Methodological Notes on the Accuracy–Confidence Relation in Eyewitness Identifications

Gary L. Wells
University of Alberta, Edmonton
Alberta, Canada

R. C. L. Lindsay
Queen’s University, Kingston, Ontario, Canada

The important question of whether confidence and accuracy are related in eyewitness identifications involves a number of methodological considerations peculiar to the eyewitness setting. Inconsistencies in the literature are attributed in part to these methodological factors. Distinctions between four types of identification error (i.e., false identification, identification of foil in perpetrator-present lineups, identification of foil in perpetrator-absent lineups, and false rejection) and two types of accuracy (accurate identification and accurate rejection) are critical to an applied analysis. Eyewitness confidence in foil identifications, although of potential theoretical interest, should not be included in the forensically relevant calculations of confidence–accuracy relationships. Choice and nonchoice responses of witnesses should be analyzed separately in terms of whether confidence has predictive utility. Applied generalizations should take account of routine real-world events such as attorney–witness briefings and other extramemorial factors that can have robust effects on confidence but not accuracy. Attempts at generalization should also take account of the imperfect relationship between witnesses’ self-rated confidence and jurors’ ascriptions of confidence. The common practice of squaring correlations to assess predictive utility is criticized and an alternative statistic is proposed.

Does the confidence with which an eyewitness proffers identification testimony bear a useful relationship to the accuracy of the identification? This nascent issue has profound practical significance, especially in light of recent studies indicating that eyewitness confidence is a strong determinant of the credibility that people ascribe to identification testimony (e.g., Lindsay, Wells, & Rumpel, 1981; Wells, Ferguson, & Lindsay, 1981; Wells, Lindsay, & Ferguson, 1979).

The available data on the accuracy–confidence relation in eyewitnessing has been discussed recently in three separate reviews (Dffenbacher, 1980; Leippe, 1980; Wells & Murray, 1984). Although all three concluded that eyewitness confidence should not be used to infer eyewitness accuracy, the literature is divided between studies that have shown no relation and those that have shown a statistically significant relation between eyewitness confidence and accuracy. Wells and Murray reported 13 studies that found a significant positive relation and 18 studies that found no relation or a negative relation. Approximately half of these 31 studies have been published since 1981, and all but three were published since 1978, attesting to the recency of the issue.

Wells and Murray (1984) proposed that the discrepancy in results across studies is due primarily to the differences in procedural and statistical techniques used by different investigators. The purpose of this article is to describe the different methods used to address the confidence–accuracy issue, evaluate the adequacy of these methods in terms of their faithfulness to forensic ecology, and recommend a uniform series of steps for researchers to follow. In doing this, we realize some risk in appearing too definitive with regard to what is and is not forensically relevant or what series of specific procedures and statistical analyses ought to prevail in the literature.
Thus, we wish to make clear from the outset that researchers could have other purposes for their research that may justify quite different procedures and analyses. We do not intend for our recommendations, based on an applied approach to legal settings, to restrict the wide variety or possible procedural and statistical treatments of theoretical issues.

False Identification Versus Other Errors

Most eyewitness researchers now realize that they must distinguish between errors of commission and omission. Errors of commission occur when the eyewitness chooses someone other than the actual perpetrator from a lineup, whereas errors of omission occur when the witness fails to make a lineup choice (a miss) when in fact the perpetrator is present in the lineup. To equate these errors is to misconstrue their differential significance in forensic settings. Although most researchers now report separately these errors across conditions, many researchers collapse across these errors in calculating confidence-accuracy correlations.

The issue becomes more complex when considering the fact that there are different models under which police construct lineups and photo spreads. Of particular concern is the distinction between a single-suspect model and a multiple-suspect model. A single-suspect model has a number of lineup members (e.g., 6, 9, 12), but only one is a suspect, while the others are distractors or “foils.” Distractors or foils are known-innocent stand-ins; they may be police officers, persons drawn from jail cells or other individuals, but their principal characteristic is that of being known to be innocent of the crime in question. A multiple-suspect model might include no foils at all: Each lineup member is a suspect in the case. For purposes of this article, we are going to assume the operation of a single-suspect model. There are two main reasons why we are assuming the operation of a single-suspect model. First, having more than one suspect in a lineup is judged unfavorably by the courts (Sobel, 1972) and operates explicitly against recent guidelines to police (e.g., Brooks, 1983). In addition, the multiple-suspect model is viewed with extreme disfavor by eyewitness experts. Ellison and Buckhout summarized this point well when they argued that a lineup in which all members are suspect is like a multiple-choice test with no wrong answers. We have no doubts that multiple-suspect lineups are sometimes conducted, but the salient unacceptability of such practices makes us focus on the single-suspect model as the most likely model in current and future cases. (We suggest that the multiple-suspect model is best regarded as a relative of the mug-shot task, a “fishing expedition” in the absence of an actual suspect.)

Assuming the single-suspect model, the identification of a foil is an error, but its significance is different from that of identifying an innocent suspect. Unlike the identification of an innocent suspect, the identification of a foil will not result in charges against the identified person. Identification of a foil has less profound consequences (e.g., “I’m sorry Mrs. Jones, you identified Officer Krupke of the vice squad”). Although the identification of a foil has some of the consequential properties of a miss, it has distinguishable and diagnostic properties that make it different (Wells & Lindsay, 1980). Foil identifications are of two types; one can identify a foil when the perpetrator is present in the lineup or one can identify a foil when the perpetrator is absent from the lineup. This distinction is subtle but is potentially important for answering questions about confidence-accuracy relations from a forensic perspective.

There are in fact four types of error: False identification of the suspect (which can occur only when the lineup’s suspect is innocent); false rejection; foil identification type $\alpha$ (when the suspect is guilty); and foil identification type $\beta$ (when the suspect is not guilty). There are, however, only two ways to be accurate: Correct identification and correct rejection.

Critical to our ultimate recommendations for calculating confidence-accuracy relationships is a distinction between “known” errors and “unknown” errors. By known and unknown we mean that one type of error is known immediately in actual cases to have occurred, namely foil identifications. Because the ultimate question is whether or not confidence can be useful in actual cases to predict accuracy, foil identifications should not be
included when calculating the relation between confidence and accuracy. It is possible, of course, that the estimate of relation between confidence and accuracy is unaffected by this distinction. We reexamined some of our own data, however, and it appears that it might make a difference. In one study, for example, we found that foil identifications of type $\alpha$ were associated with significantly less confidence than were accurate identifications, but false identifications and foil identifications type $\beta$ were associated with equal confidence to accurate identifications (Wells, Ferguson, & Lindsay, 1981). Although we do not know if the Wells et al. results are representative, it may be that eyewitnesses who choose a foil under conditions where the perpetrator is in the lineup are different from those who choose a foil when the perpetrator is absent from the lineup. In any case, the identification of a foil is a known error, and the ability to predict this from confidence-measures is forensically irrelevant. Calculations of confidence-accuracy relations using foils can be done for theoretical or other reasons, of course, but the questions of forensic utility should remove these cases from the analyses.

Removing foil identifications from the forensic analysis leaves two types of error (false rejection and false identification) and two types of accuracy (accurate identification and accurate rejection). Once again, forensic considerations indicate that only certain comparisons are meaningful. Although it is common to calculate confidence-accuracy correlations by collapsing over these two errors and collapsing over the two types of accuracy, such a procedure is unfaithful to the ecology of the question. Specifically, the question is not whether eyewitnesses who make accurate identifications are more confident than those who incorrectly make no choice. Instead, the question is whether a witness who makes an accurate identification is more confident than one who makes a false identification. It is appropriate also to ask whether correct rejections are made more confidently than are false rejections. If researchers continue to combine the two types of accuracy and compare them to a combined false identification and false rejection, much of the forensic question is lost. It is possible, for example, for there to be greater confidence in accurate rejections than false rejections but no difference in confidence between accurate identifications and false identifications.

**On Measuring Confidence**

Because the modal outcome in the eyewitness literature shows no relation between confidence and accuracy (Wells & Murray, 1984), some researchers have questioned the adequacy with which confidence has been measured (e.g., see Egeth & McCloskey, 1984, footnote 4). Researchers are resistant to accepting null-effect conclusions (Greenwald, 1975), and there have been two suggestions about other ways to measure eyewitness confidence. First, some investigators have argued that eyewitness confidence might have more diagnostic utility if measured prior to the identification act than if measured, as typically done, after the identification act. Leippe (1980), for example, argued that noncorrespondence between accuracy and confidence may be attributable to self-perception effects (as in Bern, 1972) wherein the act of identification serves to inflate confidence by virtue of the free-choice nature of the act. Tests of the hypothesis, however, have not confirmed the notion that pre-identification measures of confidence are better than postidentification measures (Murray & Wells, 1982). In addition, Murray and Wells found strong correlations between four separate measures of eyewitness confidence, suggesting strong parallel measure reliability.

The second criticism of how researchers measured eyewitness confidence was based on ecological considerations. Wells, Lindsay, and Ferguson (1979) argued that “although self-ratings of confidence on a scale may not be indicative of eyewitness accuracy, witnesses in the criminal justice system express their confidence through their overt responses to adversary cross examination. These overt indicators of confidence [which include nonverbal cues] may well be related to accuracy” (p. 441). In fact, however, studies testing this hypothesis have found no support (e.g., Lindsay et al., 1981; Wells et al., 1981; Wells et al., 1979; Wells, Lindsay, & Tousignant, 1980). Witnesses' self-rated confidence and subject-jurors' ascriptions of confidence are correlated ($r = +.55$, Wells et al., 1979), but juror-rated
confidence is no more (and perhaps less) likely to show a relation to accuracy.

Generalizing to Noisy Environments

There are theoretical reasons to believe that there exists an entire class of real-world variables that affect eyewitness confidence without having corresponding effects on eyewitness accuracy (Leippe, 1980). Rarely are these variables represented in eyewitness experiments, but some recent studies indicate that these variables contribute substantially to noncorrespondence between eyewitness accuracy and confidence. Hastie, Landsman, and Loftus (1978), for example, found that repeated questioning of witnesses had confidence-inflating effects but resulted in no increases in accuracy. Wells et al. (1981) found that "briefing" an eyewitness (after their lineup identification) regarding the questions that they might be asked under cross-examination served to inflate confidence, and this briefing eliminated the small, but statistically significant, confidence-accuracy correlation found in control conditions. Given the prevalence of attorney–witness briefings and repeated questioning of witnesses, real-world generalizations about confidence-accuracy relations based on studies that do not include these "noise" variables should be called into question.

One recent study suggests that the typical eyewitness experiment risks overestimating confidence-accuracy correlations because of a failure to represent the anxieties witnesses experience at the time of retrieval (Murray & Wells, 1982). Specifically, although many eyewitness studies use staged crimes to generate "natural" encoding conditions, subject-witnesses are informed prior to the lineup identification that the event was staged. Murray and Wells tested the effects of knowing versus not knowing at the time of identification that a theft was staged and found little effect on eyewitness accuracy but a significant effect on the confidence–accuracy correlation. Specifically, accurate witnesses were significantly more confident than inaccurate witnesses when they knew the identification was merely an experimental test, but this difference was eliminated when the witnesses believed their identification had real-world consequences for the accused.

In summary, generalizations from laboratory eyewitness experiments designed to examine the confidence–accuracy question must consider the ways in which real-world environments are subject to greater noise. Other real-world variables that may contribute to noncorrespondence between accuracy and confidence include misleading postevent information (Loftus, 1979), which can reduce accuracy without reducing confidence; the passage of time, which rarely is as great in experiments as it is in actual cases but which may reduce memory accuracy faster than it reduces confidence; commitment to earlier testimony reports, desire to secure a conviction, and other social psychological variables that likely affect expressed confidence but are unlikely to affect the accuracy of testimony. If researchers want to generalize from laboratory to real-world settings in estimating confidence–accuracy relations, some account should be taken of these noise variables. We recommend that studies be designed to explore the influence of these variables on the confidence–accuracy relation.

Statistical Considerations

Eyewitness researchers typically have used the correlation coefficient to express their findings regarding the confidence–accuracy relation. As noted previously, such analyses should be done separately for choosers and nonchoosers, and choosers should be broken down further so as to examine the relation with foil choices removed. An advantage of correlations over some other measures is that they reflect simultaneously the statistical significance of the relation (by knowing the value of $r$ and sample size) as well as the magnitude of relationship (as expressed by $r^2$).

Unfortunately, there is a misunderstanding about correlations in the applied literature. Conventional wisdom has it that the utility of a correlation for purposes of prediction is best expressed by the square of the correlation. In fact, however, it is the value of $r$ rather than $r^2$ that better reflects the utility of prediction (see Rosenthal & Rubin, 1982). For example, if 60% of confident witnesses are accurate, whereas only 40% of nonconfident witnesses are accurate, the correlation between confidence and accuracy is .20 and
Table 1

Probability of Eyewitness Identification of Innocent and Guilty "Suspects" and Diagnosticity Ratios for Varying Levels of Confidence

<table>
<thead>
<tr>
<th>Situation: criminal present (S = C) or criminal absent (S ≠ C)</th>
<th>Confidence level</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>P(IDS</td>
<td>S = C)</td>
<td>0</td>
</tr>
<tr>
<td>P(IDS</td>
<td>S ≠ C)</td>
<td>.011</td>
</tr>
<tr>
<td>Diagnosticity ratio</td>
<td>0</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Note. The data are taken from Lindsay and Wells (1980) and Wells, Ferguson, and Lindsay (1981); N = 176 subjects, 86 identifications of suspects; IDS is an identification of the suspect from the lineup; S = C means the suspect is the criminal (guilty suspect); S ≠ C means the suspect is not the criminal (innocent suspect).

the "percentage of variance accounted for" is only 4%. Although there may be many reasons to question the utility of eyewitness accuracy, the low value of squared correlations misleadingly underestimates the utility of the confidence–accuracy relation.

In previous work, we have recommended the diagnosticity ratio (Wells & Lindsay, 1980) as a measure of the applied utility of identification information. Based on a Bayesian analysis of eyewitness identification decisions, the information gained from an eyewitness-identification procedure is jointly determined by the prior probability of guilt (the probability that suspects arrested by the police are in fact guilty) and the diagnosticity ratio (the ratio of the probability of an identification of a guilty suspect divided by the probability of an identification of an innocent suspect under similar lineup conditions).

Lindsay and Wells (1980) demonstrated that higher levels of diagnosticity were obtained under fair lineup conditions as compared to biased lineup conditions. A similar analysis can be applied to evaluate the degree to which eyewitness confidence is diagnostic of eyewitness accuracy. Identification data from two of our previous studies are combined and presented in Table 1. These data provide some evidence for a confidence–accuracy relationship. The lowest levels of confidence generate diagnosticity ratios less than 1, indicating that eyewitnesses who identified a suspect and rated their confidence as 1 or 2 on a 7-point scale were more likely to be inaccurate than accurate. However, the highest diagnosticity ratio was obtained for subjects reporting a confidence level of 3, still below the midpoint on the scale. This highly inconsistent pattern may be due to the relatively low frequency of nonconfident identifications of suspects. Over 70% of identifications of suspects in these studies were accompanied by confidence ratings above the midpoint of the confidence scale (i.e., 5 or higher out of 7). As a result, the probabilities of identifying guilty and innocent suspects are determined by only a few subjects for low levels of confidence. Under such conditions, the diagnosticity ratio is likely to be highly variable and unreliable as an indication of the quality of the obtained evidence. This does not mean that low levels of confidence are associated with eyewitness unreliability—rather, it indicates that low levels of confidence occurred too infrequently in our previous studies to permit a meaningful assessment of the value of nonconfident identifications of suspects as a source of evidence.

On the other hand, somewhat larger numbers of subjects both identified suspects and indicated that their level of confidence was 5, 6, or 7 on the 7-point scale. Because the diagnosticity ratios are greater than 1 at each of these levels of confidence, confident eyewitness identifications appear to be of some value as evidence. However, to put the value of confidence as a predictor of accuracy into proper perspective, the diagnosticity ratio for particular levels of confidence must be contrasted with the overall diagnosticity ratio as an indication of the value of the evidence obtained from the identification procedure independent of the level of confidence. The data in Table 1 reveal that the diagnosticity ratios of identifications made with confidence
are only slightly higher than the overall diagnosticity ratio. In fact, an interaction test of proportions (using Langer & Abelson's, 1972, procedure) showed no two levels of confidence having different diagnosticity ratios. As a result, knowing that a witness was very confident provided little information about accuracy in these studies. A more standardized approach is required to reduce the confusion created by the confidence-accuracy issue and the plethora of measures used to study and describe it. We recommend the format of Table 1 as an efficient method for clearly communicating the results of confidence-accuracy studies to other researchers in the field.

Summary and Conclusions

The forensic import of knowing whether or not and how well eyewitness confidence predicts eyewitness identification accuracy necessitates a close examination of the methods used by researchers to calculate the relationship. Researchers should distinguish between choice and no-choice responses of witnesses and known and unknown errors of identification. Generalizations to real-world settings must consider the fact that laboratory studies are relatively noise free with regard to the haphazard presence and absence of variables that may affect confidence but not accuracy or vice versa. Small confidence-accuracy relationships found with self-rated confidence may be even less diagnostic for jurors because (they) the jurors are imperfect in estimating witnesses' self-rated confidence. On the other hand, modest correlations should not be dismissed simply because their squared values appear trivial; measures of variance-accounted-for are poor intuitive statistics for judging the predictive utility of relationships between variables.

In general, attorneys, judges, and police assume that confidence and accuracy are correlated meaningfully and usefully in eyewitness identifications (Brigham & Wolfskiel, 1983). The U.S. Supreme Court formally has endorsed the idea of using eyewitness confidence to infer eyewitness accuracy (Neil vs. Biggers, 1972). Eyewitness researchers, however, generally argue that the relationship is weak, spurious, or nonexistent (e.g., Deffenbacher, 1980; Leippe, 1980; Wells & Murray, 1984), although the literature is far from unanimous (e.g., see Brigham, Maas, Snyder, & Spaulding, 1982). Whatever the ultimate conclusion regarding the confidence-accuracy issue, currently it represents an intriguing and large gap between intuition, the legal system, and "fireside induction" (Meehl, 1977) on the one hand and recent experimental evidence on the other. By considering the methodological notes outlined herein, a greater consistency in the literature may emerge and the literature may become more directly relevant to the forensic situation.

References


Received March 19, 1984
Revision received August 6, 1984

---

Special Call for Papers on Applied Issues in the Psychology of Aging

The Psychology and Aging journal is gearing up for its first year of publication in 1986. Manuscripts have been received in the editorial office for a number of months, but more than 50% of the original submissions have been experimental. The Editor, M. Powell Lawton, and the Associate Editor, Donald H. Kausler, wish to emphasize that Psychology and Aging will be a broad-ranging publication, and manuscripts from all areas of psychology are desired.

Papers on applied issues related to aging are encouraged. As the proposed editorial policy statement outlines:

Psychology and Aging publishes original articles on adult development and aging. Such original articles include reports of research, which may be applied, biobehavioral, clinical, educational, experimental (laboratory, field, or naturalistic studies), methodological, or psychosocial. While the emphasis is on original research investigations, occasional theoretical analyses of research issues, practical clinical problems, or policy may appear, as well as critical reviews of a content area in adult development and aging. Clinical case studies that have theoretical significance are also appropriate. Brief reports are acceptable with the author’s agreement not to submit a full report to another journal; a 75–100 word abstract plus 48-space lines of text and references constitute absolute limitations on space for such brief reports.

Manuscripts should be directed to:

M. Powell Lawton
Philadelphia Geriatric Center
5301 Old York Road
Philadelphia, Pennsylvania 19141