Journal of Applied Research in Memory and Cognition 5 (2016) 192–203

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

The Effect of Retention Interval on the Eyewitness Identification Confidence–Accuracy Relationship[☆]

John T. Wixted[*]
University of California, San Diego, United States

J. Don Read
Simon Fraser University, Canada

D. Stephen Lindsay
University of Victoria, British Columbia, Canada

Individual researchers express a variety of views about the eyewitness confidence–accuracy relationship, but an argument could be made that the consensus view in the field is that (1) confidence is, at best, a weak indicator of accuracy, (2) the confidence–accuracy relationship becomes weaker still as the retention interval increases, and (3) eyewitnesses who express high confidence tend to be overconfident – perhaps even more so following a long retention interval. Here, we reanalyze the data from four previous retention interval studies in terms of *suspect ID accuracy* (Mickes, 2015). We argue that this measure is more relevant to the information sought by a court of law than either a correlation coefficient or a calibration curve (the two traditional confidence–accuracy measures). The results of our reanalysis suggest that the confidence accuracy relationship remains strong – and that high-confidence IDs remain highly accurate – even after retention intervals as long as 9 months.

Keywords: Eyewitness memory, Identification, Confidence and accuracy, Calibration

In a typical eyewitness identification (ID) experiment concerned with the relationship between confidence and accuracy, confidence is assessed immediately after an adult makes an ID from a fair lineup without undue influence from a lineup administrator. Even under those pristine testing conditions, it was once argued that confidence and accuracy are weakly related when all participants are included in the analysis (see Wells & Murray, 1984, for an early review) or are modestly related when the analysis is limited to “choosers” – namely, those who make an ID of a suspect or filler from a lineup (see Sporer, Penrod, Read, & Cutler, 1995, for a later review). These assessments were based on studies in which the confidence–accuracy relationship was measured using the point-biserial correlation coefficient. However, because that summary statistic can be deceptively low even when the confidence–accuracy relationship is undeniably strong (Juslin, Olsson, & Winman, 1996), researchers nowadays more often plot calibration curves to represent how accuracy varies as a function of confidence. Such research leaves little doubt that, under typical testing conditions, confidence is highly predictive of accuracy in the straightforward sense that a low-confidence ID implies low accuracy whereas and a high-confidence ID implies considerably higher accuracy (e.g., Brewer & Wells, 2006).

In addition to knowing the difference in accuracy between IDs made with low vs. high confidence, it is of interest to know the absolute accuracy of high-confidence suspect IDs. As noted by Sauer, Brewer, Zweck, and Weber (2010), “Highly confident identifications, when compared to those made with low confidence, are likely to have a greater impact on police investigations and jury decision making” (p. 344). With regard to jury decision making, if high-confidence IDs are only 60% correct (for

Author Note

John T. Wixted, Department of Psychology, University of California, San Diego; J. Don Read, Department of Psychology, Simon Fraser University, British Columbia, Canada; D. Stephen Lindsay, University of Victoria, British Columbia, Canada.

[☆] Please note that this paper was handled by the current editorial team of JARMAC.

[*] Correspondence concerning this article should be addressed to John T. Wixted, University of California, Department of Psychology, 9500 Gilman Drive, La Jolla, CA 92093, United States. Contact: jwixted@ucsd.edu.

example), then it is hard to imagine that any juror would regard that level of accuracy as exceeding the "beyond a reasonable doubt" standard. If, instead, high-confidence IDs are 97% correct, then some jurors might regard such an ID as exceeding that high standard. Thus, high-confidence accuracy in general – and the effect of retention interval on high-confidence accuracy in particular – are key considerations.

Although IDs made with high confidence are now known to be substantially more accurate than IDs made with low confidence, it does not necessarily follow that high-confidence IDs are necessarily reliable. In fact, there is evidence to suggest that the consensus view in the field is that high-confidence IDs are *unreliable* even under the best of conditions. Consider, for example, how the American Psychological Association (APA) recently characterized the eyewitness confidence–accuracy relationship in an amicus brief filed in support of a defendant's request to instruct the jury that "witnesses who are highly confident of their identifications are not therefore necessarily reliable" (American Psychological Association, 2014, p. 3). In that brief, the APA considered core research findings from eyewitness identification research and pointed out that "... there is overwhelming consensus as to the core findings of that research" (p. 10). With regard to the reliability of eyewitness identification assessed under typically ideal testing conditions, the amicus brief took into account studies that measured the confidence–accuracy relationship using a correlation coefficient as well as studies that measured the relationship using calibration. Its interpretation of those findings was as follows:

> "...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" (pp. 17–18).

Keep in mind that the results of these studies pertain to an *initial* ID made from a fair lineup without undue administrator influence, not to the later (potentially contaminated) ID that occurs in court. Clearly, post-identification events can contaminate memory and distort the information value of eyewitness confidence, thereby rendering any later expression of confidence unreliable. Our thesis pertains to an initial, non-leading, suspect identification task. We argue that high-confidence eyewitness identifications are much more reliable than they are characterized as being in this amicus brief (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015).

Not only are high-confidence IDs widely thought to be error prone even when obtained under ideal testing conditions, many people also believe that the trustworthiness of such IDs may become lower still as the retention interval between an event (e.g., the witnessed crime) and the first memory test (e.g., an ID made from a lineup) increases. This is believed to be true even in the absence of memory contamination. For example, in another recent APA amicus brief (American Psychological Association, 2011), the argument was made that "Empirical research also establishes that as time passes between an event and a resulting identification, the identification becomes increasingly unreliable – put simply, the memory 'decays'... Even a gap of only a few hours between exposure and identification, then, can affect the reliability of an identification" (p. 18).

The idea that eyewitness IDs become less reliable with the passage of time not only accords with common sense, it also accords with certain theoretical considerations. For example, according to Deffenbacher's (1980) optimality hypothesis, the strength of the confidence–accuracy relationship varies directly with the optimality of information processing conditions at encoding, storage, and retrieval (Bothwell, Deffenbacher, & Brigham, 1987). As information processing conditions become less optimal, measures of memory performance (e.g., *d*') will reflect that fact. Using unpublished signal-detection-based simulation studies, Deffenbacher (2008) argued that, theoretically, the confidence–accuracy relationship as measured by the point-biserial correlation coefficient weakens as *d'* declines. Because there is little doubt that *d'* declines as the retention interval increases, it follows that confidence should become an even less reliable indicator of accuracy as the retention interval increases (Deffenbacher, 2008). Of course, this argument assumes that the point-biserial correlation appropriately measures the confidence–accuracy relationship, an idea that, as noted above, is no longer widely accepted (Juslin et al., 1996).

Empirically, are eyewitnesses overconfident following a short retention interval, and does that problem become even worse as the retention interval increases? Only a few prior eyewitness identification studies have investigated how the confidence–accuracy relationship changes with the size of the retention interval. Two prior calibration studies used retention intervals that varied from 0 to 1 week (Juslin et al., 1996; Palmer, Brewer, Weber, & Nagesh, 2013), and another calibration study used retention intervals that varied from 0 to 3 weeks (Sauer et al., 2010). In addition, in a study that measured the confidence–accuracy relationship using the correlation coefficient, the retention interval varied from 3 months to 9 months (Read, Lindsay, & Nichols, 1998). As described in more detail below, these studies converge on the notion that the confidence accuracy relationship does not change appreciably for choosers as the retention interval increases despite the fact that *d'* decreases. Whether the retention interval is short or long, low confidence implies low accuracy, and high confidence implies considerably higher accuracy. Nevertheless, these studies also seem to imply that witnesses are generally overconfident following a short retention interval and may become somewhat more so as the retention interval increases.

Here, we advance the argument that these prior studies did not analyze their data in a way that is directly relevant to what a court wants to know when trying to assess the reliability of an eyewitness who has identified the suspect in an initial lineup task. Thus, the jury is still out on exactly how confidence is related to accuracy and how that relationship changes with the length of the retention interval. To provide a court with the information it seeks in a case involving an eyewitness who has identified the suspect in a lineup, the case has recently been made that the data need to be analyzed in terms of *suspect ID accuracy* (Mickes, 2015). Suspect ID accuracy differs from both the correlation coefficient and calibration. Next, we describe the critical difference between suspect ID accuracy and the other confidence–accuracy measures that are more commonly used, and then we re-analyze the relevant retention interval data from the four retention interval studies mentioned above.[1] Reanalyzing those data in terms of suspect ID accuracy is the main purpose of our article. The importance of our reanalysis lies in the fact that the message sent by the available empirical evidence turns out to be the exact opposite of the prevailing consensus view in the field, as conveyed by the APA in its recent amicus briefs (and by many individual experts who testify in courts of law) based on research using the correlation coefficient and calibration.

Suspect ID Accuracy vs. Correlation and Calibration

The relationship between confidence and accuracy has often been measured by computing the Pearson *r* correlation coefficient between the binary 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”). A correct response consists of (1) a suspect ID from a target-present lineup or (2) the rejection of a target-absent lineup, whereas an incorrect response consists of (1) a suspect ID from a target-absent lineup, (2) a filler ID from either type of lineup, or (3) the rejection of a target-present lineup. When the analysis is limited to choosers, the responses made by those who reject target-present or target-absent lineups are excluded from the analysis (e.g., Sporer et al., 1995). Because accuracy is coded as a dichotomous variable, the Pearson *r* in this case is known as a point-biserial correlation coefficient. This approach represents an effort to boil the confidence–accuracy relationship down to a single summary statistic.

As noted earlier, Juslin et al. (1996) long ago explained why the point-biserial correlation coefficient does not provide useful information to a court about the confidence–accuracy relationship. The basic problem is that the correlation can be low even when the confidence–accuracy relationship is as strong as it can possibly be (e.g., when 100% confidence is associated with 100% accuracy, 90% confidence is associated with 90% accuracy, and so on). That being the case, it is a mistake to assume that a low correlation necessarily implies a weak confidence–accuracy relationship. Juslin et al. (1996) further argued that calibration curves are more informative. A calibration curve is a plot of accuracy for each level of confidence considered separately.

However, if the goal is to inform the legal system about the relationship between eyewitness confidence and accuracy, then even a calibration curve can be misleading (Mickes, 2015; Wixted et al., 2015). To appreciate why, it is essential to consider exactly what the dependent variable is in a calibration plot and how it differs from suspect ID accuracy. For choosers, the calibration accuracy formula consists of correct IDs made with a particular level of confidence divided by correct IDs plus incorrect IDs made with that same level of confidence. A correct ID for choosers consists of a suspect ID from a target-present lineup, whereas incorrect IDs for choosers consist of (1) a suspect ID from a target-absent lineup, (2) a filler ID from a target-absent lineup, or (3) a filler ID from a target-present lineup. Thus, for IDs made with a particular level of confidence (c), calibration accuracy for that level of confidence (C_c) is equal to the number of (correct) suspect IDs from a target-present lineup ($nSID_{TP\text{-}c}$) divided by the sum of that value and the total number of errors: suspect IDs from target-absent lineups ($nSID_{TA\text{-}c}$) plus the number of filler IDs from target-absent lineups ($nFID_{TA\text{-}c}$) plus the number of filler IDs from target-present lineups ($nFID_{TP\text{-}c}$).[2] In other words:

$$C_c = \frac{nSID_{TP-c}}{nSID_{TP-c}+nSID_{TA-c}+nFID_{TA-c}+nFID_{TP-c}} \quad (1)$$

Usually, studies that use fair lineups do not have a designated innocent suspect, so $nSID_{TA\text{-}c}$ is not specifically included in the denominator. When $nSID_{TA\text{-}c}$ is removed from Eq. (1), the formula reduces to:

$$C_c = \frac{nSID_{TP-c}}{nSID_{TP-c}+nFID_{TA-c}+nFID_{TP-c}} \quad (2)$$

This is the full calibration formula for studies involving fair lineups without a designated innocent suspect, and it is the one used by the APA in its recent amicus brief. Thus, for example, when this formula is applied to the data shown in Table 3 of Sauer, Brewer and Wells (2008), IDs made with 90–100% confidence are accurate only .60 of the time (i.e., 60% correct). This formula was also used by Brewer, Keast, and Rishworth (2002) in one of the first calibration studies following Juslin et al.’s (1996) seminal report. Since then, however, most calibration studies have chosen to exclude target-present filler IDs ($nFID_{TP\text{-}c}$) from the denominator on the grounds that such IDs involve known innocents. Thus, the only errors included in most

[1] Wetmore et al. (2015) compared lineup performance on an immediate vs. delayed (48-h) test, but performance was almost identical in both conditions, and the lineups were generally unfair in that the innocent suspect in target-absent lineups was identified far more often than the fillers in 3 out of 4 conditions. Thus, we do not consider that study here.

[2] As an example of what this notation represents, $nSID_{TP\text{-}c}$ means that there was a certain number (n) of suspect IDs (SID) made from target-present ($_{TP}$) lineups with confidence level c ($_{-c}$). Similarly, $nFID_{TA\text{-}c}$ means that there was a certain number (n) of filler IDs (FID) made from target-absent ($_{TA}$) lineups with confidence level c ($_{-c}$).

calibration studies are filler IDs from target-absent lineups, so the typical calibration accuracy score is equal to:

$$C_c = \frac{nSID_{TP-c}}{nSID_{TP-c} + nFID_{TA-c}} \tag{3}$$

Although one could reasonably argue that errors are errors, so filler IDs from target-present lineups should be included in the denominator when analyzing the performance of choosers (as the APA did), this is not a critical issue because the results usually do not change much whether or not they are included.

To understand why a calibration score does not accurately inform a court that is seeking to evaluate the reliability of an eyewitness who has identified a suspect with a particular level of confidence, it is helpful to consider that some criminal prosecutions do not involve an eyewitness ID of a suspect but instead involve an eyewitness who mistakenly picked a filler (and who is therefore not asked to testify against the defendant at the trial). Is there any information provided by the fact that the witnesses in cases like that picked a filler? The answer to this question is provided by a study in which the analysis is selectively focused on filler IDs, not suspect IDs. Recently, Wells, Yang and Smalarz (2015) argued that when filler IDs are selectively analyzed in lab studies, the results suggest that a filler ID is actually somewhat probative of innocence (see also Wells & Olson, 2002). In the lab studies they considered, in which half the lineups were target-present and half target-absent (i.e., prior probability of guilt = .50), only 40% of the lineups in which a filler ID occurred contained a guilty suspect (i.e., posterior probability of guilt = .40). That being the case, these studies suggest that a filler ID, when it occurs, provides some evidence that the suspect is not guilty. More formally, in an equal base-rate study and ignoring confidence, Wells et al. (2015) argued that the probability (*p*) that the lineup is a target-present (TP) lineup given a filler ID (*FID*) is:

$$p(TP|FID) = \frac{nFID_{TP}}{nFID_{TP} + nFID_{TA}} \approx .40 \tag{4}$$

One could also compute this value separately for different levels of confidence to ask whether or not confidence in a filler ID provides any additional information about the likelihood that the suspect is guilty. Wells et al. (2015) argued that a filler ID made with high confidence is more indicative of innocence than one made with low confidence. Even though the information value of a filler ID appears to be modest, this news is nevertheless relevant to the set of criminal prosecutions in which an eyewitness identified a filler (not to the set of criminal prosecutions in which an eyewitness identified a suspect).

Other criminal prosecutions involve an eyewitness ID of a suspect, not a filler. It is this kind of case that has led to many wrongful convictions based on eyewitness misidentifications. Is there any information provided by the fact that a witness picked the suspect from a lineup? The answer to this question is provided by an eyewitness identification study in which the analysis is selectively focused on suspect IDs, not filler IDs. In contrast to the probative information provided by a filler ID, Wells et al. (2015) found that if a suspect ID occurred, approximately 77% of the lineups contained a guilty suspect (i.e., posterior probability of guilt ≈.77). More formally, in an equal base-rate study and ignoring confidence, they argued that the probability that the lineup is a target-present (TP) lineup given a suspect ID (*SID*) is:

$$p(TP|SID) = \frac{nSID_{TP}}{nSID_{TP} + nSID_{TA}} \approx .77 \tag{5}$$

Thus, using Eqs. (4) and (5), the results reported by Wells et al. (2015) indicate that, from the starting base rate of .50, the probative information provided by suspect IDs on the one hand and filler IDs on the other point in *different directions* with respect to whether or not the suspect is guilty. That being the case, if the goal is to inform the legal system about the probative value of either filler IDs (which apply to one set of cases) or suspect IDs (which apply to a different set of cases), filler IDs and suspect IDs should not be aggregated together. Yet correlation studies and calibration studies that measure the confidence–accuracy relationship for choosers usually aggregate suspect IDs and filler IDs in the same equation (see Eqs. (1)–(3)). Instead of combining them, the information value of the two kinds of IDs should be considered separately, as they are in Eqs. (4) and (5).

Just as one can ask about the information value of filler IDs made with different levels of confidence, one can ask about the information value of suspect IDs made with different levels of confidence. How accurate is a high-confidence suspect ID and how much less accurate is a suspect ID made with low confidence? Calibration studies do not provide the answer to these questions, but a confidence–accuracy characteristic (CAC) analysis does (Mickes, 2015). A CAC analysis plots *suspect ID accuracy* as a function of confidence.[3] The dependent measure in a CAC plot is equal to the probability that the lineup is a target-present (TP) lineup given a suspect ID made with a particular level of confidence (SID_c):

$$p(TP|SID_c) = \frac{nSID_{TP-c}}{nSID_{TP-c} + nSID_{TA-c}} \tag{6}$$

Note that this is the same equation as Eq. (5) except that now the values are computed separately for each level of confidence. High-confidence suspect IDs are far more indicative of guilt than low-confidence suspect IDs (Wells et al., 2015).

To use Eq. (6), one needs to know $nSID_{TP\text{-}c}$ and $nSID_{TA\text{-}c}$. The value of $nSID_{TP\text{-}c}$ is easily obtained because it is simply the number of suspects IDs from target-present lineups made with a particular level of confidence. Similarly, the value of $nSID_{TA\text{-}c}$ is easily obtained when the target-absent lineups include a designated innocent suspect. In that case, it is simply the number of suspect IDs from target-absent lineups made with a particular level of confidence. However, when a fair TA lineup is used without a designated innocent suspect, as is true in most

[3] Because filler IDs provide evidence in the direction of innocence, whereas suspect IDs provide evidence in the direction of guilt, it can be useful to analyze filler IDs and suspect IDs separately instead of bundling them together as choosers (G. Wells, personal communication). CAC analysis shows the information value of suspect IDs as a function of confidence for the 50% base rate situation typically used in eyewitness ID studies, but a complete Bayesian analysis would require a consideration of suspect ID accuracy across the full range of possible base rates, as in Fig. 8 of Wells et al. (2015).

eyewitness identification studies, the number of innocent suspect IDs is typically estimated by dividing the number of IDs from target-absent lineups by lineup size, n (e.g., Palmer et al., 2013). That is, $\sim nSID_{TA\text{-}c} = nFID_{TA\text{-}c}/n$. In that case, suspect ID accuracy for a given level of confidence, c, is given by:

$$p(TP|SID_c) = \frac{nSID_{TP-c}}{nSID_{TP-c} + \sim nSID_{TA-c}} \qquad (7)$$

This is the principal suspect ID accuracy formula that we will use in our reanalyzes. Suspect ID accuracy is not a calibration score (which requires the use of a 100-point confidence scale) and can therefore be plotted as a function of any numerical rating scale of confidence (e.g., a 1–5 scale, a 1–10 scale, a 0–100 scale, etc.). That suspect ID accuracy can be plotted as a function of any confidence scale, whereas calibration requires the use of a 100-point scale, is not the most important difference between these two measures. The most important difference between them is that the information provided by Eq. (7) differs from the information provided by a calibration score because a calibration score bundles together filler IDs (which, when they occur, somewhat decrease one's belief that the suspect is guilty) and suspect IDs (which, when they occur, more strongly increase one's belief that the suspect is guilty). To be sure, the two measures (calibration and suspect ID accuracy) are inherently correlated when a fair lineup is used, but the impression they convey about the reliability of eyewitness IDs – particularly those made with high confidence (which are the most important IDs) – can differ quite dramatically.[4]

The computational details presented above may be somewhat tedious, but they are important nonetheless. As an example of why this is an important issue, consider the fact that high-confidence suspect ID accuracy in Sauer et al. (2008) – which, as noted above and as emphasized in the aforementioned APA amicus brief, comes to only 60% correct when computed using a calibration formula (Eq. (2)) – is equal to 96.8% correct when computed using the CAC approach (see Appendix for computational details). Yet the APA used the results from this same study to explain to a court that "This confirms that many witnesses are overconfident in their identification decisions." Although that statement is technically true (because it applies when aggregating across suspect IDs and filler IDs), it is not relevant to the reliability of a high-confidence *suspect* ID, which is the information that a court cares about when evaluating the reliability of an eyewitness who has identified a suspect with a high level of confidence. To provide that information, suspect ID accuracy – not calibration – needs to be reported.

In terms of suspect ID accuracy, should high-confidence IDs made following a short retention interval be characterized as "overconfident" and do such IDs exhibit even greater overconfidence with the passage of time? The available empirical evidence, as it has been analyzed thus far, does not answer these questions because the data were aggregated across suspect IDs and filler IDs. We next plot suspect ID accuracy for data from 4 prior studies that investigated the confidence–accuracy relationship as a function of retention interval. For the sake of brevity, in each case, the data are collapsed across conditions that are not relevant to the retention interval manipulation. The story those data tell is very different from the story told by the APA in its recent amicus briefs and which the APA regards as the consensus view in the field.

Juslin et al. (1996)

Juslin et al. (1996) investigated the confidence accuracy relationship over a delay of up to 1 week. In this study, participants viewed a 90-second videotaped theft involving two perpetrators and were later asked to identify them from two 8-person photo lineups, one for each perpetrator. The fillers were selected from the photo material used by actual police officers in the course of regular police investigations. Half the participants were tested following a 1-hour retention interval and half following a 1-week retention interval. Unlike most studies, the base rate of target-present lineups was .75 (not .50), and participants were asked to make two confidence ratings – a pre-lineup confidence rating that the witness would be able to recognize the perpetrator in the lineup and a second confidence rating if a lineup member was identified as the perpetrator.

According to the calibration results, which were computed using Eq. (2) and which we estimated from their Figure 4 using WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/), in the 1-h condition, participants who were 15% confident were approximately 30% correct, whereas participants who were 95% confident were approximately 89% correct. In the 1-week condition, low-confidence accuracy was approximately 16% correct, whereas high-confidence accuracy was approximately 87% correct. Thus, the confidence–accuracy relationship was strong in both conditions (in terms of the difference in accuracy for IDs made with low vs. high confidence), but participants were somewhat overconfident at the high end of the confidence scale.

The calibration estimates were next used to compute suspect ID accuracy (Eq. (7)) using a conversion formula presented in Appendix. Exact suspect ID accuracy scores could not be computed because the raw data from this study are not available, but it was possible to obtain estimates that are, if anything, slightly lower than the true values. Those estimated suspect ID accuracy scores are shown here in Figure 1A. The results still show a conspicuously strong confidence–accuracy relationship in that low-confidence IDs are much less accurate than high-confidence IDs in both conditions. Moreover, the accuracy of high-confidence suspect IDs is 98.5% correct in the 1-h condition and 98.2% in the 1-week condition. Thus, there is nothing in these data to suggest that the confidence–accuracy relationship weakened appreciably as a function of the retention interval or that participants who made high-confidence IDs were overconfident in either retention interval condition. Instead, witnesses who made high-confidence IDs were, if anything, underconfident.

As noted by Sauer et al. (2010), there are several reasons to be cautious about interpreting these results. First, the target-present

[4] Because Eq. (7) does not consist of only raw frequency counts (instead, one value in the denominator is divided by lineup size), computing standard errors for suspect ID accuracy scores in not straightforward. In the appendix, we describe one way to compute these standard errors using a simple Monte Carlo simulation.

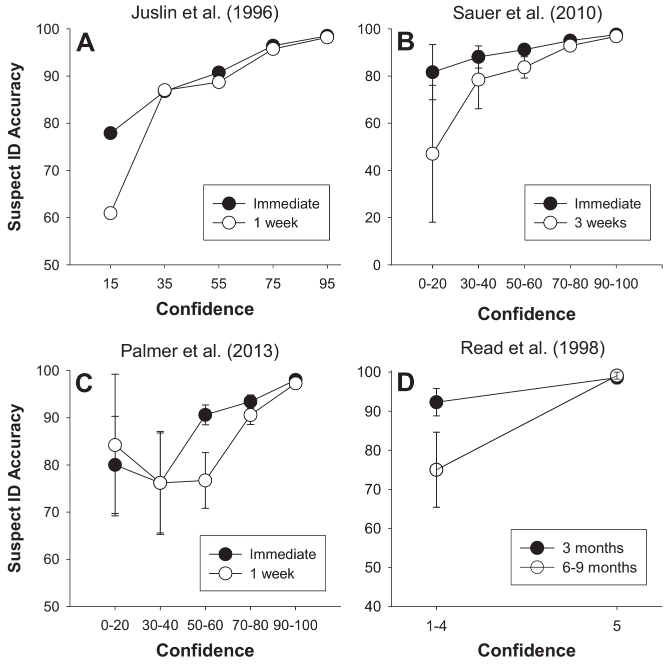

Figure 1. Suspect ID accuracy scores as a function of confidence from 4 eyewitness identification studies in which retention interval was varied. All 4 studies used 8-person photo lineups.

base rate was .75 in this study, so random-chance suspect ID accuracy would be 75% correct. As the base rate increases above .50, suspect ID accuracy scores (and calibration scores) increase. Second, the effect of the retention interval manipulation was weak (overall accuracy was reported to be 69% correct in the 1-hour condition and 64% correct in the 1-week condition). Third, an atypical 2-step confidence procedure was used. In addition to those considerations, the values shown in Figure 1A are only approximate because they are based on calibration accuracy scores that were estimated from a figure and converted to suspect ID accuracy scores after making simplifying assumptions (see Appendix). Still, as shown next, the surprising results reported by Juslin et al. (1996) are fairly typical of what has since been observed in other studies.

Sauer et al. (2010)

Sauer et al. (2010) exposed eyewitnesses in a community setting to a target individual for 10 s and then tested them either immediately or after an average delay of 3 weeks using 8-person simultaneous photo lineups. The retention interval manipulation was successful in that overall accuracy was lower following the longer retention interval condition. For example, the overall hit rate in each condition can be computed using the formula $nSID_{TP}/N_{TP}$, where N_{TP} is the number of target-present lineups, and the overall false alarm rate can be computed using the formula $\sim nSID_{TA}/N_{TA}$, where N_{TA} is the number of target-absent lineups. The hit and false alarm rates in the short retention interval condition were .73 and .04, respectively, and the corresponding values for the long retention interval condition were .60 and .05, respectively. If these values are used to compute standard *d*' scores (Mickes, Moreland, Clark, & Wixted, 2014), they come to 2.35 and 1.88, respectively.

Table 1
Overall Hit Rates (HR), Overall False Alarm Rates (FAR), and d' Scores from Experiment 3 of Read et al. (1998)

Performance measure	Retention interval		
	3	6	9
HR	0.783	0.524	0.563
FAR	0.038	0.054	0.055
d'	2.56	1.67	1.76

Applying the calibration formula used by the APA in its amicus brief to the data reported by Sauer et al. (2010, Table 1), which counts as errors fillers IDs from both target-present and target-absent lineups (Eq. (2)), IDs in the immediate condition made with 0–20% confidence were 33.3% correct, whereas IDs made with 90–100% confidence were 80.4% correct. IDs in the delayed condition made with 0–20% confidence were 6.7% correct, whereas IDs made with 90–100% confidence were 74.5% correct. Thus, in both conditions, there was a meaningful relationship between confidence and accuracy, but highly confident witnesses were, according to this calibration measure, clearly overconfident – somewhat more so in the delayed condition compared to the immediate condition. The results were similar if only filler IDs from target-absent lineups were counted (Eq. (3)), which is how Sauer et al. (2010) reported the calibration results for choosers. In that analysis, high-confidence IDs in the immediate condition were 83.3% correct, whereas high-confidence IDs in the delayed condition were 79.2% correct, which means that the witnesses in that experiment could still be reasonably characterized as being overconfident.

Figure 1B replots their data in terms of suspect ID accuracy (Eq. (7)). These values did not have to be estimated and were instead computed directly from the numbers presented in Table 1 of Sauer et al. (2001). The story is similar to the calibration-based story in the sense that confidence is a strong indicator of accuracy in both conditions, but it differs quite dramatically in what the data suggest about the accuracy of high-confidence suspect IDs. High-confidence suspect ID accuracy in the short retention interval condition was 97.6% correct, and high-confidence suspect ID accuracy in the long retention interval condition was 96.8% correct. At lower levels of confidence, suspect ID accuracy increasingly differs for the two conditions (lower accuracy for the longer retention interval condition), but the difference observed at the high end of the scale in this study was very small. Again, what makes this an important result is that calibration studies are routinely interpreted to mean that highly confident witnesses are overconfident (Sauer et al., 2008) and that problem is widely thought to become even worse as the retention interval increases. However, in terms of suspect ID accuracy in this study, no such effects were observed.

Palmer et al. (2013)

In Experiment 1 of Palmer et al. (2013), community participants viewed a target individual for 5 or 90 seconds and then completed an identification test involving an 8-person photo lineup either immediately or following a 1-week retention interval. The overall hit and false alarm rates in the short retention interval condition were .60 and .05, respectively, and the corresponding values for the long retention interval condition were .51 and .05, respectively. If these values are used to compute standard *d*' scores (Mickes et al., 2014), they come to 1.90 and 1.62, respectively.

Using the calibration formula favored by the APA in its amicus brief (Eq. (2)), IDs in the immediate condition made with 0–20% confidence were 23.1% correct, whereas IDs made with 90–100% confidence were 76.6% correct. The corresponding values in the delayed condition were 26.7% correct and 74.2% correct, respectively. In other words, there was a strong confidence–accuracy relationship in both conditions, and participants who made high-confidence IDs were again clearly overconfident. The results were similar if only filler IDs from target-absent lineups were counted as errors (Eq. (3)), which is how Palmer et al. (2013) reported their calibration results for choosers. In that analysis, high-confidence IDs in the immediate condition were 86.0% correct, whereas high-confidence IDs in the delayed condition were 81.7% correct. Again, no matter which formula is used, and as noted by Palmer et al., the results suggest that highly confident witnesses could be reasonably said to be overconfident, perhaps somewhat more so in the delayed condition than the immediate condition.

Figure 1C presents the Palmer et al. (2013) results with the data now plotted in terms of suspect ID accuracy (using Eq. (7)). These values were computed directly from the data presented in their Table 1. The results indicate that suspect ID accuracy was fairly high across the board and that high-confidence suspect ID accuracy is nearly perfect in both retention interval conditions. In the immediate conditions, high-confidence suspect ID accuracy was 98.0% correct; in the delayed condition, it was 97.3% correct.

Read et al. (1998)

The studies discussed above involved retention intervals ranging up to a few weeks, but retention intervals encountered in the real world can be much longer than that. In Experiment 3 of Read et al. (1998), they used retention intervals ranging up to *9 months*. In that study, retail store clerks in Victoria, B.C., were recruited to participate in a study investigating their memory for customers. Initially, a female confederate interviewed each clerk/witness at work for approximately 10 min. Two weeks later, a male confederate visited each clerk/witness' store to conduct another interview and to administer an 8-person photo lineup. The female confederate who had conducted the interview 2 weeks earlier was the target. The data from that test are not analyzed here. Memory for the male confederate was later tested using 8-person photo lineups following a 3-, 6-, or 9-month retention interval, and these are the conditions we focus on. The clerks were tested using either a target-present or a target-absent lineup, and confidence was assessed using a 5-point confidence scale. Table 1 presents the overall hit and false alarm rates (collapsed across confidence) for each retention interval and the corresponding *d*' values. Figure 2 presents the *d*' data as a function of retention interval along with the best-fitting power function (Wixted & Ebbesen, 1991). The data appear to be consistent with a typical forgetting curve.

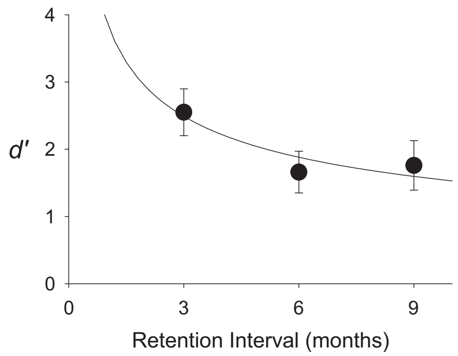

Figure 2. Forgetting function over a 9-month retention interval for the store clerk study reported by Read et al. (1998). The smooth curve represents the least squares fit of the power function, $d' = at^{-b}$, where t = retention interval, and a and b are free parameters. Error bars are standard errors estimated via Monte Carlo simulation (see Appendix).

Table 2 shows the number of suspect IDs, filler IDs and no IDs made from 60 target-present lineups as well as the estimated number of suspect IDs, and the observed number of filler IDs and no IDs made from 62 target-absent lineups across the three retention interval conditions, with each set of data broken down by confidence (1 through 5). We present the raw data here because they have not been presented before and because they can be used to illustrate exactly how suspect ID accuracy was computed. The estimated suspect IDs for target-absent lineups in Table 2 were obtained by dividing IDs made from target-absent lineups by the lineup size of 8. Note that some of the cells involved very few observations, so the data are quite variable. Table 3 shows the corresponding suspect ID accuracy scores computed using Eq. (7).

Table 2
Frequency Counts from Experiment 3 of Read et al. (1998)

	Confidence	Target present Retention interval			Target absent Retention interval		
		3	6	9	3	6	9
Suspect IDs	5	9	9	5	0.13	0.13	0.00
	4	3	0	1	0.13	0.25	0.00
	3	5	1	3	0.13	0.63	0.50
	2	1	1	0	0.25	0.13	0.38
	1	0	0	0	0.25	0.13	0.00
Filler IDs	5	0	0	0	1	1	0
	4	0	0	0	1	2	0
	3	1	1	1	1	5	4
	2	0	1	1	2	1	3
	1	1	2	0	2	1	0
No IDs	5	0	1	1	4	2	1
	4	1	0	0	5	5	4
	3	0	1	2	4	2	2
	2	1	3	0	2	0	1
	1	1	1	2	1	4	1

Note: Suspect IDs for Target Absent lineups are estimated values obtained by dividing Filler IDs for Target Absent lineups by lineup size (8).

Table 3
Suspect ID Accuracy Scores Computed from Target-Present and Estimated Target-Absent Suspect IDs in Table 2

Confidence	Retention interval		
	3	6	9
5	0.986	0.986	1.000
4	0.960	0.000	1.000
3	0.976	0.615	0.857
2	0.800	0.889	0.000
1	0.000	0.000	N/A

Given the small number of observations in some cells, we collapsed the data in two ways for the final confidence–accuracy analysis. First, because *d*' was very similar at the 6- and 9-month retention intervals, we combined the data over the two longer retention interval conditions. The *d*' score for the 3-month retention interval condition (*d*' = 2.55, *SE* = 0.350) was significantly greater than the corresponding score for the combined 6–9-month retention interval condition (*d*' = 1.71, *SE* = 0.231). Second, because suspect IDs made with confidence ratings of 1 through 4 were less frequent than suspect IDs confidence made with confidence ratings of 5, we collapsed the confidence scale into a 2-point scale consisting of low confidence (1–4) vs. high confidence (5). When combined in this manner, it was possible to compute standard errors for resulting suspect ID accuracy scores (as described in Appendix). The suspect ID accuracy data are shown in Figure 1D. Remarkably, high-confidence IDs were nearly perfectly accurate at each retention interval despite the fact that overall accuracy declined with retention interval (Figure 2). Obviously, high confidence implied high accuracy (nearly 100% correct across retention interval conditions, as in the studies considered above) and low confidence implied much lower accuracy.

Discussion

In the eyewitness identification studies considered here, we found that for both short and long retention intervals (1) confidence in a suspect ID from a photo lineup was strongly related to accuracy (as recent calibration studies have suggested) and (2) high-confidence suspect IDs were remarkably accurate (contrary to what recent calibration studies seem to suggest). In each of the four studies considered here, high-confidence IDs remained very accurate despite the unsurprising fact that overall recognition accuracy declined as the length of the retention interval increased. These findings underscore the important fact that a decline in overall memory accuracy (e.g., a decline in *d*') does not automatically imply that an ID made with high confidence also becomes appreciably less reliable.

What is new about these findings and why are the results important? The consensus view in the field is that witnesses who make a high-confidence ID from a lineup, although more accurate than witnesses who make a low-confidence ID, are nevertheless overconfident (i.e., their IDs are more error prone than their confidence would suggest). As noted earlier, in a recent amicus brief, the APA used a calibration study to make the case that high-confidence IDs (made with 90% to 100% confidence) are only 60% correct. Thus, our new message is that, in terms of suspect ID accuracy, the evidence suggests that this is not true. Instead, the available evidence suggests that high-confidence suspect ID accuracy exceeds 95% correct whether the retention interval is short or long. These new findings are important because it seems likely that eyewitness IDs made from a lineup with high confidence are the cases that are most likely to be investigated by the police and/or referred for prosecution. It is therefore important to know if the relevant research has established that such IDs are untrustworthy even under pristine testing conditions. Until now, the consensus view has been that, indeed, IDs made with high confidence are considerably more error prone than suggested by the expressed level of confidence, even under good conditions. As it turns out, the data suggest the opposite.

A natural question to ask is whether these findings would apply in the real world if lineups were administered in the same fashion as they typically are in the laboratory (e.g., fair lineups, immediate confidence ratings, and no administrator influence). In fact, the results of a recent police department field study in which eyewitness confidence was assessed under those conditions yielded suspect ID accuracy estimates that are very similar to those reported here, including the high accuracy associated with high-confidence suspect IDs (Wixted, Mickes, Dunn, Clark, & Wells, 2016). The data in this field study were not broken down by retention interval because that information was not available, but the fact that high-confidence accuracy was estimated to be very high (∼97% correct) suggests that accuracy is high across the range of retention intervals encountered during actual police investigations over the course of a year. The reanalyses presented here suggest that there is no reason to expect that it would be otherwise because high-confidence suspect IDs remained accurate even when the retention interval was as long as 6–9 months.

The *d*' scores in the studies reanalyzed here were generally fairly high (greater than 1.6 in all conditions in which it could be computed). Thus, we still do not know what the confidence accuracy relationship would be if the retention interval were long enough that performance dropped to much lower levels. However, the *d*' scores estimated by fitting a signal-detection model to the police department field data were in the same approximate range as the *d*' scores observed here (Wixted et al., 2016). Thus, based on the available evidence, it seems reasonable to suppose that the findings reported here may be generally applicable.

Expert Opinion

In a survey of eyewitness identification experts, Kassin, Tubb, Hosch, and Memon (2001) found that 90% of the respondents agreed with the following statement: "An eyewitness's confidence is not a good predictor of his or her identification accuracy." The point is often made that expert opinion in this regard contrasts with lay opinion, according to which confidence is a good predictor of identification accuracy, and that this difference in opinion is one reason why eyewitness experts should be allowed to testify in cases involving eyewitness identification. Our findings suggest that even after a long retention interval, lay opinion may actually be closer to the truth, at least for an initial ID made from a fair lineup. Then again, lay opinion may be changing. Desmarais and Read (2011) found that, according to surveys conducted from the late 1970s through the early 2000s, lay opinion has come closer to expert opinion over time. In the late 1970s and early 1980s, only 30% of the lay public endorsed the statement from Kassin et al. (2001) about the weak confidence–accuracy relationship. By the early 2000s, 56% endorsed that proposition. Thus, it seems that the lay public is slowly becoming as pessimistic about the relationship between eyewitness confidence and accuracy as the experts appear to be.

Is confidence a good predictor of accuracy? Actually, the question is not well formed because the answer depends on whether the question is being asked about confidence measured in an initial and unbiased test of memory or confidence assessed under other conditions (e.g., the ID that occurs in a court of law). On an initial test of memory, and assuming good testing conditions (e.g., fair lineup, no administrator influence, immediate confidence ratings, etc.), our analyses indicate that confidence is a remarkably good predictor of accuracy even after a retention interval as long as 6-to-9 months. Moreover, high-confidence suspect IDs appear to be highly reliable. On any subsequent test of memory, however, that strong relationship may break down, and high-confidence IDs may become less reliable. For example, it is well known that eyewitness memory can be contaminated by the very act of testing memory (e.g. Brewer & Wells, 2006; Sporer et al. (1995)). In addition, it is well known that feedback from police can inflate confidence for later IDs of the suspect (e.g., Luus & Wells, 1994; Wells & Bradfield, 1998). Furthermore, numerous studies demonstrate the general increase in the confidence held by an eyewitness over time, from witnessing the event to trial, as a function of a range of other social and cognitive variables and events (e.g., Leippe & Eisenstadt, 2007; Shaw, McClure, & Dykstra, 2007). These considerations suggest that future surveys should break the confidence–accuracy survey item into two parts, one that asks about an initial test of memory and the other that asks about a later test of memory in which the eyewitness is again asked to identify the same individual. We would disagree with the Kassin, Tubb, Hosch and Memon (2001) survey item as applied to an initial test of memory from an unbiased lineup but agree with it as apply to any later test of memory (including, and perhaps especially, the one that occurs in a court of law).

Theoretical Considerations

Our findings are inconsistent with the optimality hypothesis, according to which confidence is a less reliable indicator of accuracy under poor memory conditions (Bothwell et al., 1987; Deffenbacher, 1980). Yet, high-confidence accuracy remained extremely reliable even as memory conditions deteriorated, so much so that high-confidence suspect ID accuracy was close to 100% correct whether the retention interval was as short as 1 week or as long as 9 months. In addition, the difference between low-confidence and high-confidence IDs was, if anything, greater following a long retention interval.

How is it that the participants in the retention interval studies we reviewed were able to maintain such a high level of high-confidence accuracy even as overall performance diminished? Mickes, Hwe, Wais, and Wixted (2011) argued that this question is related to an even more fundamental question: How is it that participants – with no training whatsoever by the experimenter – are able to effectively use a confidence scale in the first place? The basic idea proposed by Mickes et al. (2011) is that our everyday experiences teach us to become experts in expressing the appropriate level of confidence in a face recognition decision (see Lindsay, Read, & Sharma, 1998, and Lindsay, Nilsen, & Read, 2000, for related ideas). More specifically, the learning process that may account for our apparent expertise in expressing appropriate levels of confidence theoretically involves differential error feedback. Young children, for example, might have a tendency to inappropriately express high confidence across the board, but if they do, they will soon find out from others that their expressions of high confidence are often in error. Such training may eventually teach them to make effective use of their own internal sense of memory strength (reserving expressions of high confidence for cases in which the memory match signal is strong). In a similar vein, Skinner (1953) once made the following argument: “Strangely enough, it is the community which teaches the individual to ‘know himself”’ (p. 261). Skinner argued that certain aspects of mental life remain undifferentiated in the absence of explicit discrimination training. To illustrate this point, he used the example of color: “Anyone who has suddenly been required to make fine color discriminations will usually agree that he now ‘sees’ colors which he had not previously ‘seen”’ (p. 260). It may be the same way with the subjective sense of memory strength. Through discrimination training involving differential error feedback, people may learn to accurately gauge the strength of their own memory match signals such that the relationship between confidence and accuracy becomes quite strong.

According to this way of thinking, the lack of an effect of a long retention interval on the accuracy of high-confidence identifications makes sense. If people learn to accurately express high confidence in a recognition decision only when their internal memory match signal is strong, and if a long retention interval weakens that signal, on average, then one would expect to find fewer expressions of high confidence as the retention interval increases. However, one would not expect to find that participants suddenly ignore their training history and start expressing high confidence despite the fact that the memory match signal is weak. Instead, one would expect to find that only those few participants who still experience a strong memory match signal despite the long retention interval will express high confidence. Those high-confidence IDs should remain accurate, and the data suggest that they do.

These considerations may also help to explain why it is relatively easy to contaminate memory following the first ID. The internal memory cues that ordinary serve us well in the course of everyday life (including on an initial lineup memory test, whether the retention interval is short or long) can sometimes be affected in ways that are not readily apparent to the witness (Dunlosky & Theide, 2013; Koriat, 2012; Leippe & Eisenstadt, 2007). For example, a witness who has been exposed to the photograph of a suspect on multiple occasions following an initial low-confidence ID (made with low confidence because the memory match signal was initially weak) may later experience a strong memory match signal (e.g., in court). If that witness fails to appreciate the effect of the intervening exposures and relies on the (usually-diagnostic) strong memory match signal, a high-confidence ID will be made in court. In effect, a source-monitoring failure will result in the witness relying on an internal memory signal that is ordinarily diagnostic but no longer is (Lindsay, 2014; Roediger & DeSoto, 2015). These theoretical considerations underscore the importance of relying only on the initial expression of confidence, prior to memory contamination.

Policy Implications

The policy implications that flow from our findings would seem to be straightforward. In agreement with recommendations that have been made for years, an initial statement of confidence made by the eyewitness as to his or her positive identification of a lineup member should always be recorded. One rationale for this recommendation in the past has been that it provides a way to track any inflation of confidence that might occur (Wells et al., 1998), but an arguably more important rationale is that eyewitness confidence on an unbiased initial test appears to be a highly reliable indicator of accuracy (Brewer & Palmer, 2010).

Although the most surprising finding from our reanalysis of prior findings is that high-confidence suspect IDs made from a lineup were highly reliable even after a 9-month retention interval, another important finding is that low-confidence IDs were much less reliable. It seems particularly important that the judicial system not overlook this fact. Most of the DNA exoneration cases involved eyewitness IDs that were made with low confidence during the initial memory test even though they were ultimately made with high confidence in court. Garrett (2011) analyzed trial materials for 161 DNA exonerees who had been misidentified by an eyewitness with high confidence in a court of law. He found that "...of these trial transcripts (92 of 161 cases), the witnesses reported that they had *not* been certain at the time of their earlier identifications" (p. 49, emphasis in original). Information about the level of eyewitness confidence at the time of the initial ID for the remaining 43% of the cases was not available. Thus, the true percentage of cases involving an initial ID made with low confidence may be considerably higher than 57%. If it were understood that (1) only the initial memory test is relevant and (2) confidence expressed on an unbiased initial memory test is a highly reliable indicator of accuracy, many of the DNA exonerees may not have been convicted in the first place (Wixted et al., 2015).

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

Appendix.

Estimating suspect ID Accuracy from a Calibration Score

Juslin et al. (1996) did not present their data in sufficient detail to directly calculate suspect ID accuracy, so we computed an estimate from the calibration data in their Figure 4. WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/) was first used to estimate C_c for each level of confidence. We then converted those scores, which included filler IDs, to scores that included only suspect IDs. Although Juslin et al. (1996) included target-present filler IDs as errors, we made the simplifying assumption that the reported scores excluded them and that the typical calibration formula shown in Eq. (3) was used. Using this equation, the conversion from C_c to suspect ID accuracy, $p(TP|SID_c)$, is straightforward. Eq. (3) is:

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

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

$$p(TP|SID_c) = \frac{C_c}{C_c + (1 - C_c)/n}$$

As an example, imagine there were 80 correct high-confidence suspect IDs from target present lineups and 80 high-confidence incorrect IDs from target-absent lineups. According to Eq. (3), C_c would equal 80/(80 + 80) = .5, but the number we want, according to Eq. (7), is 80/(80 + 80/8) = 1/(1 + 1/8) = .89. However, all we have is the reported calibration accuracy score of .50 (estimated from a figure), which, using the above formula, is converted into a suspect ID accuracy score by computing .50/[.50 + (1 − .50)/8)], which reduces to 1/(1 + 1/8) = .89. The true suspect ID accuracy score would be slightly higher because the calibration scores used in this conversion formula are slightly lower than they would be had filler IDs been excluded from the calculation of C_c, as they usually are. However, the inclusion or exclusion of target-present filler IDs usually has only a small effect, so the estimated suspect ID accuracy scores are probably not far below the true scores.

Estimating Standard Errors for Suspect ID Accuracy Scores

The standard errors associated with suspect ID accuracy scores obtained using Eq. (7) cannot be directly computed. For example, the standard error equation for binomial probabilities will not work because one term in Eq. (7) is not a frequency count but is instead a frequency count divided by lineup size. The standard errors for suspect ID accuracy scores were therefore estimated using a 10,000-trial bootstrap procedure. On each trial, the observed data from target-present lineups were randomly sampled with replacement to obtain a bootstrap sample of suspect IDs for that trial. For example, if for the observed TP data there were 150 high-confidence suspect IDs out of 500 lineups, the observed high-confidence suspect ID hit rate = 150/500 = .30. Thus, on each bootstrap trial, a high-confidence suspect ID was registered with probability .30 for each of 500 lineups (i.e., a high-confidence suspect ID would be registered approximately every third lineup, on average). The first bootstrap trial might yield 157 suspect IDs, the next bootstrap trial might yield 141 suspect IDs, and so on. Similarly, on each bootstrap trial, the observed data from target-absent lineups were randomly sampled with replacement to obtain a bootstrap sample of filler IDs for that trial. For example, if for the observed

Table A1
Raw Data and Computed Accuracy Scores from Sauer et al. (2008)

	Confidence				
	0–20	30–40	50–60	70–80	90–100
Type of ID					
$nSID_{TP}$	5	11	32	43	30
$nFID_{TP}$	6	6	25	18	12
$nFID_{TA}$	8	17	32	26	8
$\sim nSID_{TA} = nFID_{TA}/8$	1	2.125	4	3.25	1
Accuracy score					
Calibration (Eq. (2))	0.263	0.324	0.360	0.494	0.600
Calibration (Eq. (3))	0.385	0.393	0.500	0.623	0.789
Suspect ID Accuracy (Eq. (7))	0.833	0.838	0.889	0.930	0.968

TA data there were 100 high-confidence filler IDs out of 500 lineups, the observed high-confidence filler ID rate = 100/500 = .20. Thus, on each bootstrap trial, a high-confidence filler ID was registered with probability .20 for each of 500 lineups (i.e., approximately every fifth lineup yielded a high-confidence filler ID). The first bootstrap trial might yield 94 filler IDs, the next bootstrap trial might yield 101 filler IDs, and so on. After obtaining a bootstrap sample of suspect IDs and filler IDs on a given bootstrap trial, a suspect ID accuracy score was computed in exactly the same manner it was computed for the observed data using Eq. (7). Thus, for example, if there were 157 suspect IDs and 94 filler IDs on the first bootstrap trial, then suspect ID accuracy for the first bootstrap trial = 157/(157 + 94/8) = .930. Note that the bootstrap sample of 94 filler IDs was divided by lineup size (8) to estimate innocent suspect IDs from target-absent lineups. Similarly, if there were 141 suspect IDs and 101 filler IDs on the second bootstrap trial, then suspect ID accuracy for the second bootstrap trial = 141/(141 + 101/8) = .918. This process was repeated for 10,000 bootstrap trials, and the standard deviation of the 10,000 bootstrap suspect ID scores provided the estimated standard error.

The same basic approach was used to estimate the error bars in Figure 2 except that instead of computing suspect ID accuracy on each trial, we collapsed across confidence and computed the overall hit rate, false alarm rate (target-absent filler IDs divided by lineup size) and *d*'. The standard deviation of the *d'* score across 10,000 bootstrap trials provided the estimated standard error.

Calibration and Suspect ID Accuracy Scores from Sauer et al. (2008)

The raw data for choosers in Sauer et al. (2008) were taken from their Table 3 (collapsed across their Thief and Waiter conditions) and are presented in Table A1. The number of suspect IDs from target-present lineups ($nSID_{TP}$), filler IDs from target-present lineups ($nFID_{TP}$), and filler IDs from target-absent lineups ($nFID_{TA}$) were taken directly from the table, and the estimated number of suspect IDs from target-absent lineups ($\sim nSID_{TA}$) was computed by dividing filler IDs from target-absent lineups ($nFID_{TA}$) by the lineup size of 8 for each level of confidence. With these values in hand, they were simply plugged into Eqs. (2), (3) and (7) to compute the relevant accuracy scores. Eq. (2) was used by the APA in its amicus brief (American Psychological Association, 2014). Using that equation, low-confidence IDs (0–20% confident) were 26.3% correct and high-confidence IDs (90–100% confident) were 60% correct. Eq. (3) is the more commonly used calibration formula, and Eq. (7) provides suspect ID accuracy, which is the measure of interest here.

References

American Psychological Association. (2011). *Commonwealth v. Walker*. Retrieved from http://www.apa.org/about/offices/ogc/amicus/walker.aspx

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

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.

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology*, *15*, 77–96.

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.

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.

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

Deffenbacher, K. A. (2008). Estimating the impact of estimator variables on eyewitness identification: A fruitful marriage of practical problem solving and psychological theorizing. *Applied Cognitive Psychology*, *22*, 815–826.

Desmarais, S. L., & Read, J. D. (2011). After 30 years, what do we know about what jurors know? A meta-analytic review of lay knowledge regarding eyewitness factors. *Law and Human Behavior*, *35*, 200–210.

Dunlosky, J., & Thiede, K. W. (2013). Metamemory. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 283–297). New York, NY, US: Oxford University Press.

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

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.

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research: A new survey of the experts. *American Psychologist*, *56*, 405–416.

Koriat, A. (2012). The subjective confidence in one's knowledge and judgements: Some metatheoretical considerations. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of metacognition,* (pp. 213–233). New York, NY, US: Oxford University Press.

Leippe, M. R., & Eisenstadt, D. (2007). Eyewitness confidence and the confidence–accuracy relationship in memory for people. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, Vol II: Memory for people* (pp. 377–425). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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

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.

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

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.

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.

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute d', not the diagnosticity ratio. *Journal of Applied Research in Memory and Cognition*, *3*, 58–62.

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.

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: Erlbaum.

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.

Sauer, J. D., 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.

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.

Shaw, J. S., III, McClure, K. A., & Dykstra, J. A. (2007). Eyewitness confidence from the witnessed event through trial. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia's (Eds.), *Memory for events* (Vol. 1) *The handbook of eyewitness psychology* (pp. 371–397). New Jersey: Lawrence Erlbaum Associates.

Skinner, B. F. (1953). *Science and human behavior*. New York, NY: Free Press.

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.

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*, 360–376.

Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells, & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). New York, NY: Cambridge University Press.

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.

Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, *8*, 155–167.

Wells, G. L., 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.

Wetmore, S., 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.

Wixted, J. T., & Ebbesen, E. (1991). On the form of forgetting. *Psychological Science*, *2*, 409–415.

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.

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

Received 5 August 2015;
received in revised form 15 February 2016;
accepted 13 April 2016
Available online 7 June 2016