditions will not always be achieved. Scientific research has clearly established that certain non-pristine testing conditions severely compromise the information value of eyewitness confidence. We consider them here.

Initial versus later confidence. Expressions of confidence by the eyewitness beyond the confidence statement at the initial identification are potentially problematic because a variety of factors (e.g., post-ID feedback) can inflate confidence without increasing accuracy. Thus, only an initial confidence statement—one that is made

before there is much opportunity for confidence contamination to occur—provides reliable information. That fact underscores the importance of a recommendation long made by eyewitness-identification researchers and recently reiterated by the National Academy of Sciences committee: The initial confidence statement made by an eyewitness should be recorded and preserved. In this regard, another recommendation by the National Academy of Sciences committee—to videotape the witness-identification process—takes on special importance. Juries typically see an eyewitness make a high-confidence ID only in the courtroom, and they are heavily influenced by it. This is unfortunate because only the first ID, which occurred back at the beginning of the police investigation, provides diagnostic information about the reliability of the ID. With regard to its influence on jury decision making, an abstract discussion of the fact that confidence was low during an initial ID may have a hard time competing with the live expression of high confidence that occurs in the courtroom. However, if the initial lineup procedure were video recorded, jurors would have direct evidence that the eyewitnesses' initial level of confidence was low—evidence that would likely help them to understand that the ID is unreliable no matter what the witness now says.

Until relatively recently, video recording of all identification procedures was not practical for some jurisdictions because of the financial costs and video storage difficulties involved. Today, however, that is no longer true. Nevertheless, there are likely to be some cases in which witness cooperation is an issue. For example, if a witness who is critical to a case fears being video recorded (e.g., out of concern that the recording will end up on the Internet, where gang members will see who identified their comrade), then proceeding with the identification procedure without video recording it (perhaps instead audio taping it) might be advisable. Still, where possible (presumably in the large majority of cases), video recording the session will go a long way toward ensuring the integrity of the identification procedure and providing the jury with the information it needs about eyewitness confidence.

Having reliable information about the confidence of the eyewitness at the initial identification allows the defense to learn about and explain to the jury that confidence inflation has occurred. Some lab-based evidence has shown that, as one would hope, upon learning that a witness who was highly confident at trial was actually not confident at the time of the initial identification, mock jurors discounted their ratings of witness accuracy and the defendant's probability of guilt (Bradfield & McQuiston, 2004). On the other hand, Jones, Williams, and Brewer (2008) found that an "explanation" from the witness (e.g., "I was nervous at the time but now I am confident") led mock jurors to discount the low initial confidence of the witness and be more influenced by his or her later confidence. Indeed, we have concerns about how these problems would play out in pre-court and court proceedings to the extent that witnesses who were initially not confident would find some reason to explain away their initial lack of confidence and lead the court to rely on the inflated confidence that they had developed. One solution might be to adopt a hard-and-fast judicial rule stating that only the initial confidence of an eyewitness, made in good faith, is permissible in court. Another solution might be to adopt jury instructions stating that only confidence in an initial, good-faith attempt at an identification provides valid information about its reliability.

Fair versus unfair lineups. Another non-pristine testing condition that clearly compromises the information value of eyewitness confidence is an unfair lineup. Study after study has shown that if the innocent suspect in the lineup resembles the perpetrator to a greater extent than the fillers do (e.g., if the innocent suspect matches the description of the perpetrator more than the fillers do), high-confidence suspect-ID accuracy is greatly reduced (as illustrated earlier in Fig. 6). The importance of this observation is hard to overstate. If an unfair lineup is used, then the take-home message in this article does not apply. Mistakenly assuming that a high-confidence initial ID is highly accurate even when an unfair lineup is used is a recipe for wrongfully convicting the innocent.

Blind versus non-blind lineups. The blind lineup-administration procedure logically eliminates a potential source of error because the lineup administrator cannot possibly—intentionally or otherwise—steer the witness to the suspect in the lineup or provide post-ID praise to the witness for "getting it right" (thereby inflating even the initial statement of confidence). After an identification, even statements from a lineup administrator such as "you have been a really great witness" inflate the confidence of witnesses who have made a mistaken identification, but such statements do not inflate confidence if the witness knows that the lineup administrator is blind as to which lineup member is the suspect and which are fillers (Dysart, Lawson, & Rainey, 2012). In addition, there is evidence that lineup administrators influence witness confidence even when the administrators are given an unbiased script that they are supposed to follow (Garrioch & Brimacombe, 2001). Furthermore, lab data have shown that when people are assigned to the role of a lineup administrator, they tend to not make records of filler IDs when they know which lineup member is the suspect (non-blind lineup administrators), but they

faithfully make such records when they do not know the status of the identified lineup member (blind lineup administrators; see Rodriguez & Berry, 2014). These considerations explain why blind lineup administration has long been recommended by eyewitness-identification researchers and why that recommendation was also recently endorsed by the National Academy of Sciences committee.

The point is that (a) confidence is a reliable indicator of accuracy under pristine testing conditions; (b) confidence is a much less reliable indicator of accuracy under certain non-pristine testing conditions (e.g., when an unfair lineup is used or when the test is not the initial ID test); and (c) eyewitness expressions of confidence can be influenced by non-blind lineup administrators, which is an undesirable outcome no matter what its effect on accuracy might be. Obviously, confidence may or may not be a reliable indicator of accuracy under other conditions that have not yet been subjected to scientific investigation. Later, we recommend some research priorities for further investigating the eyewitness confidence-accuracy relationship.

Estimator variables and confidence in a suspect ID

In the studies reviewed here, eyewitnesses who were tested using pristine procedures appropriately adjusted their confidence downward when they were aware that no one in the lineup strongly matched their memory of the perpetrator. This is just another way of saying that there is a strong relationship between confidence and accuracy. That finding may have some non-obvious but nevertheless important implications for how people generally think about the effect of various estimator variables on eyewitness-identification accuracy. Consider, for example, how juror guidelines in Massachusetts instruct juries to think about estimator variables. Those instructions list a variety of factors than can make memory worse, on average (e.g., long retention interval, short exposure time, stress, the presence of a weapon), and they invite jurors to believe that if one or more of those factors is present, then the reliability of the ID should be regarded as less trustworthy than it otherwise would be. As intuitively appealing as this line of thinking might be, the evidence suggests that it may not be valid.

To illustrate this point, we consider the fact that a long retention interval typically results in worse overall memory performance compared to a short retention interval. Does that fact imply that a high-confidence initial ID of a suspect made after a long retention interval is less trustworthy than a high-confidence initial ID of a suspect made after a short retention interval? Not necessarily. Eyewitnesses have a sense of how well each lineup member matches their memory, and if the memory is weak, they are not likely to have high confidence. That is, as memory fades with the passage of time, eyewitnesses will be less likely to experience a strong memory-match signal when viewing the members of a photo lineup. As a result, witnesses might make more errors but, critically, those errors are likely to be associated with low confidence (because high-confidence IDs are typically made when the memory-match signal is strong, not when it is weak, as it generally would be following a long retention interval). Nevertheless, for the smaller percentage of eyewitnesses who do make a high-confidence ID despite a long retention interval, their average accuracy could be every bit as high as that for the larger percentage of eyewitnesses who make a high-confidence ID following a short retention interval.

Although additional research is certainly needed, the available evidence indicates that eyewitnesses may often appropriately adjust confidence to the prevailing memory conditions, contrary to Deffenbacher's optimality hypothesis (Deffenbacher, 1980). Palmer, Brewer, Weber, and Nagesh (2013, Experiment 1) compared immediate versus 1-week-delayed performance in a large-scale experimentally controlled field study. Not surprisingly, they reported that overall accuracy was lower following the 1-week retention interval than on the immediate test, but as shown in Figure 4l, the accuracy of high-confidence IDs was equally high either way. The same was true when overall memory strength was manipulated by varying exposure duration from 5 seconds to 90 seconds (also shown in Fig. 4l) or by varying whether or not attention was distracted during exposure (Fig. 4m). In each case, overall memory performance was weaker in one condition compared to the other, but high-confidence accuracy was the same either way. With regard to a retention-interval manipulation, Juslin et al. (1996), Read et al. (1998), and Sauer, Brewer, Zweck, and Weber (2010) all reported a similar outcome (Figs. 4h, 4n, and 4p, respectively). Note that the Read et al. (1998) results are noteworthy in that those authors used retention intervals as long as 9 months.

Similar effects are evident for several other estimator variables. For example, Carlson and Carlson (2014) and Carlson, Dias, Weatherford, and Carlson (in press) found that although the presence of a weapon clearly led to worse memory performance overall (the weapon-focus effect), it had virtually no effect on the accuracy of identifications made with high confidence (Figs. 4c and 4d). The same outcome was observed by Dodson and Dobolyi (2016) for same-race versus cross-race IDs (Fig. 4f). Cross-race IDs were associated with significantly lower recognition memory performance compared to same-race IDs, but high-confidence IDs were highly (and similarly) accurate either way.

If these results generalize to the real world, they suggest that these estimator variables may not be particularly relevant to the reliability of an initial ID made with high confidence. Although definitive conclusions cannot yet be drawn, the overall pattern of results suggests that under pristine testing conditions, estimator variables that have long been thought to compromise the reliability of a suspect ID may not do so (because eyewitnesses appropriately adjust their confidence under poorer estimator-variable conditions). Still, it would be premature to make a definitive statement regarding the effect of different estimator variables on the accuracy of IDs made with high confidence because the issue has only recently been addressed using CAC analysis. In addition, a study by Lampinen, Erickson, Moore, and Hittson (2014) investigated the effect of distance on identification accuracy. This study used an old/new recognition procedure (not a lineup) in which each witness made 16 recognition decisions. Thus, its design was far removed from the kind of forensically relevant lineup designs that we have considered here. Nevertheless, it is worth noting that according to our estimates based on the ROC data presented in their Figure 4, high-confidence accuracy was always below 90% and became noticeably worse as distance increased, falling to approximately 70% correct at the longest distances tested. Whether the same would be true for lineups is unknown, but this result underscores the fact that more work is needed to determine the effect of estimator variables on high-confidence accuracy.

Mistaken-ID rates at the level of the lineup versus the courtroom: The plea effect

At this point it is important to note that we cannot necessarily assume that the chances that a high-confidence ID is mistaken at the level of the lineup are the same as the chances that a high-confidence ID is mistaken at the level of a trial. One reason, although not the only reason, is that guilty pleas (which do not go to trial) will remove many more guilty than innocent people from trials. This *plea effect*, originally described by Wells, Memon, and Penrod (2006), yields a distribution of innocent and guilty individuals at trial that is different from the distribution at the level of the lineup.

Let us assume that witnesses who were tested using pristine procedures (fair lineup, double-blind administrator, confidence measured at time of ID, etc.) and were 95% to 100% confident have a 98% chance of being accurate. In other words, only 2% of these witnesses would be mistaken. Suppose now that we have a defendant on trial who was identified by an eyewitness who made the identification under pristine testing conditions and was 95% to 100% confident. Can we assume, in the absence of any other evidence, that at the trial level there is only about a 2% chance that the person the witness identified is an innocent person? The answer is "not necessarily," especially in the U.S. legal system. Depending on its size, the plea effect could create a situation in which the chance that the defendant is innocent is much higher than 2%.

The plea effect (Charman & Wells, 2007; Wells et al., 2006) refers to the fact that most criminal convictions never involve a trial at all but instead are obtained through guilty pleas. In fact, over 95% of criminal convictions in the United States are attained through plea deals and are never brought to trial (Ross, 2006). Because fewer than 5% of felony convictions come from people who claim innocence and choose to take their case to trial, those who do so represent a small subset of defendants. And, although innocent people sometimes plead guilty (e.g., over 20% of the DNA exoneration cases involved an innocent person who pled guilty), it seems reasonable to assume that the chances that an innocent person would take a case to trial rather than plead guilty is much greater than the chances that a guilty person would take a case to trial.

Consider 10,000 suspect IDs made with high confidence. For the sake of simplicity, let's assume that all 2% of those who were mistakenly identified with high confidence ($10{,}000 \times 0.02 = 200$ innocent suspects) are prosecuted and take their case to trial (after all, they are innocent). And, let's assume that of the 98% who were accurately identified with high confidence ($10{,}000 \times 0.98 = 9{,}800$ guilty suspects), 97% ($9{,}800 \times 0.97 = 9{,}506$) take a plea and 3% ($9{,}800 \times 0.03 = 294$) instead go to trial. If this were the case, and if jury trials always resulted in guilty verdicts, then $100\% \times 9{,}506 / (9{,}506 + 200) = 95.1\%$ of guilty verdicts would arise through plea deals. Moreover, among those who took their case to trial (200 innocent suspects and 294 guilty suspects), the chances of the defendant being guilty based on the eyewitness-identification evidence alone would be slightly less than $100\% \times 294 / (294 + 200) \approx 60\%$. In other words, what is a mere 2% mistaken-identification rate at the level of the lineup becomes a 40% chance of innocence among cases that make it to trial. That reduction in accuracy at trial is, of course, offset by an increased level of accuracy associated with high-confidence IDs that ended in a plea deal instead of going to trial. In this example, because all of the innocent suspects went to trial, 100% of the defendants who were identified with high confidence and who accepted a plea bargain would be guilty.

Obviously, these numbers will change depending on the assumptions that are made. For example, instead of being equally likely to be forwarded for prosecution (as assumed in the example above), guilty suspects may be

more likely than innocent suspects to be forwarded for prosecution. This might occur because guilty suspects are more likely to have independent corroborating evidence against them compared to innocent suspects. In addition, the 95% plea rate, which is based on all cases, may be an overestimate for eyewitness-identification cases because defense attorneys might believe that they have a better chance of acquittal in cases involving eyewitness-identification evidence than in many other types of cases. If we assume that suspects who have been identified with high confidence are twice as likely to be forwarded for prosecution if they are guilty than if they are innocent (because of a disparity in corroborating evidence), that 25% of guilty suspects choose jury trials (in hopes of discrediting eyewitness evidence), and that 50% of jury trials end in guilty verdicts, then 85% of all guilty verdicts in cases involving eyewitness identification would arise from plea bargains, and high-confidence ID accuracy at trial would be 96% correct.

Although the precise numbers cannot be known, it is important to appreciate that the plea effect changes the ratio of the innocent to the guilty among those who actually go to trial. The more the plea effect increases the ratio of the innocent to the guilty at trial, the less trustworthy a high-confidence ID becomes at trial (and the more trustworthy a high-confidence ID becomes for those who choose to accept a plea bargain).

The distinction between eyewitness-identification accuracy at the level of the lineup and eyewitness-identification accuracy at the level of cases that go to trial is important. An eyewitness expert giving trial testimony, for example, should be careful to not equate the mistaken-identification rate at the level of the lineup with the chances that the defendant is guilty in a particular case that made it to trial. A similar caution applies to the base-rate issue discussed previously (i.e., a high-confidence accuracy score estimated from a study that used a 50% target-present base rate does not directly apply to a jurisdiction that might have a much lower base rate). At the same time, these considerations do not undermine the general conclusion of the current article, namely that high-confidence eyewitness identifications made using pristine testing procedures have a very low rate of error.

Priorities for future research

How to collect a confidence statement from an eyewitness. Although confidence in an initial ID is highly predictive of accuracy, no police department field study has specifically investigated different methods for recording initial confidence. Should a confidence statement be taken in the witness's own words (as in Klobuchar et al., 2006), or should confidence be recorded using an explicit 3-point rating scale (as in Wixted et al., 2016)—or should a 100-point scale be used? Given the clear information value of initial confidence, this issue seems important to pursue.

How to create a fair lineup. Unfair lineups seriously degrade the information value of eyewitness confidence. One way to minimize the chances of creating an unfair lineup is to ensure that every member of the lineup matches the description of the perpetrator provided by the witness. However, this is a subjective process, and even an investigator who is trying to follow that directive might unintentionally create an unfair lineup. Indeed, in one condition of a recent police department field study (the blinded condition in Wixted et al., 2016), the lineups assessed by mock witnesses were found to be unfair in that the suspect in the lineup was selected, on average, more than the fillers based solely on the description. But even when care is exercised to make sure that all fillers match the description of the perpetrator that was provided by the witness, the lineup might not be fair. This is because eyewitnesses' verbal descriptions of perpetrators are often vague or incomplete, and sometimes the description does not even match the suspect (Luus & Wells, 1991). Some have proposed that the fillers should be matched to the suspect on major physical characteristics rather than just those contained in the eyewitness's description of the perpetrator (e.g., Lindsay, Martin, & Webber, 1994). Others have proposed that the fillers be selected based on their overall similarity to the suspect (Clark & Tunnicliff, 2001). Some have found that it is possible to make the fillers too similar to the suspect (which protects innocent suspects but reduces the chances of identifying perps; see Wells et al., 1993). And, as discussed earlier in this article, when someone becomes a suspect based on similarity to a composite or a surveillance image, simply matching fillers to the eyewitness's verbal description of the suspect is not sufficient. Clearly, the general idea that poor lineup fillers place innocent suspects at risk and confound our ability to rely on confidence is not in question, and we see evidence of this in the CACs shown in Figure 6. 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?

The effect of estimator variables on confidence. An important goal for future research will be to determine if the conclusions discussed above with respect to estimator variables apply to other estimator variables that are relevant to eyewitness IDs in the real world (e.g., high stress vs. low stress). The fact that estimator variables have an effect on overall memory accuracy is beyond

dispute; what remains unknown is what effect they have on the confidence-accuracy relationship when the data are subjected to CAC analysis. This is an important issue to specifically investigate because variables that impair overall memory accuracy do not necessarily have any effect on the accuracy of suspect IDs made with high confidence (instead, they may affect only the frequency of high-confidence suspect IDs).

Exploring other ways of sorting between guilty and innocent suspects. The standard approach to assessing eyewitness-identification confidence is to ask the eyewitness how confident she or he is in the identification that was made. But research by Sauer, Brewer, and Weber (2008) found that collecting a witness confidence statement for each lineup member (rather than only the one who was chosen) provided a more informative indicator of recognition. Following on this finding, more recent research has shown promising results for procedures in which eyewitnesses do not pick someone out of a lineup at all but instead make a confidence judgment about whether each lineup member is the perpetrator (e.g., Brewer, Weber, Wootton, & Lindsay, 2012; Sauer, Brewer, & Weber, 2008; Sauer, Brewer, & Weber, 2012) or rate how well each face matches their memory of the perpetrator (Sauer, Weber, & Brewer, 2012). Results from profile analyses and classification algorithms have shown that such methods may be superior to the traditional eyewitness-identification task. Other work has examined decision time and shown that eyewitnesses make accurate identifications consistently faster than they make mistaken identifications (e.g., Dunning & Perretta, 2002; Sauer, Brewer, & Wells, 2008; Sporer, 1993). Our point here is simply that we do not want to close off the possibility that there might be other approaches to assessing the probability of a suspect's guilt that work even better than traditional methods.

Conclusion

According to the available data, the relationship between confidence and accuracy for an initial ID from an appropriately administered lineup is sufficiently impressive that it calls into question the very notion that eyewitness memory is generally unreliable. Eyewitness memory can certainly *become* unreliable as a result of influences introduced by the legal system (feedback, repeated exposure to the suspect, misinformation, biased lineup composition, etc.), but the same is true of any kind of evidence, including DNA evidence. A contaminated eyewitness memory test, like a contaminated DNA test, is not reliable. However, the available research suggests that when pristine testing procedures are used, an initial ID made with high confidence is highly indicative of accuracy. Perhaps even more importantly, an initial ID made with low confidence—whether testing conditions are pristine or not—is highly error prone. A better appreciation of that simple fact might have prevented most of the DNA exonerees from being convicted in the first place. Thus, instead of disregarding eyewitness confidence altogether, the legal system should draw a distinction between initial confidence that was obtained using pristine testing procedures and confidence obtained later or under conditions known to compromise the confidence-accuracy relationship.

Appendix A

An illustration of a strong confidence-accuracy relationship despite a low point-biserial correlation

Twenty years ago, Juslin, Olsson, and Winman (1996) explained that the point-biserial correlation coefficient is problematic for assessing the confidence-accuracy relationship because its value can be low even when eyewitnesses exhibit perfect calibration (such that 100% confidence implies 100% accuracy, 90% confidence implies 90% accuracy, etc.). However, they did not illustrate what the point-biserial correlation coefficient actually measures, nor did they reanalyze any of the prior data to show what those data look like when analyzed in a more appropriate way. This may account for why, to this day, scientists continue to rely on the point-biserial correlation coefficient to measure the relationship between confidence and accuracy and why the legal system does so as well. Here, we explain what this statistic actually measures and why it should no longer be used if the goal is to inform the legal system about the reliability of a suspect ID made with a particular level of confidence. Again, it is a perfectly valid statistic when used for other purposes (Rosnow, Rosenthal, & Rubin, 2000), and it does signal a strong relationship between confidence and accuracy when its value is high (D. S. Lindsay, Nilsen, & Read, 2000; D. S. Lindsay, Read, & Sharma, 1998). However, for the purpose of predicting eyewitness-identification accuracy from an eyewitness's expression of confidence, it can be misleading because it does not necessarily indicate a weak relationship between confidence and accuracy when its value is low (as has been assumed by researchers and the legal system alike).

Table A1 presents hypothetical data generated by 30 "choosers" who have made an ID from a lineup and rated confidence using a 5-point confidence scale (1 = *low confidence*, 5 = *high confidence*). Choosers make one of four possible decisions: correctly identifying a suspect from a target-present lineup, incorrectly identifying a suspect from a target-absent lineup, incorrectly identifying a filler from a target-present lineup, or incorrectly identifying a filler from a target-absent lineup. Thus, all of the

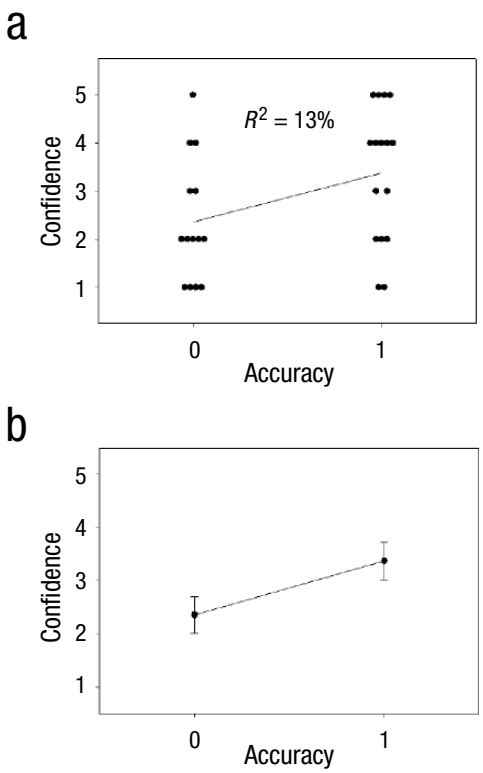

Fig. A1. Hypothetical data generated by 30 "choosers" who have made an ID from a lineup and rated confidence using a 5-point scale, with confidence plotted as a function of accuracy. In panel (a), each point is a data point from one participant, and the line is the line of best fit. In panel (b), the points represent the data averaged across confidence, and the bars represent standard errors.

accuracy scores of "1" in Table A1 correspond to suspect IDs from target-present lineups (which is the only correct response for a chooser). For simplicity, these hypothetical data are conceptualized as having come from a mock-crime study in which it is known whether the suspect in the lineup is innocent or guilty (e.g., as illustrated in Fig. 1). The data in Table A1 have been chosen to illustrate a point about the correlation between confidence and accuracy, not to reflect what typical data necessarily look like. The correlation between the 30 accuracy scores and the 30 corresponding confidence scores shown in the two rightmost columns of Table A1 is .36, which is slightly lower than the generally accepted value of .41 for choosers.

Figure A1a illustrates the fact that computing a point-biserial correlation coefficient is tantamount to fitting a straight line through the data when confidence is plotted as a function of accuracy coded in binary format (0 = inaccurate, 1 = accurate). Each point represents one participant, and the points for different participants that would fall atop one another have been slightly spread out on the accuracy dimension to show how many participants are associated with each confidence-accuracy pair. The best-fitting line is the one that minimizes the sum of the squared deviations (vertically) between the line and the 30 individual data points. It is difficult to imagine how judges and juries could extract useful information about the likely reliability of a particular suspect ID (e.g., one made with high confidence) from data analyzed in this manner.

Figure A1b shows the same data except that the confidence ratings have been averaged together to make a more interpretable graph. This figure clearly shows that the average level of confidence is higher for correct IDs than for incorrect IDs. In fact, this is how the data were plotted in Figure 1 of Sporer, Penrod, Read, and Cutler's (1995) seminal article. When the data are analyzed in this manner, the result is presumably more interpretable to judges and juries. However, a problem with Figure A1b is that it plots the unaveraged dependent measure (accuracy coded as 0 or 1) on the *x*-axis and the averaged predictor variable (confidence) on the *y*-axis. This would be the appropriate way to plot the data if you knew, for each eyewitness, whether his or her ID was correct or incorrect and wanted to estimate his or her likely level of confidence. If that were the question of interest, then the point-biserial correlation coefficient would be a reasonable effect-size statistic to help conceptualize the results of a *t* test (for example) comparing average confidence for correct decisions versus average confidence for incorrect decisions (Rosnow et al., 2000). Yet this is not the question of interest, because in actual practice, the situation is reversed: An eyewitness provides a confidence rating associated with an ID (this is the predictor variable, which is not averaged), and the legal system wants to make the best estimate as to the likely accuracy of that ID (this is the dependent variable, and it equals the average level of accuracy associated with each level of confidence that an eyewitness might express). This logic suggests, as Juslin, Olsson, and Winman (1996) pointed out, that plotting average accuracy (on the *y*-axis, as the dependent measure) versus different levels of confidence (on the *x*-axis, as the independent measure) is the sensible way of representing the data and addressing the question of interest. Only when plotted this way are the data presented in a manner that provides an answer to the critical question asked by the criminal justice system: Given that an eyewitness has a particular level of confidence in his or her ID, how accurate is that ID likely to be?

Figure A2a shows the same data plotted in Figure 2a except that the axes have been reversed to plot the independent variable (confidence) on the *x*-axis and the dependent variable (accuracy) on the *y*-axis. Obviously, because the information in Figures A1a and A2a is the same, the best-fitting line corresponds to the same point-biserial correlation coefficient (.36) as in Figure A1a. However, even with the variables appropriately reversed, the data do not yet provide much in the way of useful

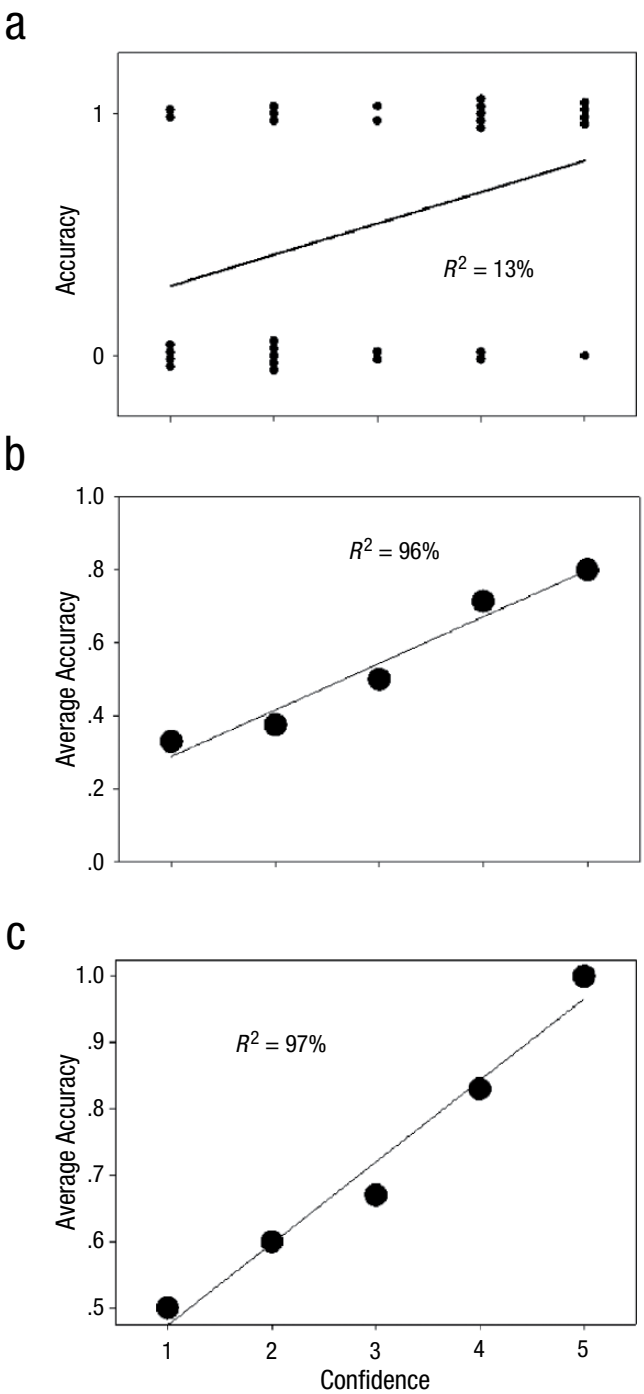

Fig. A2. Hypothetical data generated by 30 "choosers" who have made an ID from a lineup and rated confidence using a 5-point scale, with accuracy plotted as a function of confidence. In panel (a), each point is a data point from one participant, and the line is the line of best fit. In panel (b), the points represent the data averaged across accuracy. Panel (c) shows the confidence-accuracy characteristic curve for when only suspect IDs are included in the accuracy calculation.

information to courts of law. Figure A2b shows the same data as Figure A2a except that the binary accuracy scores associated with each level of confidence have been averaged together. Now the data are depicted in a way that is useful to judges and juries. How accurate is an ID made with the highest level of confidence (a rating of 5) likely to be? How accurate are medium-confidence IDs (e.g., ratings of 3)? And how accurate are low-confidence IDs (e.g., ratings of 1)? The answers to these questions are meaningful to judges and jurors (Mickes, 2015), and all of this information is available in Figure A2b. By contrast, the point-biserial correlation coefficient (obtained by fitting the data in Figures A1a and A2a with a straight line) does not provide this information.

For these hypothetical data, which yield a point-biserial correlation of .36, IDs made with high confidence (a rating of 5) are 80% correct, whereas IDs made with low confidence (a rating of 1) are only 33% correct. Thus, a point-biserial correlation that is even less than the magnitude of the widely accepted estimate for choosers (i.e., .41) is consistent with high-confidence IDs being far more accurate than low-confidence IDs. But even this improved analysis underestimates the reliability of eyewitness identification for the same reason that the calibration curves do. What is the problem?

Of the 30 hypothetical choosers shown in Table A1, 22 picked a suspect and the other eight picked a filler (as indicated in Column 2). Imagine that in none of these 22 cases is there any incriminating evidence against the suspect other than the evidence that might be provided by the eyewitness. In this example, eight eyewitnesses chose a filler, thereby ending any further consideration of the suspects in those lineups. But 22 of them identified a suspect, and those 22 identifications are the ones that would go forward as direct evidence of the suspect's guilt. Some of these identifications involve a suspect ID made with high confidence and others involve a suspect ID made with low confidence. The judges and juries in those cases would be interested in knowing whether or not such IDs are reliable. Stated differently, their question is as follows: Of the eyewitness-identification cases that end up before judges and juries (which are limited to identified suspects), what does confidence tell us about the reliability of the ID? Note that this is a question about the 22 cases that go forward to the prosecution using eyewitness identification as direct evidence of the suspect's guilt, not about the full set of 30 cases involving choosers. The answer to this question is provided by limiting the analysis not just to choosers but to *choosers who identify a suspect*—just as the legal system limits its consideration to choosers who identify a suspect.

Table A2 presents the hypothetical data from the 22 choosers who identified a suspect (i.e., it presents the data that would be of interest to judges and jurors), and it now highlights the six choosers in this example who incorrectly identified an innocent suspect from a target-absent lineup. Although none of those six choosers identified an innocent suspect with high confidence (i.e., with

Table A1. Hypothetical Confidence-Accuracy Data From 30 Participant-Witnesses

Witness	Pick type	Lineup type	Accuracy	Confidence
1	Suspect	TP	1	4
2	Suspect	TP	1	4
3	Suspect	TA	0	1
4	Suspect	TP	1	1
5	Suspect	TA	0	2
6	Suspect	TP	1	4
7	Filler	TA	0	5
8	Suspect	TP	1	3
9	Filler	TA	0	2
10	Filler	TP	0	2
11	Suspect	TP	1	5
12	Suspect	TP	1	4
13	Filler	TA	0	1
14	Suspect	TP	1	4
15	Filler	TP	0	1
16	Filler	TA	0	2
17	Suspect	TP	1	2
18	Suspect	TP	1	5
19	Filler	TA	0	4
20	Suspect	TP	1	2
21	Suspect	TP	1	3
22	Suspect	TA	0	3
23	Suspect	TP	1	5
24	Suspect	TP	1	1
25	Suspect	TA	0	1
26	Suspect	TA	0	4
27	Suspect	TP	1	5
28	Suspect	TP	1	2
29	Filler	TA	0	3
30	Suspect	TA	0	2

Table A2. Hypothetical Confidence-Accuracy Data From the 22 Choosers From Table A1

Witness	Pick type	Lineup type	Accuracy	Confidence
1	Suspect	TP	1	4
2	Suspect	TP	1	4
3	Suspect	TA	0	1
4	Suspect	TP	1	1
5	Suspect	TA	0	2
6	Suspect	TP	1	4
8	Suspect	TP	1	3
11	Suspect	TP	1	5
12	Suspect	TP	1	4
14	Suspect	TP	1	4
17	Suspect	TP	1	2
18	Suspect	TP	1	5
20	Suspect	TP	1	2
21	Suspect	TP	1	3
22	Suspect	TA	0	3
23	Suspect	TP	1	5
24	Suspect	TP	1	1
25	Suspect	TA	0	1
26	Suspect	TA	0	4
27	Suspect	TP	1	5
28	Suspect	TP	1	2
30	Suspect	TA	0	2

Note: Confidence ratings range from 1 (*low confidence*) to 5 (*high confidence*). Accuracy is coded as 0 for inaccurate and 1 for accurate. TP = target present; TA = target-absent.

a rating of 5), four of the other 16 witnesses did identify a guilty suspect with high confidence (Witnesses 11, 18, 23, and 27). Thus, high-confidence suspect-ID accuracy in this hypothetical example is perfect (4 correct, 0 incorrect).

Figure A2c shows the results of this analysis when the data are limited to the 22 choosers in Table A1 who identified a suspect. Obviously, the relationship between confidence and accuracy for these hypothetical data is still very strong, in the sense that high-confidence IDs are far more accurate than low-confidence IDs (as illustrated in Fig. 3b). High-confidence suspect IDs are 100% accurate, whereas low-confidence suspect IDs are only 50% accurate (close to chance). Thus, not only is confidence highly diagnostic of accuracy, high-confidence suspect IDs in this hypothetical example are extremely accurate (as accurate as they could possibly be). Keep in mind that these are the very same data that when analyzed using the point-biserial correlation coefficient and including choosers who identify fillers (as in Fig. A1a) yield a value of .36. Even when the point-biserial correlation coefficient is computed for choosers who made suspect IDs (i.e., even when computed using the data in Table A2), its value is only .39. Thus, the correlation coefficient does not convey the information of interest to judges and juries. The data shown in Figure A2c do.

Figure A2c shows a confidence-accuracy characteristic curve (Mickes, 2015). Such a curve plots suspect-ID accuracy as a function of confidence that has been assessed using any numerical scale (in this example, a 1-to-5 scale). Suspect-ID accuracy is computed separately for each level of confidence, *c*, and is computed from the number of suspect IDs from target-present (*TP*) lineups, $nSID_{TP\text{-}c}$, and the number of suspect IDs from

target-absent (*TA*) lineups, $nSID_{TA\text{-}c}$. More specifically, suspect ID accuracy for a given level of confidence is equal to $nSID_{TP\text{-}c}$ / ($nSID_{TP\text{-}c}$ + $nSID_{TA\text{-}c}$). In the example above, for high-confidence IDs (i.e., c = 5), $nSID_{TP\text{-}5}$ = 4 and $nSID_{TA\text{-}5}$ = 0, so high-confidence suspect-ID accuracy is 4 / (4 + 0) = 1.0. This accuracy score differs from the usual dependent measure in calibration studies, in which filler IDs are included in the denominator (as in Fig. A2b). Obviously, including filler IDs lowers the estimated accuracy score, although in this case it has little effect on the overall correlation between confidence and accuracy. However, the correlation is not relevant for what judges and juries want to know, because the correlation could be perfect and yet high-confidence IDs could still (hypothetically) be only 60% accurate. Hence our concentration on the confidence-accuracy characteristic and the probability correct associated with high- and low-confidence suspect IDs.

Appendix B

Estimating suspect-ID accuracy from a calibration score

Most of the calibration studies we reviewed did not present their data in sufficient detail to directly calculate suspect-ID accuracy, so we computed an estimate from the calibration data reported in figures. WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/) was first used to estimate C_c (proportion correct, C, for each level of confidence, c). We then converted those scores, which included filler IDs, to scores that included only suspect IDs. The conversion from C_c to suspect-ID accuracy, $p(TP|SID_c)$, is straightforward. Using the most common calibration formula (which excludes filler IDs from target-present lineups), calibration for a given level of confidence is:

$$C_c = \frac{nSID_{TP-c}}{nSID_{TP-c} + nFID_{TA-c}} \tag{1}$$

To convert C_c to suspect-ID accuracy, we use the following formula:

$$p\left(TP \mid SID_c\right) = \frac{c_c}{c_c + \left(1 - c_c\right)/n} \tag{2}$$

where n = lineup size. As an example, imagine a study using eight-person lineups in which there were 80 correct high-confidence suspect IDs from target-present lineups and 80 high-confidence incorrect IDs from fair target-absent lineups that did not have a designated innocent suspect. Thus, $nSID_{TP\text{-}high}$ = 80 and $nFID_{TA\text{-}high}$ = 80. In that case, calibration for high-confidence IDs (Equation 1) would equal 80 / (80 + 80) = .50. However, to compute suspect-ID accuracy, the number of high-confidence filler IDs from target-absent lineups, $nFID_{TA\text{-}high}$, is divided by lineup size to estimate the number of innocent-suspect IDs from target-absent lineups, $nSID_{TA\text{-}high}$, where $nSID_{TA\text{-}high}$ = $nFID_{TA\text{-}high}$ / n. Note that suspect-ID accuracy is given by:

$$\frac{nSID_{TP-c}}{nSID_{TP-c} + nSID_{TA-c}}$$

Thus, for this example, suspect-ID accuracy (the proportion of suspect IDs that were correct) is 80 / (80 + 80 / 8), which reduces to 1 / (1 + 1 / 8) = .89. However, all we have is the reported calibration accuracy score of .50 (estimated from a figure). Using the above formula (Equation 2), the calibration score is converted into a suspect-ID accuracy score by computing .50 / [.50 + (1 − .50) / 8)], which reduces to 1 / (1 + 1 / 8) = .89. Thus, Equation 2 gives us the right answer (i.e., the same answer we came up with by directly computing suspect-ID accuracy from the raw counts of suspect IDs and filler IDs—the kind of information we do not have access to in most studies). Equation 2 was used to compute suspect-ID accuracy from the calibration scores for each level of confidence—scores that were estimated from the reported figures. All of the studies involved a base rate of approximately 50% (i.e., 50% of the lineups were target-present lineups, and 50% were target-absent lineups).

Appendix C

A short primer on base rates in lineups

The probability that some proposition is true (e.g., that a suspect is guilty) given the result of an evidentiary test (e.g., identification by a witness in a lineup test) is a function of both the diagnostic value of the evidence (e.g., the reliability of the identification) and the base-rate (or prior) probability that the proposition is true. This is often counterintuitive, and people commonly assume that the probability that a proposition is true is equal to the diagnostic value of the evidence without regard to the base rate. Consider, for example, a prostate exam that gives a positive result 98% of the time when there is cancer (a 98% hit rate) and a positive result only 2% of the time when there is no cancer (a 2% false-positive rate). Armed with such information, most people will assume that a positive result indicates a 98% chance of cancer. But that would be true only if one were sampling from a population of men for whom the base rate of prostate cancer was 50% to begin with. Suppose, however, the test is conducted on relatively young men for whom the base rate for prostate cancer is a mere 1%. In the 1%-base-rate

population, a positive test result would yield a probability of cancer of slightly less than 5%, not 98%.

The influence of base rates is somewhat counterintuitive, but the math is not particularly difficult. Consider, for example, that in the 1%-base-rate population of young males, 999 of every 1,000 males tested would not have cancer. However, because there is a 2% false-positive rate for the test, 20 of these young males would have a false-positive result (2% of 999 = 19.98). The one male with cancer among the 1,000 young males would almost certainly yield a positive result as well. So, 21 of the young males would have a positive test result, but only one of the 21 would actually have cancer. Hence, the probability that any one of these young males who had a positive result actually has cancer would be only about 1 in 21, or 4.8%.

This same base-rate issue applies to police lineup tests. Specifically, the probability that a suspect is guilty given that the witness identified that suspect is a function of both the diagnostic value of the evidence and the base-rate probability that a lineup's suspect is guilty. Imagine one extreme jurisdiction (the "Bumbling Detectives PD") in which none of the lineups that police conduct include the guilty suspect (i.e., the target-present base rate is 0%). With a 0% base rate, even a miniscule false-positive rate yields only mistaken identifications and no accurate identifications. Now imagine the other extreme (the "Perfect Detectives PD"), a jurisdiction in which the suspect in a lineup is always the perpetrator (i.e., the target-present base rate is 100%). With a 100% base rate, even a high false-positive rate would yield no false positives on suspect identifications: Every ID of a suspect would be accurate.

When the base rate is 0%, the accuracy rate for identifications of the suspect is 0%, and when the base rate is 100%, the accuracy rate for identifications of suspect is 100%. Of course, real base rates for target-present lineups in police departments will lie somewhere between these two extremes. And, as one moves from the 0% base rate to the 100% base rate, the probability that the identified suspect is the perpetrator follows a Bayesian curve (not a straight line)—a prior-by-posterior probability curve.

Consider the prior-by-posterior curves that we created for the Wetmore et al. (2015) data as displayed in Figure 8. We used Bayes's theorem to calculate each point in these curves. Here, we show how three specific points on the moderate-confidence curve were calculated—one at the 30% base rate, one at the 50% base rate, and one at the 80% base rate.

The vertical axis in Figure 8 is the probability that the suspect is the perpetrator given that the witness identified the suspect from the lineup, which is what we are trying to estimate. We use the expression $p(SP|IDS)$ to represent the probability that the suspect is the perpetrator (SP) given an identification of the suspect (IDS). We use the expression $p(IDS|SP)$ to represent the probability of identification of the suspect (IDS) given that the suspect is the perpetrator (SP). In effect, $p(IDS|SP)$ is the hit rate. Likewise, $p(IDS|SNP)$ is the probability of identification of the suspect (IDS) given that the suspect is *not* the perpetrator (SNP). In effect, $p(IDS|SNP)$ is the false-alarm rate. The term $p(SP)$ is the target-present base rate (or prior probability that the suspect is the perpetrator). The term $p(SNP)$ is, in effect, the target-absent base rate, which is $1 - p(SP)$. We can then put the data into a version of Bayes's theorem as shown below.

$$p(SP \mid IDS) = \frac{p(IDS \mid SP) \times p(SP)}{\left(p(IDS \mid SP)\, p(SP)\right) + \left(p(IDS \mid SNP)\, p(SNP)\right)}$$

In the Wetmore et al. (2015) data, $p(IDS|SP)$ (i.e., the hit rate) for moderate-confidence witnesses was 72.3%, and $p(IDS|SNP)$ (i.e., the mistaken-identification rate) was 10.6%. These two values do not change as a function of the base rate. In effect, these two values constitute the diagnosticity of IDs by the moderate-confidence witnesses. Using the Bayesian expression in Equation 1, the probability that the suspect is the perpetrator given that the witness identified the suspect for the 50% base rate is:

$$p(SP \mid IDS) = \frac{.723 \times .50}{(.723 \times .50) + (.106 \times .50)} = .872$$

Suppose, however, that the base rate was 80%. The probability that the suspect is the perpetrator given that the witness identified the suspect for the 50% base rate is:

$$p(SP \mid IDS) = \frac{.723 \times .80}{(.723 \times .80) + (.106 \times .20)} = .965$$

And, if the base rate was 30%, the probability that the suspect is the perpetrator given that the witness identified the suspect is:

$$p(SP \mid IDS) = \frac{.723 \times .30}{(.723 \times .30) + (.106 \times .70)} = .745$$

Each of these three points on the moderate-confidence curve can be observed in Figure 8.

The degree to which base-rate changes (e.g., from 30% to 80%) moderate the probability that an identified suspect is guilty depends on the diagnosticity of the witness. As diagnosticity increases, the effect of the base rate diminishes. For example, for moderate-confidence witnesses in the Wetmore et al. data, moving from a 30% base rate to an 80% base rate changed the probability that the suspect was the perpetrator from 74.5% to 96.5%, a change of over

20 percentage points. But for high-confidence witnesses, moving from the 30% base rate to the 80% base rate changed the probability that the suspect was the perpetrator from 87.7% to 98.5%, a change of less than 11 percentage points. And for low-confidence witnesses, moving from the 30% base rate to the 80% base rate changed the probability that the suspect was the perpetrator from 62.8% to 94.0%, a change of over 30 percentage points.

Another observation about base rates of note here is that the confidence of the witness makes more difference to our ability to trust the identification when the base rate is in the lower ranges than when the base rate is in the upper ranges. Using the Wetmore et al. (2015) data, for example, when the base rate is 35%, the probability that an identified suspect is guilty for low-confidence witnesses is 22% lower than it is for high-confidence witnesses. When the base rate is 90%, however, the probability that an identified suspect is guilty for low-confidence witnesses is only 2% lower than it is for high-confidence witnesses. This means that when jurisdictions have lineups with relatively low target-present base rates, the importance of eyewitness confidence is even greater than when their lineups' base rates are higher.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Notes

1. Note that in these studies and in most of the ones we consider later, there are more IDs made with high than low confidence, so suspect-ID accuracy scores for low-confidence IDs tend to be more variable than for high-confidence IDs.
2. Only 5% of witnesses made IDs at the highest level (a rating of 7) of confidence, which makes the sample size unstable for isolating this one level of confidence. Hence, we combined confidence levels 6 and 7.

References

American Psychological Association. (2014). *Commonwealth v. Gomes and Commonwealth v. Johnson*. Retrieved from http://www.apa.org/about/offices/ogc/amicus/gomes-johnson.aspx

Behrman, B.W., & Davey, S.L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior*, *25*, 475–491.

Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior*, *29*, 279–301. doi:10.1007/s10979-005-3617-y

Berard, P. (2014). Eyewitness testimony at heart of court cases. *New England Psychologist*. Retrieved from http://www.nepsy.com/articles/leading-stories/eyewitness-testimony-at-heart-of-court-cases/

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology*, *72*, 691–695. doi:10.1037/0021-9010.72.4.691

Bradfield, A., & McQuiston, D. E. (2004). When does evidence of eyewitness confidence inflation affect judgments in a criminal trial? *Law and Human Behavior*, *28*, 369–387. doi:10.1037/1076-898x.8.1.44

Brady v. Maryland (1983). 373 U.S. 83.

Brainerd, C. J., & Reyna, V. F. (2002). Recollection rejection: How children edit their false memories. *Developmental Psychology*, *38*, 156–172. doi:10.1037//0012-1649.38.1.156

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York, NY: Oxford University Press.

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

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

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science*, *23*, 1208–1214. doi:10.1177/0956797612441217

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

Brewer, N., & Wells, G. (2011). Eyewitness identification. *Current Directions in Psychological Science*, *20*, 24–27. doi:10.1177/0963721410389169

Brodes v. State, 614 SE 2d 766 (2005).

Brown, E., Deffenbacher, K., & Sturgill, W. (1977). Memory for faces and the circumstances of the encounter. *Journal of Applied Psychology*, *62*, 311–318. doi:10.1037/0021-9010.62.3.311

Buratti, S., & Allwood, C. M. (2012). Improved realism of confidence for an episodic memory event. *Judgment and Decision Making*, *7*, 590–601.

Carlson, C. A., & Carlson, M. A. (2014). An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *3*, 45–53. doi:10.1016/j.jarmac.2014.03.004

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

Charman, S. D., & Wells, G. L. (2007). Applied lineup theory. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Memory for people* (pp. 219–254). Mahwah, NJ: Lawrence Erlbaum.

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Highly dissimilar fillers increase confidence in lineup

identifications. *Law and Human Behavior*, *35*, 479–500. doi:10.1007/s10979-010-9261-1

Charman, S. D., & Wells, G. L. (2012). The moderating effect of ecphoric experience on post-identification feedback: A critical test of the cues-based inference conceptualization. *Applied Cognitive Psychology*, *26*, 243–250. doi:10.1002/acp.1815

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

Clark, S. E., & Wells, G. L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior*, *32*, 406–422. doi:10.1007/s10979-007-9115-7

Clifford, B. R., & Scott, J. (1978). Individual and situational factors in eyewitness testimony. *Journal of Applied Psychology*, *63*, 352–359. doi:10.1037/0021-9010.63.3.352

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, *23*, 482–505. doi:10.1002/acp.1490

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, *4*, 243–260. doi:10.1007/bf01040617

Deffenbacher, K. A., Brown, E. L., & Sturgill, W. (1978). Some predictors of eyewitness memory accuracy. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 219–226). New York, NY: Academic Press.

Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied*, *19*, 345–357. doi:10.1037/a0034596

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

Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10- to 12-second rule for distinguishing accurate from inaccurate positive identifications. *Journal of Applied Psychology*, *87*, 951–962. doi:10.1037/0021-9010.87.5.951

Dysart, J. E., Lawson, V. Z., & Rainey, A. (2012). Blind lineup administration as a prophylactic against the post-identification feedback effect. *Law and Human Behavior*, *36*, 312–319. doi:10.1037/h0093921

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

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

Garrioch, L., & Brimacombe, C. A. E. (2001). Lineup administrators' expectations: Their impact on eyewitness confidence. *Law and Human Behavior*, *25*, 299–315. doi:10.1023/a:1010750028643

Garry, M., Manning, C. G., Loftus, E. F., & Sherman, S. J. (1996). Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin & Review*, *3*, 208–214. doi:10.3758/bf03212420

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*, 140–152. doi:10.1037/a0015082

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *1*, 221–228. doi:10.1016/j.jarmac.2012.09.003

Horry, R., & Brewer, N. (2016). How target-lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General*, *145*, 1615–1634. doi.10.1037/xge0000227

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

Innocence Project. (2013). Brief of amicus curiae the Innocence Project in support of Thomas and Raymond Highers motion in limine for an eyewitness-identification expert. The People of the State of Michigan vs. Raymond Highers & Thomas Highers (Case No. 87-6345).

Innocence Project. (2016). DNA Exonerations in the United States. Retrieved January 17, 2017. http://www.innocenceproject.org/dna-exonerations-in-the-united-states/

Jones, E. E., Williams, K. D., & Brewer, N. (2008). "I had a confidence epiphany!": Obstacles to combating post-identification confidence inflation. *Law and Human Behavior*, *32*, 164–176. doi:10.1007/s10979-007-9101-0

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1304–1316. doi:10.1037/0278-7393.22.5.1304

Kassin, S. M. (1998). Eyewitness identification procedures: The fifth rule. *Law and Human Behavior*, *23*, 649–654. doi:10.1023/a:1025702722645

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

Klobuchar, A., Steblay, N. K., & Caligiuri, H. L. (2006). Improving eyewitness identification: Hennepin County's blind sequential lineup pilot project. *Cardozo Public Law, Policy, and Ethics Journal*, *2*, 381–414.

Lampinen, J. M., Erickson, W. B., Moore, K. N., & Hittson, A. (2014). Effects of distance on face recognition: Implications for eyewitness identification. *Psychonomic Bulletin & Review*, *21*, 1489–1494. doi:10.3758/s13423-014-0641-2

Lampinen, J. M., Neuschatz, J. S., & Cling, A. D. (2012). *The psychology of eyewitness identification*. New York, NY: Psychology Press.

Leach, A., Cutler, B. L., & Van Wallendael, L. (2009). Lineups and eyewitness identification. *Annual Review of Law and*

Social Science, *5*, 157–178. doi:10.1146/annurev.lawsocsci.093008.131529

Leippe, M. R., Wells, G. L., & Ostrom, T. M. (1978). Crime seriousness as a determinant of accuracy in eyewitness identification. *Journal of Applied Psychology*, *63*, 345–351. doi:10.1037//0021-9010.63.3.345

Lindsay, D. S. (2014). Memory source monitoring applied. In T. Perfect & D. S. Lindsay (Eds.), *The SAGE handbook of applied memory* (pp. 59–75). London, England: Sage.

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior*, *24*, 685–697. doi:10.1023/a:1005504320565

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science*, *9*, 215–218. doi:10.1111/1467-9280.00041

Lindsay, R. C. L. (1986). Confidence and accuracy in eyewitness identification from lineups. *Law and Human Behavior*, *10*, 229–239. doi:10.1007/bf01046212

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 and Human Behavior*, *18*, 527–541. doi:10.1007/bf01499172

Lindsay, R. C. L., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior*, *4*, 303–313. doi:10.1007/bf01040622

Luus, C. A. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior*, *15*, 43–57. doi:10.1007/bf01044829

Luus, C. A. E., & Wells, G. L. (1994). The malleability of eyewitness confidence: Co-witness and perseverance effects. *Journal of Applied Psychology*, *79*, 714–723. doi:10.1037/0021-9010.79.5.714

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago Face Database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*, 1122–1135.

Manson v. Braithwaite, 432 U.S. *98*, 1977.

Massachusetts Court System. (2015). Criminal model jury instructions, 9160-Identification. Retrieved from http://www.mass.gov/courts/docs/courts-and-judges/courts/district-court/jury-instructions-criminal/6000-9999/9160-defenses-identification.pdf

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

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous vs. sequential lineups. *Journal of Experimental Psychology: Applied*, *18*, 361–376. doi:10.1037/a0030609

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140*, 239–257. doi:10.1037/a0023007

Münsterberg, H. (1908). *On the witness stand: Essays on psychology and crime*. Garden City, NY: Clark, Boardman.

National Institute of Justice. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: Office of Justice Programs, U.S. Department of Justice.

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: The National Academies Press.

Neal, T. M. S., Christiansen, A., Bornstein, B. H., & Robicheaux, T. R. (2012). The effects of mock jurors' beliefs about eyewitness performance on trial judgments. *Psychology, Crime & Law*, *18*, 49–64. doi:10.1080/1068316x.2011.587815

New Jersey Courts. (2012a). Identification - in court and out of court identifications [New Jersey criminal model jury charges]. Retrieved from http://www.judiciary.state.nj.us/pressrel/2012/jury_instruction.pdf

New Jersey Courts. (2012b, July 19). *Supreme Court releases eyewitness identification criteria for criminal cases* [Press release]. Retrieved from http://www.judiciary.state.nj.us/pressrel/2012/pr120719a.htm

Palmer, M., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*, 55–71. doi:10.1037/a0031602

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, *1*, 817–845. doi:10.1037/1076-8971.1.4.817

Police Executive Research Forum. (2013). A national survey of eyewitness identification procedures in law enforcement agencies. Retrieved from http://www.policeforum.org/assets/docs/Free_Online_Documents/Eyewitness_Identification/a%20national%20survey%20of%20eyewitness%20identification%20procedures%20in%20law%20enforcement%20agencies%202013.pdf

Quinlivan, D. S., Neuschatz, D. S., Wells, G. L., Cutler, B. L., McClung, J. E., & Harker, D. (2012). Do pre-admonition suggestions moderate the effect of the unbiased-lineup instructions? *Legal and Criminological Psychology*, *17*, 165–176. doi:10.1348/135532510x533554

Read, J. D., Lindsay, D. S., & Nichols, T. (1998). The relation between confidence and accuracy in eyewitness identification studies: Is the conclusion changing? In C. P. Thomson, D. Bruce, J. D. Read, D. Hermann, D. Payne, & M. P. Toglia (Eds.), *Eyewitness memory: Theoretical and applied perspectives* (pp. 107–130). Mahwah, NJ: Lawrence Erlbaum.

Read, J. D., Tollestrup, P., Hammersley, R., McFadzen, E., & Christensen, A. (1990). The unconscious transference effect: Are innocent bystanders ever misidentified? *Applied Cognitive Psychology*, *4*, 3–31. doi:10.1002/acp.2350040103

Read, J. D., Yuille, J. C., & Tollestrup, P. (1992). Recollections of a robbery: Effects of alcohol and arousal upon recall and person identification. *Law and Human Behavior*, *16*, 425–446. doi:10.1007/bf02352268

Reinitz, M. T., Seguin, J. A., Peria, W., & Loftus, G. R. (2012). Confidence-accuracy relations for faces and scenes: Roles of features and familiarity. *Psychonomic Bulletin & Review*, *19*, 1085–1093. doi:10.3758/s13423-012-0308-9

Report of the Special Master, State v. Henderson, No. A-8–08. (2011, June 18). Retrieved from http://www.judiciary

.state.nj.us/pressrel/HENDERSON%20FINAL%20BRIEF%20.PDF%20%2800621142%29.PDF

Rodriguez, D. N., & Berry, M. A. (2014). The effect of line-up administrator blindness on recording of eyewitness identification decisions. *Legal and Criminological Psychology*, *19*, 69–79. doi:10.1111/j.2044-8333.2012.02058.x

Roediger, H. L., & DeSoto, K. A. (2015). Understanding the relation between confidence and accuracy in reports from memory. In D. S. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. Roediger (Eds.), *Remembering: Attributions, processes, and control in human memory: Papers in honor of Larry L. Jacoby* (pp. 347–367). New York, NY: Psychology Press.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, *3*, 377–386. doi:10.1017/s0140525x00075506

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, *11*, 446–453. doi:10.1111/1467-9280.00287

Ross, J. E. (2006). The entrenched position of plea bargaining in United States legal practice. *The American Journal of Comparative Law*, *54*, 717–732. doi:10.2307/20454559

Sauer, J., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General*, *137*, 528–554. doi:10.1037/a0012712

Sauer, J., Brewer, N., & Weber, N. (2012). Using confidence ratings to identify a target among foils. *Journal of Applied Research in Memory and Cognition*, *1*, 80–88. doi:10.1016/j.jarmac.2012.03.003

Sauer, J., Brewer, N., & Wells, G. L. (2008). Is there a magical time boundary for diagnosing eyewitness identification accuracy in sequential line-ups? *Legal and Criminological Psychology*, *13*, 123–135. doi:10.1348/135532506x159203

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

Sauer, J., Weber, N., & Brewer, N. (2012). Using ecphoric confidence ratings to discriminate seen from unseen faces: The effects of retention interval and distinctiveness. *Psychonomic Bulletin & Review*, *19*, 490–498. doi:10.3758/s13423-012-0239-5

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

Smalarz, L., & Wells, G. L. (2015). Contamination of eyewitness self-reports and the mistaken-identification problem. *Current Directions in Psychological Science*, *24*, 120–124. doi:10.1177/0963721414554394

Smith, H. M. J., & Flowe, H. D. (2014). ROC analysis of the verbal overshadowing effect: Testing the effect of verbalization on memory sensitivity. *Applied Cognitive Psychology*, *29*, 159–168. doi:10.1002/acp.3096

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology*, *78*, 22–33. doi:10.1037/0021-9010.78.1.22

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*, 315–327. doi:10.1037/0033-2909.118.3.315

State v. Guilbert. Supreme Court of Connecticut, 306 Conn. 218 (2012).

State v. Lawson (2012). 352 Or. 724, 291 P.3d 673.

State v. Mitchell, 275 P.3d 905 (Kan. 2012).

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

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

Thompson, J. (2000, June 18). I was certain, but I was wrong. *The New York Times*. Retrieved from http://www.nytimes.com/2000/06/18/opinion/i-was-certain-but-i-was-wrong.html

Thompson-Cannino, J. T., Cotton, R., & Torneo, E. (2009). *Picking Cotton: Our memoir of justice and redemption*. New York, NY: St. Martin's Press.

Tollestrup, P. A., Turtle, J. W., & Yuille, J. C. (1994). Actual witnesses to robbery and fraud: An archival analysis. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 144–162). New York, NY: Cambridge University Press.

U.S. Department of Justice. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: Office of Justice Programs.

Valentine, T., & Davis, J. P. (2015). *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and CCTV*. West Sussex, UK: John Wiley.

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

Wells, G. L. (1978). Applied eyewitness testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology*, *36*, 1546–1557. doi:10.1037/0022-3514.36.12.1546

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, *14*, 89–103. doi:10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist*, *48*, 553–571. doi:10.1037/0003-066x.48.5.553

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. doi:10.1037/0021-9010.83.3.360

Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitnesses' recollections: Can the post-identification feedback effect be moderated? *Psychological Science*, *10*, 138–144. doi:10.1111/1467-9280.00121

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. L. (1981). The tractability of eyewitness confidence and its implication for triers of fact. *Journal of Applied Psychology*, *66*, 688–696. doi:10.1037/0021-9010.66.6.688

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior*, *3*, 285–293. doi:10.1007/bf01039807

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness non-identifications. *Psychological Bulletin*, *88*, 776–784. doi:10.1037/0033-2909.88.3.776

Wells, G. L., & Luus, E. (1990). Police lineups as experiments: Social methodology as a framework for properly-conducted lineups. *Personality and Social Psychology Bulletin*, *16*, 106–117. doi:10.1177/0146167290161008

Wells, G. L., Memon, A., & Penrod, S. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, *7*, 45–75. doi:10.1111/j.1529-1006.2006.00027.x

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). On the selection of distractors for eyewitness lineups. *Journal of Applied Psychology*, *78*, 835–844. doi:10.1037//0021-9010.78.5.835

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *23*, 603–647. doi:10.1023/a:1025750605807

Wells, G. L., & Turtle, J. W. (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin*, *99*, 320–329. doi:10.1037/0033-2909.99.3.320

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior*, *39*, 99–122. doi:10.1037/lhb0000125

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*, 8–14. doi:10.1016/j.jarmac.2014.07.003

Wilson, J. P., Hugenberg, K., & Bernstein, M. J. (2013). The cross-race effect and eyewitness identification: How to improve recognition and reduce decision errors in eyewitness situations. *Social Issues and Policy Review*, *7*, 83–113. doi:10.1111/j.1751-2409.2012.01044.x

Wise, R. A., Safer, M. A., & Maro, C. M. (2011). What U.S. law enforcement officers know and believe about eyewitness factors, eyewitness interviews, and identification procedures. *Applied Cognitive Psychology*, *25*, 488–500. doi:10.1002/acp.1717

Wixted, J. T., Mickes, L., Clark, S. E., Dunn, J. C., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, USA*, *113*, 304–309. doi:10.1073/pnas.1516814112

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

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