Journal of Experimental Psychology: Applied
2012, Vol. 18, No. 4, 361–376

1076-898X/12/$12.00 DOI: 10.1037/a0030609

Receiver Operating Characteristic Analysis of Eyewitness Memory: Comparing the Diagnostic Accuracy of Simultaneous Versus Sequential Lineups

Laura Mickes
University of California, San Diego

Heather D. Flowe
University of Leicester

John T. Wixted
University of California, San Diego

A police lineup presents a real-world signal-detection problem because there are two possible states of the world (the suspect is either innocent or guilty), some degree of information about the true state of the world is available (the eyewitness has some degree of memory for the perpetrator), and a decision is made (identifying the suspect or not). A similar state of affairs applies to diagnostic tests in medicine because, in a patient, the disease is either present or absent, a diagnostic test yields some degree of information about the true state of affairs, and a decision is made about the presence or absence of the disease. In medicine, receiver operating characteristic (ROC) analysis is the standard method for assessing diagnostic accuracy. By contrast, in the eyewitness memory literature, this powerful technique has never been used. Instead, researchers have attempted to assess the diagnostic performance of different lineup procedures using methods that cannot identify the better procedure (e.g., by computing a diagnosticity ratio). Here, we describe the basics of ROC analysis, explaining why it is needed and showing how to use it to measure the performance of different lineup procedures. To illustrate the unique advantages of this technique, we also report 3 ROC experiments that were designed to investigate the diagnostic accuracy of simultaneous versus sequential lineups. According to our findings, the sequential procedure appears to be inferior to the simultaneous procedure in discriminating between the presence versus absence of a guilty suspect in a lineup.

Keywords: eyewitness memory, ROC analysis, simultaneous lineups, sequential lineups

Using a memory test to identify the presence or absence of a perpetrator in a lineup is much like using a medical test to diagnose the presence or absence of a disease in a patient. In both cases, the relevant tests typically yield true positives (correctly identifying the perpetrator or correctly identifying the presence of a disease) and, unfortunately, false positives (incorrectly identifying an innocent suspect or incorrectly identifying the presence of a disease). To judge the performance of one diagnostic test relative to another, both outcomes need to be taken into consideration. In the field of medicine, diagnostic tests are typically evaluated in terms of these two outcomes by conducting an analysis of the receiver operating characteristic (ROC); in the field of eyewitness memory, different methods are used. The goals of this article are (a) to explain why, under conditions that often prevail in the eyewitness memory literature, ROC analysis is the only way to determine whether one lineup procedure is diagnostically superior to another; (b) to show how ROC analysis can be performed on lineup data; and (c) to report new ROC data comparing simultaneous versus sequential lineup procedures.

We begin our inquiry into these matters by defining some terms that are used throughout this article. In a typical eyewitness memory study, participants first observe an actor in the role of a perpetrator committing a staged crime; later, they attempt to identify the perpetrator from a lineup. A typical six-member lineup consists of one suspect and five foils. Some participants view a lineup in which the suspect is, in fact, the perpetrator (target-present lineups), but other participants view a lineup in which the suspect is an innocent person who resembles the perpetrator (target-absent lineups). The proportion of target-present lineups from which the guilty suspect is correctly identified (i.e., the proportion of true positives) is called the *hit rate* (HR), and the proportion of target-absent lineups from which the innocent suspect is incorrectly identified (i.e., the proportion of false positives) is called the *false alarm rate* (FAR). Because the foils in a lineup are not suspects and are therefore known to be innocent, choosing a foil is treated as the functional equivalent of not choosing anyone

Laura Mickes and John T. Wixted, Department of Psychology, University of California, San Diego; Heather D. Flowe, Department of Psychology, University of Leicester, Leicester, United Kingdom.

Supported in part by grant number SES-1155248 from the National Science Foundation.

This work is dedicated to the memory of our friend and colleague Ebbe Ebbesen. We thank Vivian Hwe and Daniel Klein for their assistance in filming the video and for their help in collecting the data, and we also thank Daniel Bajic for his assistance in computing the partial AUCs.

Correspondence concerning this article should be addressed to John T. Wixted, Department of Psychology, 0109, University of California, San Diego, La Jolla, CA 92093-0109. E-mail: jwixted@ucsd.edu

(i.e., foil choices are not counted as either hits or false alarms). Together, the HR and FAR characterize the diagnostic performance of a lineup procedure.

Ideally, when two lineup procedures are compared, one procedure would outperform the other by yielding both a higher HR and a lower FAR. Under those conditions, no special analytical technique would be needed to determine which procedure is better. However, Clark (2012) recently reviewed the effects of several commonly recommended lineup procedures that often yield a more ambiguous outcome. These recommended procedures include (a) warning the witness that the perpetrator may not be in the lineup (as opposed to not warning the witness), (b) using a sequential lineup procedure (as opposed to the standard simultaneous procedure), (c) using foils that match the suspect description (as opposed to using foils that might allow the suspect to stand out), and (d) ensuring that the lineup administrator does not influence the witness's decision (e.g., using an administrator who is blind to the suspect's identity as opposed to using a nonblind administrator). Compared with the lineup procedures they would replace, the recommended lineup procedures yield a lower FAR, which is a desirable effect, but they also tend to yield a lower HR, which is an undesirable effect.

Determining the better lineup procedure when one yields both a lower FAR and a lower HR compared with the other is not straightforward. As described in more detail below, ROC analysis can render a clear verdict under these conditions, but it has never been used for that purpose. Instead, the performance of different lineup procedures has been assessed by comparing their respective *diagnosticity ratios* (or a closely related measure of probative value). The diagnosticity ratio is equal to HR/FAR, and the higher that ratio is, the better the lineup procedure is judged to be. As an example, in a recent meta-analysis, Steblay, Dysart, and Wells (2011) reviewed the diagnostic performance of simultaneous and sequential lineups (R. C. L. Lindsay & Wells, 1985). In the *simultaneous procedure*, the members of the lineup are presented together (this has long been the standard police procedure); in the *sequential procedure*, the members are presented one at a time for individual recognition decisions, and the test effectively stops when someone is identified as the perpetrator (if the sequential test continues beyond that point, only the first identification typically counts). Steblay et al. reported that the average HR and FAR for the simultaneous lineup procedure equal 0.52 and 0.28, respectively, whereas the corresponding values for the sequential lineup procedure equal 0.44 and 0.15, respectively.[1] Thus, on average, the sequential procedure yields both a lower HR and a lower FAR—an ambiguous outcome in terms of identifying the better procedure. However, because the diagnosticity ratio for the sequential lineup procedure (0.44/0.15 = 2.93) is higher than that of the simultaneous lineup procedure (0.52/0.28 = 1.86), the sequential procedure was judged to be superior. This result is not always obtained (e.g., in a large-scale study conducted online, Gronlund, Carlson, Dailey, & Goodsell, 2009, found that the diagnosticity ratios were similar for the two procedures), but a higher diagnosticity ratio associated with the sequential procedure has been observed in a number of studies. The three other recommended lineup procedures listed above have also been judged to be superior to the lineup procedures they would replace when using the same approach (i.e., comparing diagnosticity ratios; Clark, 2012).

On the surface, the reasoning that has been used to establish which lineup procedure is superior makes sense. For example, when switching from the simultaneous to the sequential lineup, the fact that the diagnosticity ratio increases means that the percentage decrease in the HR (from 0.52 to 0.44, a 15% decrease) is less than the percentage decrease in the FAR (from 0.28 to 0.15, a 46% decrease). Intuitively, the cost seems worth the benefit. In addition, a witness's identification decision obtained using a procedure associated with a higher diagnosticity ratio is more probative of guilt (i.e., one can be more certain that an identified suspect is, in fact, the perpetrator) compared with an identification decision obtained from a procedure associated with a lower diagnosticity ratio. However, despite these apparent indicators of diagnostic superiority, an inquiry into the nature of ROC analysis, a well-established technique grounded in signal-detection theory (Green & Swets, 1966; Swets, Dawes, & Monahan, 2000), reveals that a higher diagnosticity ratio does not actually identify the superior procedure. Indeed, the field of medicine long ago abandoned the use of the diagnosticity ratio (where it is usually referred to as either the *likelihood ratio* or the *positive likelihood ratio*) and has come to instead rely almost exclusively on ROC analysis (Lusted, 1971a, 1971b; Metz, 1978). We argue that a similar change in emphasis is needed in the field of eyewitness memory, and we begin our case by describing how ROC analysis is routinely used in the field of medicine to evaluate the performance of competing diagnostic tests.

ROC Analysis in the Medical Literature

A diagnostic test—whether a lineup test or a medical test—yields four outcomes of interest, two of which (true positives and false positives) were mentioned above. These four outcomes are illustrated using a standard 2 × 2 table shown in Figure 1. In the medical literature, the term *sensitivity* is used to refer to the number of people with the disease who test positive (true positives) divided by the total number of people tested who have the disease. Thus, sensitivity is synonymous with the HR in the eyewitness memory literature. The term *specificity* is used to refer to the number of people without the disease who test negative (true negatives) divided by the total number of people tested without the disease. Thus, 1 – specificity (i.e., the proportion of people without the disease who nevertheless test positive; that is, the proportion of false positives) is synonymous with the FAR in the eyewitness memory literature.

An ROC is a plot of different sensitivity versus 1 – specificity pairs (i.e., a plot of HR vs. FAR pairs) associated with a single test. What makes ROC analysis possible is the fact that the results of a diagnostic test typically fall on a continuum. For example, a blood test might yield a result that falls on a scale that ranges from 0 to 100. Imagine that a test result greater than 50 is used to identify individuals who have a particular disease (i.e., the cutoff is set to 50) and that 63% of people who actually

[1] These values were taken from Table 3 of Steblay et al. (2011) because those data came from published studies that used adults as subjects and used a full Simultaneous/Sequential × Perpetrator-Present/Perpetrator-Absent design. For the FARs, we used the values representing "identification of designated innocent suspect," although filler identification rates for target-absent lineups taken from studies that did not designate an innocent suspect could be used to illustrate the same points.

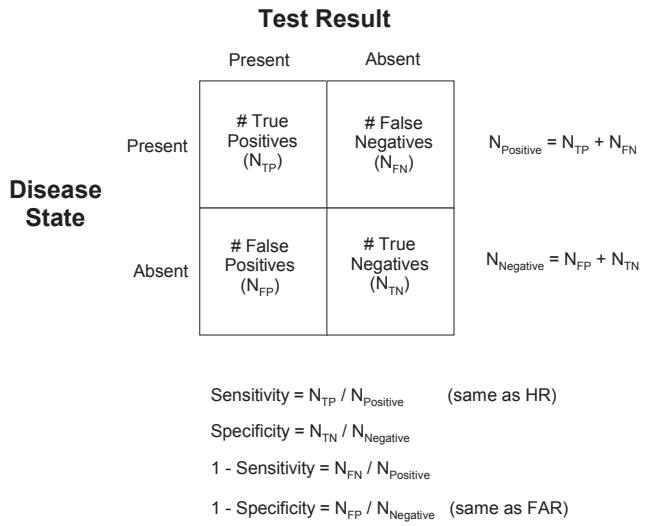

Figure 1. The four outcomes of a diagnostic test illustrated in a 2 × 2 table. For a lineup, an "absent" test result includes both lineup rejections and foil choices. HR = hit rate; FAR = false alarm rate; N_{TP} = number of true positives; N_{FN} = number of false negatives; N_{FP} = number of false positives; N_{TN} = number of true negatives; $N_{Positive}$ = number who have the disease; $N_{Negative}$ = number who do not have the disease.

have the disease yield a score greater than 50 (and are therefore correctly diagnosed as having the disease), but so do 16% of people who do not have the disease (and are therefore incorrectly diagnosed as having the disease). Thus, sensitivity = 0.63 and 1 – specificity = 0.16. This pair of values (HR = 0.63, FAR = 0.16) would correspond to 1 point—sometimes called an *operating point*—on the ROC. Additional points on the ROC could be obtained simply by choosing different cutoff values. Using a lower (i.e., more liberal) cutoff of 30 would correctly identify a higher percentage of people who actually have the disease (e.g., 80%), but it would also mistakenly identify a higher percentage of people do not have the disease (e.g., 40%). Thus, when a lower criterion is used, both sensitivity and 1 – specificity would be higher (i.e., HR = 0.80, FAR = 0.40), and this pair of values would correspond to a second operating point on the ROC. By contrast, using a higher (more conservative) cutoff of 70 would have the opposite effect, identifying fewer people who have the disease (e.g., 43%) and fewer people who do not have the disease (e.g., 4%). Thus, both sensitivity and 1 – specificity would be lower (HR = 0.43, FAR = 0.04), and this pair of values would correspond to a third operating point on the ROC.

By varying the cutoff across a range of scores produced by the diagnostic test, a researcher can obtain a range of operating points that collectively defines the diagnostic performance of the test. The ROC is simply a plot of these operating points—that is, a plot of sensitivity versus 1 – specificity values—associated with a range of cutoffs for one particular test. A hypothetical example of such an ROC is shown in Figure 2, and it illustrates an important point: The performance of a diagnostic test is not defined by a single pair of hit and false alarm rates but is instead defined by a range of hit and false alarm rates as the cutoff is varied (i.e., it is defined by its ROC). This will turn out to be a key consideration because, invariably, researchers have attempted to compare the diagnostic performance of competing lineup procedures based on a single HR-FAR pair obtained from each procedure.

The diagonal line on the ROC shown in Figure 2 indicates the performance of a test that provides no diagnostic information whatsoever because, for points that fall on that line, sensitivity = 1 – specificity (or, equivalently, HR = FAR). At the other extreme, a perfect test would yield a single point that falls at the upper left corner (where sensitivity and specificity both equal 1.00, or, equivalently, where HR = 1.00 and FAR = 0.00). In practice, most diagnostic tests yield a curvilinear trajectory of points that fall somewhere in between those two extremes, as is true of the ROC shown in Figure 2. The three ROC points discussed above are labeled "a" (HR = 0.63, FAR = 0.16), "b" (HR = 0.80, FAR = 0.40), and "c" (HR = 0.43, FAR = 0.04).

To say that one diagnostic test is more accurate than another is to say that it yields an ROC curve that falls closer to the ideal (i.e., farther above the diagonal line) than the other. The ROC performance of a test is usually measured by the *area under the curve* (AUC), which equals 0.50 for a test that yields no information (i.e., that yields ROC data along the diagonal) and 1.00 for a perfect test that yields a single point in the upper left corner (Hanley & McNeil, 1982). When two tests are compared using ROC analysis, the more accurate diagnostic test is the one that yields a higher ROC (and, therefore, a higher AUC). The test that yields a higher ROC is more accurate in the sense that it is better able to discriminate the presence versus absence of a disease compared with the other test (van Erkel & Pattynama, 1998). No single point on the ROC (i.e., no single HR-FAR pair) can adequately characterize the performance of a diagnostic test because a single point is compatible with a variety of different ROC curves that could be drawn through it. This is why it is generally not possible to effectively compare two diagnostic tests using a single HR-FAR pair generated by each one.

The ROC example shown in Figure 2 was (hypothetically) based on a lab test that yields an objective result along a continuum (e.g., a blood glucose level), but the same method can be used when the test result must be judged subjectively instead (Metz,

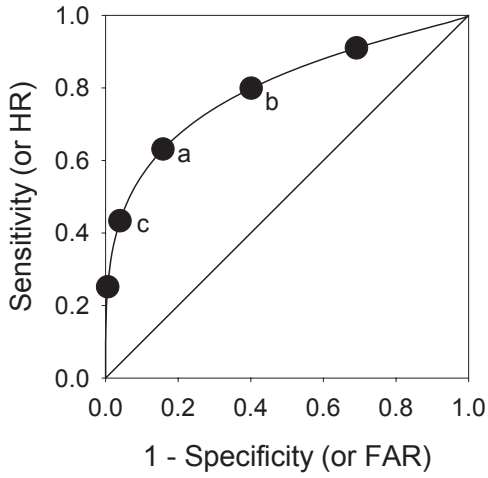

Figure 2. Hypothetical receiver operating characteristic (ROC) data. The points labeled "a," "b," and "c" represent three pairs of hit and false alarm rates associated with different cutoffs, as discussed in the text.

1978). In the field of radiology, for example, x-rays, mammograms, or magnetic resonance images are typically subjectively evaluated by radiologists for evidence of a disease. A subjective rating of confidence for the presence of a disease is often made using a numerical rating scale, and different cutoffs on that scale can be used to compute the different sensitivity versus 1 – specificity pairs that define the ROC. The use of ROC analysis in radiology is, in many ways, very similar to its potential use in the field of eyewitness memory.

Consider the case of a radiologist attempting to diagnose the presence or absence of a malignant tumor in a mammogram. The radiologist is in a role analogous to that of the eyewitness, and the mammogram is in a role analogous to that of the lineup. Pisano et al. (2005) compared the efficacy of two different diagnostic procedures, film mammography versus digital mammography (analogous to comparing two different lineup procedures) using ROC analysis. The radiologists in that study were presented with either film or digital mammograms and asked to supply confidence ratings using a 7-point scale ranging from 1 = *definitely not malignant* to 7 = *definitely malignant*. These confidence ratings provided the (semi) continuous scale from which sensitivity and specificity were computed according to different cutoffs. One pair of values for the ROC was computed by using the most conservative cutoff of 7. For this cutoff, sensitivity equals the number of participants correctly identified as having a malignancy with a confidence rating of 7 divided by the total number of participants with a malignancy (verified by a contemporaneous biopsy test or by later follow-up). Similarly, 1 – specificity equals the number of participants incorrectly identified as having a malignancy with a confidence rating of 7 divided by the total number of participants without a malignancy. This sensitivity versus 1 – specificity pair yields an operating point that falls toward the lower left of the ROC. The next pair was computed by using a slightly more liberal cutoff of 6 on the confidence scale. For this cutoff, sensitivity equals the number of participants correctly identified as having a malignancy with a confidence rating of 6 or 7 divided by the total number of participants with a malignancy. Similarly, 1 – specificity equals the number of participants incorrectly identified as having a malignancy with a confidence rating of 6 or 7 divided by the total number of participants without a malignancy. This pair is the second operating point on the ROC (i.e., the next point up and to the right). By cumulating responses starting from ever lower points on the confidence scale, a full range of pairs can be computed and then plotted to reveal the ROC for that test.

Figure 3 shows the ROC data comparing these two diagnostic methods as reported by Pisano et al. (2005). The upper left panel (Panel A) shows the overall results, and it is clear that the two procedures yield virtually the same ROC (i.e., the two procedures are equivalent in terms of diagnostic accuracy). However, the ROC data plotted in the other three panels (Panels B, C, and D) show that the use of digital mammography yields more accurate results than film mammography in three different subgroups of women. In other words, for these three subgroups, the digital ROC falls closer to the upper left corner (and has a correspondingly higher AUC) than the film ROC. A higher ROC is what serves to identify digital mammography as the more accurate procedure for these subgroups, and it is important to emphasize that this conclusion has nothing to do with the diagnosticity ratio associated with either test (a point we revisit later in this article).

Although empirical ROCs have not been reported in the eyewitness memory literature, Clark, Erickson, and Breneman (2011) recently presented theoretical ROC curves predicted by a model of eyewitness memory called WITNESS (also see Ebbesen & Flowe, 2002). ROC analysis is already widely used for theory testing in basic experimental studies of memory (e.g., Mickes, Wixted, & Wais, 2007), and there is no reason why it could not serve the same purpose in applied studies of eyewitness memory (e.g., to test the predictions of different versions of the WITNESS model). However, beyond its potential contributions to theory development, the practical (and theory-free) benefits of ROC analysis are potentially far-reaching as well.

Since the technique was introduced to medicine in the early 1970s, more than 10,000 ROC analyses have been published in that field according to a PubMed Clinical Query using the search terms *sensitivity*, *specificity*, and *ROC* (search settings: category = diagnosis, scope = narrow). Restricting the search by adding the term *radiology* yields nearly 2,000 articles in that field alone. In a tutorial overview of ROC analysis published in the journal *Clinical Chemistry*, Zweig and Campbell (1993) puzzled over the fact that clinical laboratorians had yet to use the example set by radiologists in the field of medicine:

> Of the 18 papers mentioned earlier, only 5 included ROC plots. Others had some data on sensitivity, specificity, efficiency and/or predictive value, but without ROC plotting. Why such an elegant but simple tool has been underutilized by laboratorians is a puzzle. It is widely recognized in medicine as a powerful way to represent the accuracy of a signal detection system. (p. 568)

In the years since, hundreds of such ROC analyses have been reported in *Clinical Chemistry* alone. Why ROC analysis has never been used by eyewitness memory researchers is a similar puzzle, one that we hope to solve by showing exactly how to do it and illustrating why it is necessary. Because ROC analysis involves the use of confidence ratings, we begin by briefly reviewing the complicated history of previous attempts to relate confidence and accuracy in eyewitness identification (a history that may help to explain why eyewitness memory researchers have thus far been reluctant to embrace confidence-based ROC analysis).

Confidence and Accuracy in Eyewitness Identification

In a lineup procedure, the "test result" is provided by the eyewitness. That is, after identifying someone from the lineup, an eyewitness can report some degree of confidence that the identified individual is actually the perpetrator. Thus, subjective confidence is the continuum along which the test result falls and for which different cutoffs can be set.

Confidence is a proxy for the diagnosticity of the memory signal in the mind of the eyewitness, just as confidence is a proxy for the diagnosticity of the perceptual signal in the mind of the radiologist. In each case, the more diagnostic the signal is thought to be, the higher the observer's confidence rating will be. In practice, a signal that is subjectively more diagnostic tends to be objectively more diagnostic as well (i.e., as confidence increases, so does accuracy), and that fact is what makes confidence-based ROC analysis an effective technique.

In the eyewitness memory literature, confidence-based ROC analysis has never been used to compare different lineup proce-

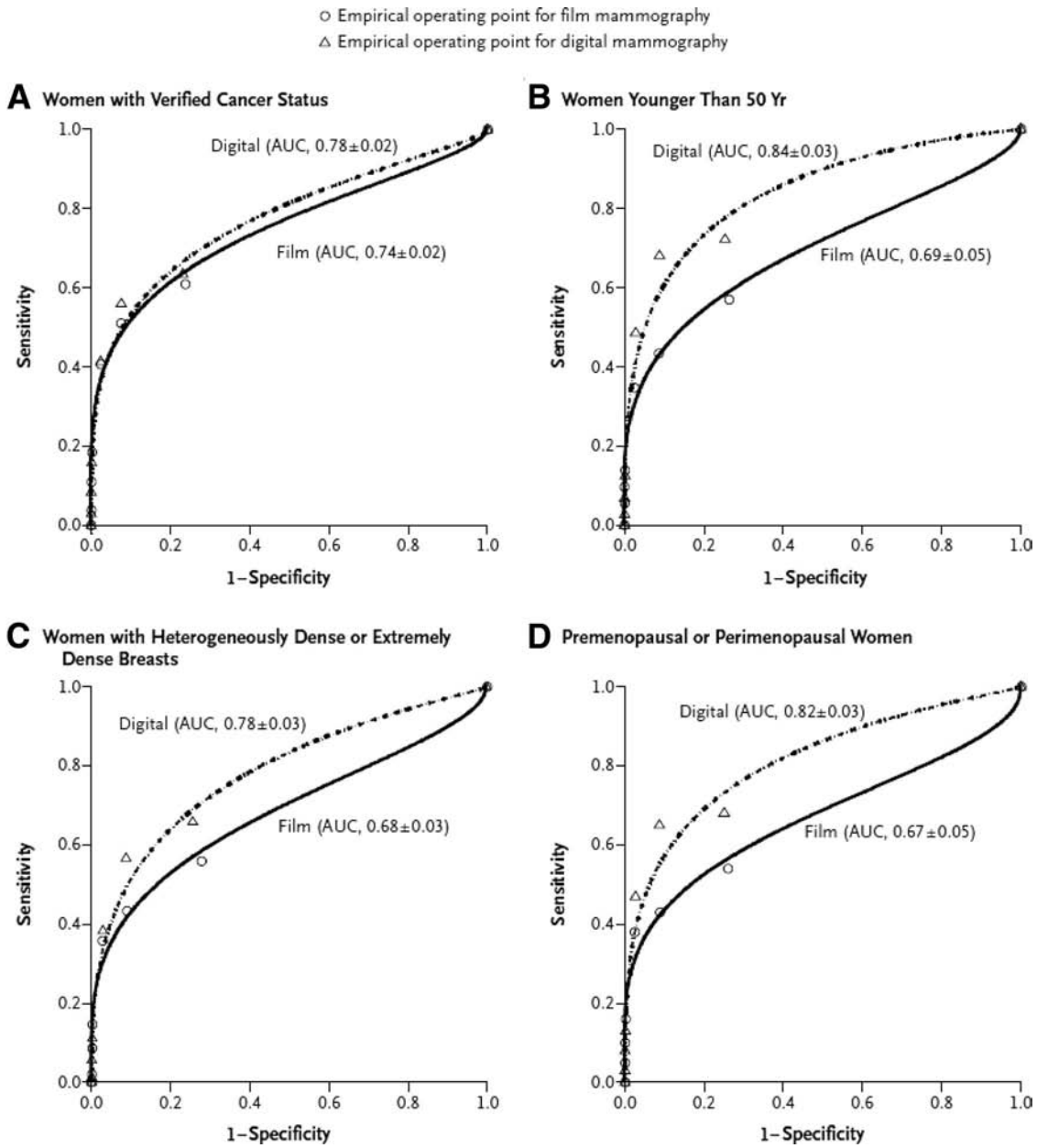

Figure 3. Confidence-based receiver operating characteristic (ROC) data reported by Pisano et al. (2005), who tested the ability of radiologists to detect a malignancy in film mammograms versus digital mammograms (AUC = area under the ROC curve). From *New England Journal of Medicine*, E. D. Pisano, C. Gatsonis, E. Hendrick, M. Yaffe, J. K. Baum, S. Acharyya,. . . M. Rebner, (2005). Diagnostic Performance of Digital Versus Film Mammography for Breast-Cancer Screening, Vol. No. 353, p. 1778. Copyright © 2005 Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

dures, perhaps because it was long thought that the relationship between confidence and accuracy is weak. For example, in one review of the literature, Wells and Murray (1984) found that the point-biserial correlation between confidence and accuracy was only .07. Largely on that basis, they concluded that ". . . the eyewitness confidence–accuracy relation is weak under good laboratory conditions and functionally useless in forensically representative settings" (p. 165). In another review of the literature, Bothwell, Deffenbacher, and Brigham (1987) found that the correlation was only .25, which is somewhat better than what Wells and Murray reported, but is still quite low. Later work showed that the correlation is modestly higher if the analysis is limited only to those who make a positive identification from a lineup (Sporer, Penrod, Read, & Cutler, 1995), but even after taking that fact into consideration, Penrod and Cutler (1995) concluded that eyewitness confidence ". . . is a weak indicator of eyewitness accuracy even

when measured at the time an identification is made and under relatively 'pristine' laboratory conditions" (p. 830).

However, when confidence ratings are taken at the time a positive identification is first made from a lineup, the relationship between confidence and accuracy is now known to be quite strong. As explained by Juslin, Olsson, and Winman (1996), the previous misunderstanding of this issue arose because a largely uninformative statistic (namely, the point-biserial correlation coefficient) was used to measure the relationship of interest. The problem with that approach is that the correlation coefficient can be very low even when the relationship between confidence and accuracy is very strong. The correlation coefficient can be high under certain conditions (e.g., D. S. Lindsay, Read, & Sharma, 1998), but the point is that it is often very low even when confidence and accuracy are strongly related. Thus, it is the statistic itself, not the relationship between confidence and accuracy, that is problematic (Roediger, Wixted, & DeSoto, 2012).

Using a more appropriate calibration approach, recent laboratory studies have shown that there is in fact a strong relationship between confidence and accuracy when eyewitnesses identify someone from a lineup (e.g., Brewer, Keast, & Rishworth, 2002; Brewer & Wells, 2006; Juslin et al., 1996). For example, Brewer and Wells (2006) summarized their findings on the confidence–accuracy (CA) relationship as follows: "Despite the modest CA correlations, plotting confidence against proportion correct for choosers clearly indicated a positive relationship between confidence and accuracy for both sets of stimulus materials under all experimental conditions" (p. 22). They also found that the CA relation was noticeably weaker for those who did not identify someone from a lineup—that is, for nonchoosers (cf. Sporer et al., 1995). A similar asymmetry between positive and negative recognition decisions is also observed in standard list-learning studies of memory (e.g., Mickes, Hwe, Wais, & Wixted, 2011, Figure 5a). Thus, there no longer appears to be a wide gulf separating basic list-learning studies of memory (which, for decades, have found a strong relationship between confidence in positive recognition decisions and the accuracy of those decisions) and studies of eyewitness memory (which now report a similar result). Even so, lingering doubts about this issue—doubts that originated from early studies using the point-biserial correlation coefficient—may account for the reluctance of eyewitness memory researchers to embrace confidence-based ROC analysis.

How to Construct an Eyewitness Memory ROC

The kind of data needed to perform ROC analysis were reported in the study discussed above by Brewer and Wells (2006). They used the simultaneous lineup procedure,[2] and witnesses made confidence judgments using a 100-point confidence scale, with ratings of 100% indicating *absolute certainty* that the identified individual was the perpetrator and ratings of 1% indicating only *slight confidence* that the identified individual was the perpetrator. We used the data provided by choosers (i.e., by those who made a positive identification from the lineup) to construct two empirical lineup ROCs.

Figure 4A shows the ROC computed from the "Thief Lineups" data reported in Table 9 of Brewer and Wells (2006). The HR and FAR pair plotted at the lower left of the ROC was computed by treating suspect identifications as "hits" or "false alarms" only if they were made with a confidence of 90% or higher. Of 600 lineups containing a guilty suspect (target-present lineups), only 54 identifications of the suspect were made with that level of confidence. Thus, the high-confidence HR = 54/600 = 0.09. Of 600 lineups that did not contain a guilty suspect (target-absent lineups), only 10 incorrect identifications were made with that high level of confidence. In actual police lineups, incorrect identifications of an innocent suspect matter far more than incorrect identifications of a foil. However, in the study by Brewer and Wells, the target-absent lineups did not include a particular member who was designated as the suspect, and no member looked more like the perpetrator than

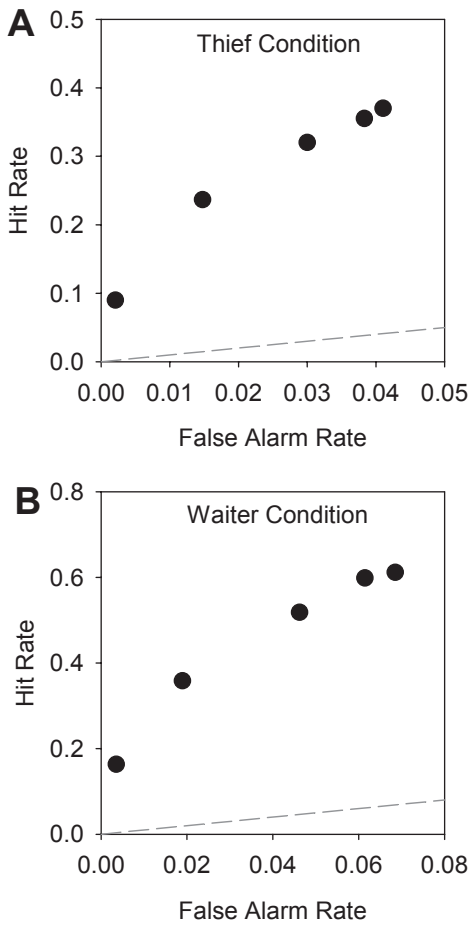

Figure 4. Confidence-based receiver operating characteristics constructed from the confidence data reported in Table 9 of Brewer and Wells (2006) for the "thief" condition (A) and "waiter" condition (B). The dashed line represents the diagonal line of chance performance. A simultaneous lineup procedure was used in this experiment.

[2] Studies in the applied literature that investigate the relationship between confidence and accuracy often plot calibration curves. Calibration curves are distinct from ROC analysis. For example, calibration curves from two conditions can be identical even though the same two conditions yield different confidence-based ROC curves. Similarly, two conditions that yield different calibration curves can yield data that fall on the same confidence-based ROC (this would happen if responding in one condition were more conservative than responding in another across all levels of confidence).

any other member. To obtain an estimate of what the FAR would be if an innocent suspect had been designated, we divided the number of incorrect identifications made to target-absent lineups by the total number of faces in the lineup (eight in this case), as Brewer and Wells also did to compute confidence-specific diagnosticity ratios. Thus, the high-confidence FAR = (10/8)/600 = 0.002. The remaining points on the ROC were computed by using ever lower cutoff values on the confidence scale. That is, the next pair of hit and false alarm rates was computed by treating as "hits" or "false alarms" only those identifications made with a confidence rating of 70% or higher; the next point was based on identifications made with a confidence rating of 50% or higher, and so on. Figure 4B shows the ROC computed in similar fashion from the "Waiter Lineups" data reported in Table 9 of Brewer and Wells.

Lineup ROCs look somewhat different from diagnostic ROCs in the medical literature because the HR and FAR do not each span the range from 0 to 1. For example, a diagnostic test in medicine that uses the most liberal cutoff will have a HR of 1.0 and a FAR of 1.0 (i.e., everyone tested will be diagnosed as having the disease), but the highest FAR for a lineup will be lower. Consider, for example, an eight-member lineup consisting of one suspect and seven foils, as in Brewer and Wells (2006). In a fair lineup involving an innocent suspect (one who does not look more like the perpetrator than the other seven members of the lineup), the maximum FAR—which would be obtained if participants used such a liberal confidence cutoff that they always identified someone from a target-absent lineup—would be 1/8, or 0.125. Thus, unlike the ROC data shown in Figures 2 and 3, in which the FAR on the *x*-axis ranges from 0 to 1, the FAR for a (fair) eight-member lineup will range from only 0.0 to 0.125. In addition, unless memory is perfect, the HR will be less than 1.0 even if participants always identify someone from a target-present lineup. Generally speaking, a lineup ROC looks like a truncated version of the ROCs shown in Figures 2 and 3 (cf. Clark et al., 2011).

Despite the unique features of eyewitness ROCs, the logic of ROC analysis for diagnosing the presence or absence of a perpetrator in a lineup is the same as the logic of ROC analysis for diagnosing the presence or absence of a disease in a patient. In both cases, the more accurate diagnostic test—that is, the test that is better able to discriminate guilty suspects from innocent suspects—is the one that yields an ROC curve that falls farther above the diagonal line of chance performance and closer to the upper left corner.

In a typical study of eyewitness memory, each participant supplies only a single recognition decision. Thus, the eyewitness ROC represents data pooled over individuals. This differs from ROC analysis based on standard laboratory studies of memory for lists of stimuli, where each participant can supply hundreds of recognition decisions. Under those conditions, ROC analysis can be performed separately for each individual. However, because the legal system usually deals with individual witnesses who supply only a single recognition decision, a pooled ROC represents an appropriate level of analysis.

Whereas ROC analysis can be used to determine whether one procedure is diagnostically superior to another, the methods that have been used for decades in the eyewitness memory literature to make this determination cannot do so (even in principle). The typical method is to compute a *single* HR-FAR pair for each procedure and then to compare their respective diagnosticity ratios (HR/FAR) to identify the better procedure. As indicated earlier, in the medical literature, the diagnosticity ratio is often referred to as the likelihood ratio. In their informative tutorial/review article, Zweig and Campbell (1993) noted that "The likelihood ratio is not a particularly good tool for assessing test performance or for comparing test performance" (p. 571). We next explain why it is not a particularly good tool for assessing test performance in the eyewitness memory domain either, and why it even has the potential to mistakenly identify the inferior procedure as being superior.

The Problem With Comparing Single HR-FAR Pairs

The ROC data shown in Figure 4 demonstrate that any lineup procedure (in this case, the simultaneous lineup procedure) is associated with a wide range of HR and FAR pairs, not with a single HR-FAR pair. The data in Table 1 further show that the HR and FAR pairs plotted in Figure 4 are associated with a correspondingly wide range of diagnosticity ratios. Thus, it is not possible to characterize the performance of a lineup procedure using a single diagnosticity ratio, yet this is precisely what the field has been trying to do since the sequential procedure was first introduced by R. C. L. Lindsay and Wells (1985).

Table 1 reveals a clear empirical trend that also happens to be typical of standard (i.e., nonlineup) recognition memory data. Specifically, as confidence in a positive identification increases, the diagnosticity ratio increases as well (which is simply another way of saying that as confidence increases, accuracy increases as well). Although this trend has long been known to be true of recognition memory tested using a list of words (e.g., Stretch & Wixted, 1998), Clark et al. (2011) predicted that the same trend would likely be true of recognition memory tested using a lineup (based on simulated data generated by their WITNESS model). The data in Table 1 show that the predicted trend is, in fact, true of memory tested using a simultaneous lineup.

In some ways, this result should not be surprising. After all, every recognition decision is based on some confidence scale. Minimally (and typically), a 2-point confidence scale is used, although it is not usually conceptualized as such, and numerical ratings are not typically recorded. In a typical lineup investigation, participants have two response options (either identifying someone

Table 1
Receiver Operating Characteristic Data and Corresponding Diagnosticity Ratios Computed From Confidence Ratings Reported in Table 9 of Brewer and Wells (2006)

Confidence cutoff (%)	HR	FAR	DR
Thief condition			
<90	0.090	0.002	43.2
<70	0.237	0.015	16.0
<50	0.320	0.030	10.7
<30	0.355	0.038	9.3
<0	0.370	0.041	9.0
Waiter condition			
<90	0.163	0.004	46.1
<70	0.358	0.019	18.9
<50	0.518	0.046	11.2
<30	0.598	0.061	9.7
<0	0.612	0.069	8.9

Note. HR = hit rate; FAR = false alarm rate; DR = diagnosticity ratio.

from the lineup or not), and those two options can be conceptualized on a numerical confidence scale as follows: 1 = *No, I am not confident enough to identify anyone from the lineup as the perpetrator*, and 2 = *Yes, I am confident enough to identify lineup Member 3 as the perpetrator*. In that sense, every recognition decision involves an expression of confidence. Moreover, accuracy is higher when the effective confidence rating is 2 (i.e., when an identification is made) than it would be if the witness were asked to guess who the perpetrator might be when the effective confidence rating is 1 (i.e., when no identification is made). The data in Table 1 simply show that the same CA trend is evident when a more fine-grained confidence scale is used.

To say that the diagnosticity ratio increases as confidence increases (as shown in Table 1) is to say that the diagnosticity ratio increases as responding becomes more conservative. These are two ways of describing the same relationship because to adopt a more conservative decision rule is to require a more diagnostic memory signal before identifying someone from the lineup, and when a more diagnostic memory signal is used to identify someone from the lineup, the decision is made with higher confidence. Similarly, for two different lineup procedures, if Lineup Procedure A yields more conservative responding than Lineup Procedure B, it means that a witness exposed to Lineup Procedure A requires a more diagnostic memory signal before identifying someone (i.e., before declaring "Yes, I am confident enough to identify lineup Member 3 as the perpetrator") compared with a witness exposed to Lineup Procedure B. If so, then all else being equal, the diagnosticity ratio should be higher for Lineup Procedure A compared with Lineup Procedure B (just as, within a lineup procedure, the diagnosticity ratio is higher for decisions made with high confidence compared with decisions made with low confidence).

Drawing on signal-detection theory to conceptualize lineup memory performance, Ebbesen and Flowe (2002) argued that the lower hit and false alarm rate associated with the sequential procedure compared with the simultaneous procedure may reflect nothing more than a conservative shift in the decision criterion. Other researchers have presented evidence in support of this idea (e.g., Gronlund et al., 2009; Meissner, Tredoux, Parker, & MacLin, 2005). Recently, Steblay et al. (2011) agreed with other researchers that the sequential lineup procedure may yield more conservative responding than the simultaneous lineup procedure, but they argued that the sequential procedure is nevertheless superior because it is also associated with a higher diagnosticity ratio. In other words, the higher diagnosticity ratio associated with the sequential procedure was the basis for declaring a *sequential superiority effect*. The problem with this line of reasoning is that more conservative responding and a higher diagnosticity ratio are two sides of the same coin. Thus, a higher diagnosticity ratio does not provide an additional piece of information that goes beyond the observation that a particular lineup procedure induces more conservative responding (and it certainly does not establish the superiority of that procedure).

When comparing the performance of two lineup procedures, instead of computing their respective diagnosticity ratios, a more informative strategy would be to determine which procedure is better able to discriminate innocent from guilty suspects in a lineup. Although the diagnosticity ratio computed from a single pair of hit and false alarm rates offers no useful information in that regard, an alternative strategy might be to use the HR-FAR pairs from each procedure to compute d' scores instead. Conceptually, comparing d' for different lineup procedures makes much more sense than comparing their respective diagnosticity ratios. The reason is that d' is a measure of the ability to distinguish between the two states of the world (in this case, between innocent and guilty suspects in a lineup). If one lineup procedure yields a higher d' than the other, then that lineup procedure would be the superior of the two. If the two lineup procedures instead yielded the same d', then neither would be superior to the other.

Although conceptually on the right track, a problem with this approach is that a d' score is nothing more than a theoretical proxy for the full ROC. In other words, theoretical assumptions are used to estimate from a single HR-FAR pair what the rest of the ROC would look like had the ROC data actually been collected. Using this approach, Palmer and Brewer (2012) recently fit a theoretical signal-detection model to single pairs of hit and false alarm rates produced by simultaneous and sequential lineups in previously published studies. They found that both procedures yield approximately the same d', from which one might infer that the two procedures would yield approximately equivalent ROCs (contrary to the idea that there is a sequential superiority effect) and that they differ only in that the sequential procedure yields more conservative responding.

The analysis reported by Palmer and Brewer (2012) represents an important step forward in that it used signal-detection concepts to clearly separate accuracy—that is, how far an ROC falls above the diagonal line of chance performance—from response bias—that is, where a point falls on an ROC (Ebbesen & Flowe, 2002; Meissner et al., 2005). However, actually performing ROC analysis would be a far better way to investigate this issue because it is not dependent on detailed theoretical assumptions, whereas the analysis reported by Palmer and Brewer is dependent on numerous minimally tested assumptions that eyewitness memory researchers are free to dispute (e.g., the target and lure distributions are assumed to be Gaussian in form and to have the same variance, an *integration* decision rule is assumed instead of an *independent observations* decision rule, the decision rule is assumed to be the same for both lineup procedures, etc.). These theoretical assumptions are necessary because each study reviewed by Palmer and Brewer reported only a single pair of hit and false alarm rates for each of the two lineup procedures. Under those conditions, a theory is needed to estimate what the rest of the ROC would look like had it actually been measured. In other words, the theory does the work of specifying a hypothetical ROC curve that passes through the single HR-FAR pair that was empirically measured. However, there is no reason to rely on a theory to estimate the ROC when the ROC itself can be plotted using confidence ratings supplied by the participants. In fact, the reason why ROC analysis has become the standard method of comparing diagnostic procedures in the medical field is precisely because it provides the sought-after information without relying on debatable theoretical assumptions.

These same considerations apply to the A' statistic, which can also be computed from a single pair of hit and false alarm rates. This measure of discrimination is similar to d', but it is often considered to be superior because it is ostensibly nonparametric

and is therefore theory-free. However, like d', A' also relies on detailed parametric assumptions. The difference is that the theoretical assumptions underlying d' are clearly specified (even if they are debatable in the lineup situation), whereas the theoretical assumptions underlying A' are usually left unspecified. The fact that the theoretical assumptions on which the (parametric) A' measure depends are typically hidden from view does not mean that they do not exist. Not only does A' depend on parametric theoretical assumptions, those assumptions appear to be implausible once they are spelled out (Macmillan & Creelman, 1996; see also Verde, Macmillan, & Rotello, 2006, Figure 2).

Instead of relying on theoretical assumptions, a better approach to evaluating the diagnostic accuracy of competing lineup procedures is to compute a full range of HR-FAR pairs for each (so that their ROCs can be directly compared). In the case of simultaneous versus sequential lineups, such an analysis might actually validate the sequential superiority effect. We illustrate that possibility in Figure 5A by presenting the aggregate simultaneous and sequential data from the meta-analysis reported by Steblay et al. (2011) as if the sequential data fall on a higher ROC than the simultaneous data. Although the higher diagnosticity ratio associated with the sequential procedure does not establish its superiority, Figure 5A shows that ROC analysis might nevertheless do so. Alternatively, ROC analysis might reveal that the two points from the meta-analysis reported by Steblay et al. fall on the same ROC (as illustrated in Figure 5B), in which case the data would show that the sequential procedure induces more conservative responding than the simultaneous procedure (Ebbesen & Flowe, 2002; Gronlund et al., 2009; Meissner et al., 2005). But there is also a third possibility, one that has not yet been considered in the eyewitness memory literature. As shown in Figure 5C, the same two data points that have been interpreted as supporting a sequential superiority effect (see Figure 5A) and that others have interpreted as simply reflecting more conservative responding (see Figure 5B) are in fact also compatible with a *simultaneous* superiority effect (see Figure 5C), and the truth of the matter cannot be known in the absence of empirical ROC analysis.

Because the existing evidence does not indicate whether one lineup procedure is diagnostically superior to the other, we ran three experiments involving simultaneous and sequential lineups and then compared them using ROC analysis. Although we focus on simultaneous versus sequential lineups, we emphasize that the ROC method is generally applicable, and its use is essential for determining whether one lineup procedure is diagnostically superior to another when the two procedures yield hit and false alarm rates that differ in the same direction. In the experiments reported next, participants viewed a simulated crime (the theft of a laptop computer from an unoccupied office) and were then randomly assigned to a lineup condition (simultaneous or sequential). The lineups had six members (Experiments 1A and 1B) or eight members (Experiment 2), and each participant viewed only one lineup. The perpetrator was present in the lineups viewed by a random half the participants (target-present lineups) and was not present in the lineups viewed by the other half (target-absent lineups). Any participant who identified someone from the lineup (simultaneous or sequential) was asked to supply a confidence rating; if the simultaneous lineup as a whole was rejected, or if an individual was rejected in the sequential lineup, a confidence rating was made for that decision as well.

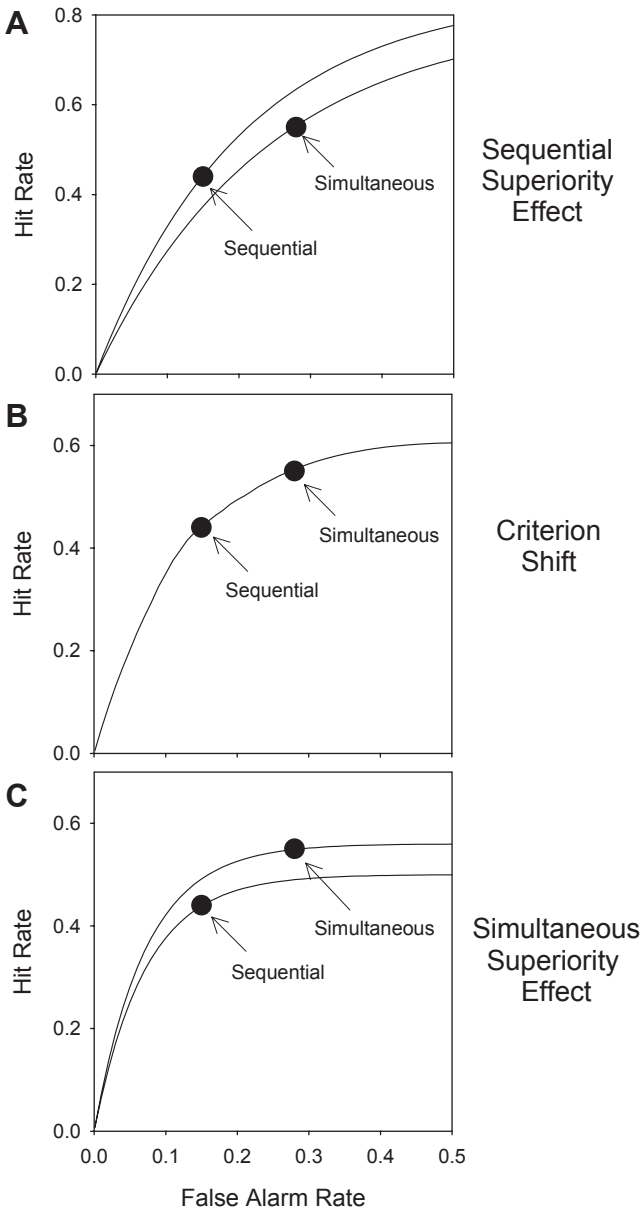

Figure 5. Average hit and false alarm rate data for simultaneous and sequential lineup procedures taken from Table 3 of the meta-analysis reported by Steblay et al. (2011). The two pairs of hit and false alarm rates might fall on different receiver operating characteristics (ROCs), with the sequential procedure yielding the higher ROC (A). This pattern would indicate a sequential superiority effect in terms of diagnostic accuracy. Alternatively, the same two points might fall on the same ROC (B), a result that would support the criterion shift interpretation. Finally, the same two points might fall on different ROCs, with the simultaneous procedure yielding the higher ROC (C). This pattern would indicate a simultaneous superiority effect in terms of diagnostic accuracy.

Experiment 1

The first experiment was run twice, once in a laboratory using undergraduate subjects and once again over the Internet using a somewhat more diverse group of participants. Because the two

experiments were otherwise procedurally identical, we describe them together and refer to them as Experiment 1a and Experiment 1b, respectively.

Method

Participants. The participants were 598 undergraduates at the University of California, San Diego (Experiment 1a) and 631 individuals from around the world who completed the task over the Internet (Experiment 1b). The demographic characteristics of the participants are presented in Table 2. The undergraduate students received course credit for their participation; the online participants received $0.30 each for their participation. All participants were treated in accordance with the ethical standards of the American Psychological Association. Experiment 1a was conducted at the University of California, San Diego and was approved by the University of California, San Diego Institutional Review Board. Experiment 1b was coordinated from the University of Leicester (using Amazon Mechanical Turk) and was approved by its institutional review board.

Materials.

Video. We recorded a short video of a 22-year-old White male (the perpetrator) walking past an unoccupied office. The door to the office was open, and a laptop computer could be plainly seen sitting on a desktop. After walking past the open door in the video, the perpetrator backs up and enters the office. He inspects the laptop, closes it, picks it up off the desk, and walks toward the door to leave. As he is leaving the office, he looks both ways along an adjacent hallway (apparently to ensure that no one is looking) and then hurries away with the laptop computer. A viewer of the video gets a clear look at the face of the perpetrator as he is scanning the hallway before leaving. A photograph of the perpetrator's face was the target in all target-present lineups.

Foils. All of the foils in target-present lineups were White males who matched the description of the perpetrator. A separate group of 22 participants watched the video and then completed a form listing the perpetrator's physical attributes, including gender, eye color, hair color, ethnicity, height, and weight. We then entered the range of values for each of these attributes (and an age range of 20 to 30 years) into the Florida Department of Corrections Offender Network database (http://www.dc.state.fl.us/AppCommon/) to retrieve description-matched photographs. A large number of matching photographs was retrieved, and, for each participant, five photographs were randomly selected to serve as foils. The position of the target in the target-present lineups was randomly determined for each participant in both the simultaneous and sequential conditions.

Target-absent lineups were constructed in a manner identical to that of target-present lineups except that all six members were foils who matched the description of the perpetrator. Thus, there was no designated innocent suspect in the target-absent lineups.

Procedure. The participants were randomly assigned to one of the four lineup conditions (simultaneous target-present condition, simultaneous target-absent condition, sequential target-present condition, and sequential target-absent condition). In all cases, participants were instructed to watch the video closely because they would have to answer questions about it afterward. The video was followed by a distractor task involving 10 anagrams of U.S. states so that performance could not be based on working memory (following Gronlund et al., 2009). After completing the anagrams, the participants were told that they would view a lineup that may or may not contain the perpetrator from the video. The simultaneous lineup participants were then shown a six-member lineup. Participants were instructed to select the person they thought was the perpetrator or to choose "not present" if they thought the perpetrator was not in the lineup.

The sequential lineup participants were instructed that each person in the lineup would be presented one at a time. As they viewed a photograph, they were to decide whether or not that person was the perpetrator. They were instructed that they would make a decision for each member of the lineup, but that only their first "yes" choice would count. No mention was made of how many lineup members would be viewed, although all members of the lineup were shown. The sequential condition could have been run with a "stop rule," whereby participants are informed that once they identify a face, no further photographs would be shown (see McQuiston-Surrett, Malpass, & Tredoux, 2006, for a review). We elected to use the above protocol instead, as it is in keeping with the majority of eyewitness memory studies.

After making their single decision for the simultaneous lineup (a positive decision identifying someone or a negative decision rejecting the lineup), participants rated their confidence. For a positive decision, a confidence rating was made using a 100-point scale (where 10 = *guessing* and 100 = *absolutely certain that the identified individual is the perpetrator*). For a negative decision, participants expressed their confidence in that decision using the same scale (where 10 = *guessing* and 100 = *absolutely certain the perpetrator is not in the lineup*). For the sequential lineup, participants supplied confidence ratings in this manner for each of the six photographs in the lineup even though they had been told that only their first identification (with a confidence rating of 10 or more) would count.

Table 2
Demographic Information for Participants in Experiments 1a, 1b, and 2

Demographic	Experiment 1a (*n*)	Experiment 1b (*n*)	Experiment 2 (*n*)
Gender			
Male	155	398	354
Female	443	233	202
Age (years)			
18–24	583	208	182
25–32	12	240	213
33–40	3	90	94
41–50	0	51	40
51–60	0	26	12
<60	0	13	10
No answer	0	3	5
Ethnicity			
Latin/Hispanic	77	8	9
Black/African	8	12	10
White/European	110	130	148
Asian/Indian	337	428	350
Native American	0	0	3
Other/unknown	66	53	36

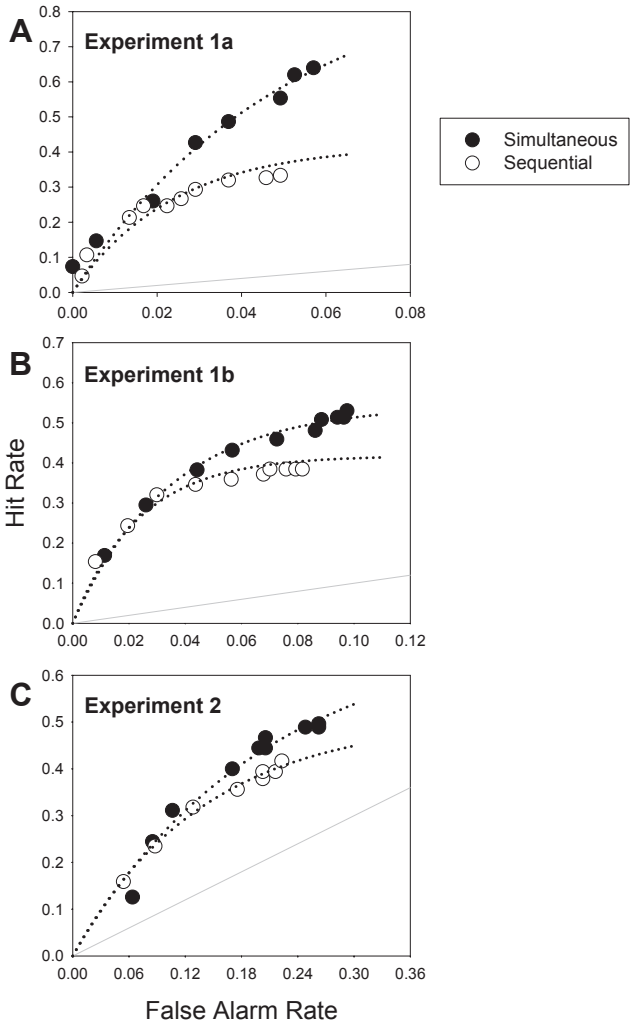

Figure 6. Confidence-based receiver operating characteristics (ROCs) from three experiments in which memory for a perpetrator in a simulated crime was tested using either a simultaneous lineup procedure (filled symbols) or a sequential lineup procedure (open symbols). (A) The participants in Experiment 1a were undergraduates tested in a laboratory, and fair lineups were used. (B) The participants in Experiment 1b were tested over the Internet, but the procedure was otherwise identical to that of Experiment 1a. (C) The participants in Experiment 2 were also tested over the Internet, but the procedure differed from that of Experiment 1 in that target-absent lineups included an innocent suspect who more closely resembled the perpetrator than the foils did.

Results

The confidence-based ROC data were analyzed in the manner described earlier for the confidence data taken from Brewer and Wells (2006). Figure 6A shows the ROC data from the laboratory study involving undergraduates from the University of California, San Diego (Experiment 1a), and Figure 6B shows the ROC data from those who participated over the Internet (Experiment 1b). The results are visually similar in that the sequential ROC is, if anything, inferior to the simultaneous ROC. No further analysis is needed to appreciate the fact that these data weigh against the notion of a sequential superiority effect.

Is the apparent simultaneous superiority effect significant? To address this question, we computed AUC values for each lineup procedure. However, unlike standard ROC analysis involving the full range of hit and false alarm rates from 0 to 1, partial ROC analysis is appropriate here. Partial AUC values were computed and compared using the statistical package pROC (Robin et al., 2011). Instead of computing the full ROC, this program computes the AUC over a partial range, which is appropriate here because the range of FARs for lineup-based ROCs extends from 0 to a value less than 1. For each ROC analysis, we selected a FAR range from 0 to q, where q was set to a value slightly greater than the maximum FAR obtained for the simultaneous ROCs.

For the lab data from Experiment 1a (see Figure 6A), the partial AUC for the simultaneous lineup (0.13) was significantly greater than the partial AUC sequential ROC (0.09), $D = 2.02$, $p < .05$. D is defined as (AUC1 – AUC2)/s, where s is the standard error of the difference between the two AUCs estimated by the bootstrap method (with the number of bootstraps set to 10,000). For the online data from Experiment 1b (see Figure 6B), the partial AUC for the simultaneous lineup (0.22) was again greater than the partial AUC sequential ROC (0.20), but the difference was not significant, $D = 0.70$. Thus, a simultaneous superiority effect was statistically supported in the first case only.

We also conducted several additional analyses to see whether we could find evidence of a sequential superiority effect. For example, we reanalyzed the sequential data by counting as correct any suspect identification that was made from a target-present lineup even if it was not the first choice (so long as the subsequent identification of the suspect was made with a higher level of confidence than an earlier identification of a foil). This had virtually no effect on the data from Experiment 1a (i.e., a simultaneous superiority effect was still evident), but it largely eliminated the visual difference between the simultaneous and sequential ROCs in Experiment 1b. Still, there was no evidence of a sequential superiority effect. We also tried eliminating all participants from the sequential target-present condition in which the suspect appeared in the first position, but this had only small effects on the ROC data.

Table 3 shows the hit and false alarm rates associated with the ROC data plotted in Figures 6A (Experiment 1a) and 6B (Experiment 1b), along with their corresponding diagnosticity ratios. Although the ratios are somewhat variable, they show once again that the diagnosticity ratio increases as responding becomes more conservative. This result again illustrates why the diagnosticity ratio is of no help in determining which lineup procedure is superior.

Experiment 2

The first experiment was designed to have fair lineups in the sense that no foil in either the simultaneous or sequential target-absent lineups was intentionally more similar to the perpetrator than any other foil. Instead, all of the foils were selected to match the description of the perpetrator. Thus, one foil would be more similar to the perpetrator than another because of chance only. In Experiment 2, we compared simultaneous and sequential lineups using target-absent lineups that were intentionally unfair in that an innocent suspect who more closely resembled the perpetrator than the foils was designated. We used unfair lineups because prior evidence has raised the possibility that the sequential procedure may be particularly useful under those conditions (Carlson, Gron-

Table 3
Receiver Operating Characteristic Data (Hit and False Alarm Rates) and Corresponding Diagnosticity Ratios for Experiments 1a, 1b, and 2

	Simultaneous lineup			Sequential lineup		
Confidence	HR	FAR	DR	HR	FAR	DR
Experiment 1a						
100	0.073	0.000	>100	0.047	0.002	23.5
90	0.147	0.006	24.5	0.107	0.003	35.7
80	0.260	0.019	13.7	0.213	0.013	16.4
70	0.427	0.029	14.7	0.247	0.017	14.5
60	0.487	0.037	13.2	0.247	0.022	11.2
50	0.553	0.049	11.3	0.267	0.026	10.3
40	0.620	0.053	11.7	0.293	0.029	10.1
30	0.640	0.057	11.2	0.320	0.037	8.6
20	0.640	0.057	11.2	0.327	0.046	7.1
10	0.640	0.057	11.2	0.333	0.049	6.8
Experiment 1b						
100	0.169	0.011	15.4	0.154	0.008	19.3
90	0.295	0.026	11.3	0.244	0.020	12.2
80	0.383	0.044	8.7	0.321	0.030	10.7
70	***0.432***	***0.057***	7.6	0.346	0.044	7.9
60	0.459	0.073	6.3	**0.359**	**0.056**	6.4
50	0.481	0.086	5.6	0.372	0.068	5.5
40	0.508	0.088	5.8	0.385	0.07	5.5
30	0.514	0.094	5.5	0.385	0.076	5.1
20	0.514	0.096	5.4	0.385	0.079	4.9
10	**0.530**	**0.098**	5.4	0.385	0.082	4.7
Experiment 2						
100	0.126	0.064	2.0	0.159	0.054	2.9
90	0.244	0.085	2.9	0.159	0.054	2.9
80	0.311	0.106	2.9	0.159	0.054	2.9
70	0.400	0.170	2.4	0.235	0.088	2.7
60	0.444	0.199	2.2	0.318	0.128	2.5
50	0.444	0.206	2.2	0.356	0.176	2.0
40	0.467	0.206	2.3	0.379	0.203	1.9
30	0.489	0.248	2.0	0.394	0.203	1.9
20	0.489	0.262	1.9	0.394	0.216	1.8
10	0.496	0.262	1.9	0.417	0.223	1.9

Note. Values in bold and bold-italic type are discussed in the text. HR = hit rate; FAR = false alarm rate; DR = diagnosticity ratio.

lund, & Clark, 2008; R. C. L. Lindsay et al., 1991). Thus, we set out to answer the question of whether a sequential superiority effect would be observed according to ROC analysis when unfair lineups are used.

Method

Participants. The participants were 556 individuals from around the world who completed the task over the Internet and received $0.25 each for their participation. The demographic characteristics of the participants are presented in Table 2. The experiment was orchestrated from the University of Leicester and was approved by its institutional review board.

Materials.

Video. The video was the same as that used in Experiment 1.

Foils. The foils were selected in the same manner as in Experiment 1 with one exception. Each target-absent lineup for both the simultaneous and sequential conditions included a photograph of a designated innocent suspect who resembled the perpetrator. The innocent suspect was, in fact, an altered photo of the perpetrator himself. The photo was altered using Photoshop by changing the hair color, skin tone, nose shape, and face shape. The position of the innocent look-alike in the target-absent lineup was randomly determined for each participant. Both target-present and target-absent lineups had eight members each.

Procedure. The procedure was identical to that used in Experiment 1.

Results and Discussion

As expected, the suspect was chosen in target-absent lineups far more often than the other foils were chosen. In the simultaneous lineups, the FAR (number of suspect choices divided by the number of target-absent lineups) was 0.26, whereas the foil FAR (number of foil choices divided by the number of target-absent lineups and then divided again by 5 because these choices were distributed across five foils) was 0.10. In the sequential lineups, the FA was 0.22, whereas the foil FAR rate was 0.10. Thus, this was an unfair lineup.

The confidence-based ROC data were analyzed as before and are shown in Figure 6C. As might be expected, performance was considerably worse in Experiment 2 (with much higher FARs) compared with Experiment 1b, but, once again, no sequential superiority effect was observed. Instead, the two ROCs appear quite similar, but a slight (nonsignificant) advantage for the simultaneous procedure is still apparent. Thus, even when an unfair lineup was used, there was still no evidence for a sequential superiority effect. Table 3 shows the hit and false alarm rates associated with the ROC data plotted in Figure 6C (Experiment 2), along with their corresponding diagnosticity ratios. Once again, the diagnosticity ratios increase as responding becomes more conservative (i.e., as confidence increases), but their absolute values are much lower than in Experiments 1a and 1b.

The Diagnosticity Ratio does not identify the more accurate lineup procedure. Although the difference between the partial AUCs for the simultaneous and sequential lineups was significant only in Experiment 1a, if anything, all three experiments point to a sequential *inferiority* effect, which is the opposite of what has been often concluded in the past. This might seem like a contradiction with prior findings, but the fact that the sequential procedure may be diagnostically inferior to the simultaneous procedure is not necessarily incompatible with prior research suggesting that the sequential procedure yields a higher diagnosticity ratio than the simultaneous procedure. Consider, for example, the ROC data from Experiment 1b shown in Table 3. Imagine that, in the absence of confidence ratings, responding was more conservative for the sequential procedure than for the simultaneous procedure. More specifically, imagine that when the simultaneous procedure is used, participants make an identification when their confidence that the individual is the perpetrator is at least 10% (i.e., a liberal criterion is used) but that when the sequential procedure is used, participants do not make an identification unless their confidence that the individual is the perpetrator is at least 60% (i.e., a conservative criterion is used). As shown in Table 3, for the simultaneous procedure, this would yield hit and false alarm rates of 0.530 and 0.098, respectively, and a diagnosticity ratio of 5.4. These hit and false alarm rate values are shown in boldface type in Table 3. For the sequential procedure, it would yield hit and false alarm rates of 0.359 and 0.056, respectively, and a diagnosticity ratio of

6.4. These hit and false alarm rate values are also shown in boldface type in Table 3.

Under these conditions, the single pair of hit and false alarm rates obtained from each procedure would exhibit the hallmarks of what has been termed the *sequential superiority effect*. Specifically, the percentage drop in the HR associated with switching from the simultaneous to the sequential procedure (from 0.530 to 0.359, a 32% decrease) is smaller than the percentage drop in the FAR (from 0.098 to 0.056, a 43% decrease). In addition, because the diagnosticity ratio is higher for the sequential procedure (6.4 for the sequential procedure, 5.4 for the simultaneous procedure), a suspect identified using the sequential procedure would be more likely to be the perpetrator than a suspect identified using the simultaneous procedure (which is another way of saying that the sequential procedure would yield information with higher probative value). Nevertheless, the corresponding ROC data shown in Figure 6B indicate that this state of affairs can arise even when the sequential procedure is the inferior diagnostic procedure.

It might be argued that the lower FAR associated with the sequential procedure (0.056 in this example) is worth the cost of a lower HR no matter what. If so, then one might still prefer the sequential procedure even if it happens to yield data that fall on a lower ROC. However, if a FAR of 0.056 is the goal, a better solution would be to use the simultaneous lineup procedure in conjunction with a more conservative decision rule. An example of a more conservative decision rule would be to count identification decisions using the simultaneous procedure only if they were made with a confidence level of 70% or more. In that case, as shown in Table 3, the hit and false alarm rates for the simultaneous procedure would be 0.432 and 0.057, respectively (these values are shown in bold-italic type in Table 3), and the diagnosticity ratio would be 7.6. This is a better outcome than can be achieved using the sequential procedure with a similar FAR. Thus, there is never any reason to use a diagnostically inferior lineup procedure. If a higher diagnosticity ratio (or a lower FAR) is desired, it can always be achieved by using the diagnostically superior procedure in conjunction with a more conservative decision criterion. This could be achieved either by accepting only eyewitness identifications made with a relatively high level of confidence or by using lineup instructions that require eyewitnesses to adopt a more conservative decision rule before choosing someone from a lineup (Wixted & Mickes, 2012).

The Diagnosticity Ratio does not identify the optimal HR versus FAR trade off. These considerations raise an important question: Once the diagnostically superior lineup procedure is identified using ROC analysis, what level of confidence yields the appropriate trade off between the HR and the FAR? Does the diagnosticity ratio offer any guidance on this issue? In fact, é the diagnosticity ratio is not useful for this purpose either. It does not identify the optimal HR-FAR trade-off point because, as shown in Tables 1 and 3, the diagnosticity ratio continues to increase as responding becomes ever more conservative. Using the diagnosticity ratio as a guide, one would always conclude that responding should be as conservative as possible, no matter how low the HR and FAR might be.

If a higher diagnosticity ratio, per se, is not the goal, what, then, is the goal of decision making in the eyewitness memory domain? Perhaps the most rational decision-making goal is to maximize overall *expected value*, in which case the optimal operating point on the ROC would be the one that comes closest to achieving that goal (Clark, 2012). The considerations involved in determining the value-maximizing operating point on the ROC were worked out long ago (Green & Swets, 1966, pp. 20–23) and are widely discussed in the medical literature (e.g., Halpern, Albert, Krieger, Metz, & Maidment, 1996; Lusted, 1971a; Metz, 1978; Swets, 1979; Zweig & Campbell, 1993). The optimal operating point on the ROC is determined by two variables: (1) the prevalence, or base rate, of the "signal" in the population (e.g., the prevalence of the disease among people tested, or the prevalence of a guilty suspect in police lineups), and (2) the relative cost of the four outcomes of a diagnostic test (i.e., true positives, false positives, true negatives, and false negatives).

These two variables combine to determine how conservative or how liberal the decision criterion should be to maximize expected value. In other words, these two variables determine the (criterion) level of eyewitness confidence associated with the optimal operating point on the ROC.

The equation that determines the optimal decision criterion is as follows (Green & Swets, 1966; Swets, 1979):

$$\beta = \frac{P(innocent)}{P(guilty)} \cdot \frac{V_{TN} + C_{FP}}{V_{TP} + C_{FN}}, \quad (1)$$

where β is the optimal decision criterion expressed as a likelihood ratio, $P(innocent)$ is the prior probability that a police lineup contains an innocent suspect, $P(guilty)$ is the prior probability that the lineup instead contains a guilty suspect and is equal to 1 – $P(innocent)$, and V and C represent the values and costs associated with true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP). Note that Equation 1 ignores the cost of foil choices by assuming them to be negligible. To the extent that one believes that foil choices are also costly (e.g., Steblay et al., 2011), additional terms would need to be added to Equation 1 to reflect those costs.

How should β (the decision criterion) be interpreted? Although its precise meaning is somewhat involved, its essence is easy to understand. The larger β is, according to this equation, the more conservative the criterion needs to be—that is, the higher confidence should be on the 0-to-100 scale before making an identification—to maximize expected value. By contrast, the smaller β is, the more liberal the criterion should be—that is, the lower confidence should be on the 0-to-100 scale—to maximize expected value. The key to determining β (and thus the key to identifying the optimal operating point on the ROC) involves specifying the two ratios on the right side of Equation 1. Unfortunately, behavioral science cannot provide much in the way of useful information when it comes to computing these ratios.

With regard to the first ratio, $P(innocent)/P(guilty)$, the question is this: What are the prior odds that a police lineup contains an innocent suspect? As those odds increase, a more conservative criterion (yielding a relatively low HR and FAR) would be needed to maximize expected value. As the odds decrease, a more liberal criterion (yielding a relatively high HR and FAR) would be needed instead. However, it is not possible to specify the prior odds with any degree of precision. Brewer et al. (2002) obtained expert opinion about this issue from police officers "who combined considerable experience in detective work with a formal university education (e.g., psychology, law)" (p. 47). These police officers

argued that *P*(*innocent*) was unlikely to exceed 0.10 because lineups are constructed only when investigating officers have strong grounds for believing that the suspect is, in fact, the person who committed the crime. According to that estimate, the prior odds that the lineup contains an innocent suspect (i.e., the first ratio in the above equation) would be 0.10/0.90, or 0.11. Other estimates provided by applied psychologists range as high as *P*(*innocent*) = 0.50 (Brewer & Wells, 2006), in which case the prior odds would be 0.50/0.50 = 1.0. To maximize expected value, the first estimate calls for a more liberal setting of the confidence criterion than the second. The problem is that the base rate information that is needed for determining whether a more conservative or a more liberal criterion should be used is not available, and the base rate value may differ across jurisdictions. Indeed, it is hard to know how such information could be obtained because there is no "gold standard" for lineups (i.e., no way to determine, across the full range of police lineups, the proportion that contain an innocent suspect). By contrast, in medicine, a gold standard often does exist in the form of a biopsy, which can be used to gather information about the base rate of a disease in the population.

What about the second ratio, $(V_{TN} + C_{FP})/(V_{TP} + C_{FN})$? The costs involved in mistakenly identifying an innocent suspect (C_{FP}) are obviously high and include both economic and moral considerations; the costs associated with freeing a guilty suspect (C_{FN}) do as well. All else being equal, the higher the cost associated with identifying an innocent suspect relative to the cost of not identifying a guilty suspect, the larger the ratio and the more conservative the criterion would need to be to maximize expected value. However, as with information about base rates, these are considerations about which scientific research does not have much to say, and opinions about these relative costs will undoubtedly differ across well-meaning individuals. Under such conditions, the opinions of policymakers matter more than the opinion of scientists. What scientists can do (and what policymakers cannot do) is to establish which lineup procedures are diagnostically more accurate using ROC analysis. Our own data suggest that when the comparison is between the simultaneous and sequential lineup procedures, the simultaneous procedure might be more accurate than the sequential procedure (a possibility that has not been considered before).

General Discussion

Going forward, ROC analysis can and should be used on a routine basis to determine whether one lineup procedure is diagnostically more accurate than the other (Wixted & Mickes, 2012). Plotting the ROC for different lineup procedures—and, therefore, determining whether one procedure is better able to discriminate innocent from guilty suspects in a lineup—is something that clearly falls within the purview of behavioral science. By contrast, scientific research will have less to say about the optimal operating point on the ROC because the information needed to identify that point is either not available (e.g., information about base rates) or depends to a large extent on subjective values (e.g., the relative costs of hits and false alarms and, perhaps, foil choices). Although it seems reasonable to educate policymakers about how the critical variables interact (captured by Equation 1), any decision about the optimal HR-FAR trade off should, in our view, be left primarily to them.

Clark (2012) recently delved into the complexities associated with weighing the costs and benefits associated with lineup procedures that yield hit and false alarm rates that differ in the same direction (also see Ebbesen & Flowe, 2002). We have built on his observations by drawing a sharp distinction between (a) determining the diagnostic accuracy of different lineup procedures based on ROC analysis and (b) determining the value associated with different HR-FAR pairs on the ROC curve associated with the more accurate lineup procedure. By focusing on the diagnosticity ratio associated with a single HR-FAR pair, eyewitness memory researchers have not clearly distinguished between these two issues. Moreover, once those issues are distinguished, it becomes apparent that the diagnosticity ratio does not usefully inform either one.[3] Thus, in our view, the practice of relying on the diagnosticity ratio to determine the better lineup procedure should be abandoned (Wixted & Mickes, 2012), as it has been in the medical field.

Are there concerns about using confidence ratings to construct an ROC? One concern might be that confidence ratings are largely uninformative because of the low CA correlation as measured by the point-biserial correlation coefficient. However, the size of that correlation is not particularly informative (Juslin et al., 1996), and, as noted by Wells, Olson, and Charman (2002), ". . . it is probably more forensically valid to use calibration and overconfidence/underconfidence measures rather than correlations" (p. 152). Using a calibration approach, Brewer and Wells (2006) reported that ". . . confidence assessments obtained immediately after a positive identification can provide a useful guide for investigators about the likely accuracy of an identification" (p. 11). Moreover, U.S. Department of Justice guidelines recommend that confidence ratings be taken by police officers immediately after an identification is made from a lineup (Technical Working Group for Eyewitness Evidence, 1999). Such confidence ratings are used to inform subsequent investigative activity and are also used in courts of law. Thus, in practice, confidence ratings are already used to distinguish between the presence or absence of a guilty suspect in a lineup, and confidence-based ROC analyses are relevant to that widespread practice.

The earlier use of the point-biserial correlation coefficient to argue that the relationship between confidence and accuracy is weak is not unlike the later use of the diagnosticity ratio to argue that the sequential procedure is superior to the simultaneous procedure. In both cases, the use of an inappropriate statistic led to what appears to be an erroneous—but, unfortunately, widely publicized—conclusion. With regard to confidence and accuracy, recent studies using a more appropriate calibration approach have shown that the relationship between them is not weak but is instead often quite strong (even when the correlation coefficient is low). With regard to simultaneous versus sequential lineups, the sequential procedure is not superior to the simultaneous procedure; instead, it may be the other way around. In the many previous studies that reported a single pair of hit and false alarm rates for each lineup procedure, there is evidence that the sequential procedure

[3] Although it cannot be used to identify the better lineup procedure, the diagnosticity ratio can be used to compute the posterior odds of guilt because, according to Bayes's rule, the *posterior* odds of guilt = the *prior* odds of guilt times the diagnosticity ratio. In Brewer and Wells (2006), for example, half of the lineups contained a guilty suspect and half did not, so the prior odds of guilt were 1/1 = 1. Thus, the diagnosticity ratios shown in Table 1 also provide the posterior odds of guilt (i.e., the odds that, for each level of confidence, a positively identified suspect is guilty).

induces more conservative responding and, perhaps, a higher diagnosticity ratio compared with the simultaneous procedure (Steblay et al., 2011). However, such findings have no bearing on which procedure is diagnostically superior. By contrast, using a more appropriate ROC approach, we found evidence for a simultaneous superiority effect. Similarly, based on a reanalysis of data they had previously published, Gronlund et al. (in press) also recently reported ROC evidence for a simultaneous superiority effect (not a sequential superiority effect). Thus, the initial ROC-based evidence from two different laboratories suggests that switching from the simultaneous lineup procedure to the sequential lineup procedure may be moving in the wrong direction. Only time will tell whether this ends up being the typical empirical result.

We emphasize that we are not arguing that the procedural recommendations associated with what Clark (2012) refers to as the "reform movement" are necessarily wrong. In some cases, the suggested reform strongly appeals to one's sense of fair play even if it is not necessarily called for on the basis of scientific research (e.g., the suspect should not stand out from the foils in an obvious way). The point, instead, is that, based on the existing evidence (consisting of single pairs of hit and false alarm rates associated with different procedures), one cannot know—yet one should know—whether the recommended procedures are diagnostically more or less accurate than the alternative procedures they would replace. In the case of simultaneous versus sequential lineups, our ROC data raise the possibility that the sequential procedure may be the diagnostically inferior procedure. Obviously, more work is needed to determine whether the same conclusion holds under a wider range of conditions (e.g., involving longer retention intervals, different foil selection procedures, different types of crimes, multiple "laps" through the sequential lineup, etc.). However, until such time as it is established that the sequential lineup procedure typically yields a higher ROC than the simultaneous lineup procedure (contrary to what our data and other recent data suggest), it seems prudent for police departments that have not already switched to the sequential procedure to refrain from doing so.

Future research may sometimes show that two competing lineup procedures yield different points on the same ROC (i.e., in which case, diagnostic accuracy would be the same for both procedures). It is easy to imagine, for example, that this is the result that would be observed when comparing different sets of instructions, one that warns the participant that the perpetrator may not be in the lineup and one that includes no such warning. If the different instructions yield points that fall on the same ROC (indicating that one set of instructions induced more conservative responding than the other), then the debate over which instruction should be used would be all about choosing the optimal trade-off point on the ROC. But if that turns out to be the case, then the optimal trade off would largely be a function of unknown base rates combined with the values and beliefs of the recommenders and would not follow directly from the outcome of scientific research.

Although we have focused on the practical utility of ROC analysis, the potential for advancing theory should be reiterated. Indeed, ROC analysis is widely used to differentiate between competing theories of recognition memory in the experimental psychology literature (e.g., Hautus, Macmillan, & Rotello, 2008; Heathcote, 2003; Mickes, Johnson, & Wixted, 2010). However, as Clark et al. (2011) observed, theoretical accounts of eyewitness memory are in short supply. Such theories are needed, in part because they could suggest ways to create lineup procedures that are diagnostically more accurate (i.e., procedures that do more than simply induce more conservative responding) compared with existing procedures. That theories of eyewitness memory can be useful to applied researchers should already be evident in the fact that the WITNESS model anticipated that the diagnosticity ratio would naturally increase as responding becomes more conservative (Clark et al., 2011). This possibility was unknown to other researchers who had no theoretical guidance while trying to determine which lineup procedure is better (and who settled on the diagnosticity ratio to do that). In any case, as theories of eyewitness memory become more developed, ROC analysis offers an important way to test their predictions.

In summary, we recommend that eyewitness memory researchers conduct confidence-based ROC analyses whenever their goal is to differentiate between two lineup procedures that yield hit and false alarm rates that differ in the same direction. As noted by Clark (2012), this pattern is common among various lineup procedures that have been recommended by eyewitness memory researchers in the past. It is somewhat sobering to realize that it is currently unknown whether the recommended procedures, no matter how sensible they might seem, are diagnostically inferior, diagnostically equivalent, or diagnostically superior to the alternative lineup procedures they would replace. Because research-based lineup reforms are already well underway in the legal system, investigating the effect of the recommended procedures on the ROC is not only important, but in our view, such investigations are urgently needed.

References

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

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

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

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 14,* 118–128. doi:10.1037/1076-898X.14.2.118

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7,* 238–259.

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35,* 364–380. doi:10.1007/s10979-010-9245-1

Ebbesen, E. B., & Flowe, H. D. (2002). *Simultaneous v. sequential lineups: What do we really know?* Retrieved from http://www2.le.ac.uk/departments/psychology/ppl/hf49/SimSeq%20Submit.pdf

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

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

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S., Wooten, A. & Graham, M. (in press). Showups versus lineups:

An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*.

Halpern, E. J., Albert, M., Krieger, A. M., Metz, C., & Maidment, A. D. (1996). Comparison of receiver operating characteristic curves on the basis of optimal operating points. *Academic Radiology, 3,* 245–253. doi:10.1016/S1076-6332(96)80451-X

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36.

Hautus, M., Macmillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review, 15,* 889–905. doi:10.3758/PBR.15.5.889

Heathcote, A. (2003). Item recognition memory and the ROC. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1210–1230. doi:10.1037/0278-7393.29.6.1210

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

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

Lindsay, R. C. L., Lea, J. A., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., & Seabrook, C. (1991). Biased lineups: Sequential presentation reduces the problem. *Journal of Applied Psychology, 76,* 796–802. doi:10.1037/0021-9010.76.6.796

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70,* 556–564. doi:10.1037/0021-9010.70.3.556

Lusted, L. B. (1971a). Decision-making studies in patient management. *New England Journal of Medicine, 284,* 416–424. doi:10.1056/NEJM197102252840805

Lusted, L. B. (1971b, March 26). Signal detectability and medical decision-making. *Science, 171,* 1217–1219. doi:10.1126/science.171.3977.1217

Macmillan, N. A., & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review, 3,* 164–170. doi:10.3758/BF03212415

McQuiston-Surrett, D. E., Malpass, R. S., & Tredoux, C. G. (2006). Sequential vs. simultaneous lineups: A review of methods, data, and theory. *Psychology, Public Policy, and Law, 12,* 137–169. doi:10.1037/1076-8971.12.2.137

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition, 33,* 783–792. doi:10.3758/BF03193074

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283–298. doi:10.1016/S0001-2998(78)80014-2

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

Mickes, L., Johnson, E. M., & Wixted, J. T. (2010). Continuous recollection vs. unitized familiarity in associative recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 843–863. doi:10.1037/a0019755

Mickes, L., Wixted, J. T., & Wais, P. (2007). A direct test of the unequal variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14,* 858–865. doi:10.3758/BF03194112

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

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

Pisano, E. D., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J. K., Acharyya, S., . . . Rebner, M. (2005). Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England Journal of Medicine, 353,* 1773–1783. doi:10.1056/NEJMoa052911

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12, 77.* doi:10.1186/1471-2105-12-77

Roediger, H. L., Wixted, J. T., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. Sinnott-Armstrong (Eds.), *Memory and law* (pp. 84–118). New York, NY: Oxford University 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. doi:10.1037/0033-2909.118.3.315

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17,* 99–139. doi:10.1037/a0021650

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1397–1410. doi:10.1037/0278-7393.24.6.1397

Swets, J. A. (1979). ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology, 14,* 109–121. doi:10.1097/00004424-197903000-00002

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26. doi:10.1111/1529-1006.001

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

van Erkel, A. R., & Pattynama, P. M. (1998). Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. *European Journal of Radiology, 27,* 88–94. doi:10.1016/S0720-048X(97)00157-5

Verde, M. E., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d′, A_z, and A′. *Perception & Psychophysics, 68,* 643–654. doi:10.3758/BF03208765

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., Olson, E. A., & Charman, S. D. (2002). The confidence of eyewitnesses in their lineup identifications from lineups. *Current Directions in Psychological Science, 11,* 151–154. doi:10.1111/1467-8721.00189

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon "probative value" and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7,* 275–278. doi:10.1177/1745691612442906

Zweig, M. H., & Campbell, G. (1993). Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39,* 561–577.

Received December 8, 2011
Revision received July 17, 2012
Accepted August 14, 2012 ■