The Field of Eyewitness Memory Should Abandon Probative Value and Embrace Receiver Operating Characteristic Analysis

John T. Wixted and Laura Mickes

Perspectives on Psychological Science 2012 7: 275
DOI: 10.1177/1745691612442906

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(3) 275–278

DOI: 10.1177/1745691612442906
http://pps.sagepub.com
SAGE

The Field of Eyewitness Memory Should Abandon Probative Value and Embrace Receiver Operating Characteristic Analysis

John T. Wixted and Laura Mickes
University of California, San Diego

Abstract

Clark (2012) highlights an important issue that has received inadequate attention in the eyewitness memory literature: lineup procedures that reduce the false identification rate (a desirable effect) often tend to reduce the correct identification rate as well (an undesirable effect). Determining which procedure is diagnostically superior under those conditions is not easy. Clark (2012) showed that the procedure with the lower false identification rate could be associated with higher overall costs to society once costs and benefits are both taken into consideration. Beyond the issue of cost, we argue that Clark's (2012) observation has far reaching implications for evaluating the diagnostic performance of a lineup procedure. Specifically, the field of eyewitness memory has attempted to differentiate between lineup procedures by using various measures of probative value (such as the diagnosticity ratio). However, contrary to intuition, probative value is not a relevant consideration. Instead, lineup procedures should be compared using receiver operating characteristic analysis, as is routinely done in other applied fields (such as radiology).

Keywords

eyewitness memory, lineup identification, false identification

The performance of a lineup procedure is typically characterized by its correct identification rate (i.e., the proportion of target-present lineups in which the guilty suspect is correctly identified) and its false identification rate (i.e., the proportion of target-absent lineups in which the innocent suspect is incorrectly identified). We will refer to these two measures as the hit rate (HR) and the false alarm rate (FAR), respectively.

Ideally, when two lineup procedures are compared, Procedure A would outperform Procedure B by yielding a lower FAR and a higher (or at least the same) HR. Under those conditions, Procedure A would obviously be better than Procedure B. However, Clark (2012) points out that several commonly recommended lineup procedures often yield a more ambiguous outcome. The recommended lineup procedures often yield both a lower FAR (a desirable effect) and a lower HR (an undesirable effect) than do the lineup procedures they would replace. Determining which of two lineup procedures is better when one yields both a lower FAR and a lower HR is not straightforward. As noted by Clark (2012), the field has generally tried to make this determination by relying on probative value (PV), but this is not a valid strategy.

PV is Not Relevant

PV refers to the likelihood that a suspect identified from a lineup is guilty, and it can be computed in several conceptually similar ways (e.g., HR/FAR, which is called *the diagnosticity ratio*). Our comment will focus on this critical issue. Clark (2012) highlighted the limitations of the PV approach, with his main point being that the lineup procedure associated with higher PV does not necessarily maximize utility once costs and benefits are taken into account. Both the guilty base rate and the utilities associated with the different decision outcomes (hits, misses, correct rejections and false alarms) must be specified in order to compute overall utility, but the guilty base rate among police lineups is unknown, and the utilities associated with different decision outcomes are subjective. These unfortunate realities cannot be circumvented by using PV in an effort to determine which lineup procedure is better in terms of overall costs and benefits.

If PV cannot determine which procedure yields higher utility, what can it do? Clark (2012) allowed for the possibility that despite its limitations, PV has something useful to offer when it comes to comparing different lineup procedures. This is an intuitively reasonable way to think because if Procedure A, with its lower HR and FAR, has higher PV than Procedure B, it means that the decrease in the FAR associated with

Corresponding Author:
John T. Wixted, Department of Psychology, University of California, San Diego, La Jolla, CA 92093
E-mail: jwixted@ucsd.edu

Procedure A is proportionally greater than the corresponding decrease in the HR. However, we suggest that the PV of a lineup procedure does not merely suffer from limitations. Instead, contrary to intuition (and contrary to established practice in the field of eyewitness memory), it is an *irrelevant* consideration when trying to decide which of two lineup procedures is better. This is true even though studies of eyewitness memory have shown that a suspect identified using a procedure with higher PV is more likely to be guilty than a suspect identified using a procedure associated with a lower PV. We argue that it is misleading to think along these lines because a lineup procedure is not associated with a single PV, so it does not make sense to say that Procedure A has higher PV than Procedure B.

A Lineup Procedure Is Characterized by a Range of PVs

The idea that a lineup procedure is associated with a single PV—one that can be meaningfully compared with the single PV associated with a different lineup procedure—is rooted in the idea that a lineup procedure can be characterized by a single HR–FAR pair. However, the fact that experimenters are in the habit of measuring only a single hit and false alarm rate pair should not be taken to mean that it is the proper way to characterize the performance of a lineup procedure.

Consider a hypothetical experiment investigating the performance characteristics of the simultaneous lineup procedure. After viewing a simulated crime, but before being presented with a lineup, witnesses in one experimental condition (the neutral condition) might be told that it is just as important to remove suspicion from an innocent suspect as it is to identify a guilty suspect (i.e., they would be told that these two outcomes are equally important). Under these conditions, the HR and FAR might be .50 and .20, respectively. Witnesses in a second condition might be told that it is far more important to remove suspicion from an innocent suspect than it is to identify a guilty suspect. Under those instructions, they would likely be more cautious about making an identification from the lineup (i.e., they would be more conservative), in which case the HR and FAR would both be lower (e.g., HR = .30, FAR = .06). Witnesses in a third condition might be told that it is far less important to remove suspicion from an innocent suspect than it is to identify a guilty suspect. Under those instructions, they would likely feel freer to make an identification (i.e., they would be more liberal), and the HR and FAR would likely both be higher (e.g., HR = .65, FAR = .50).

Note that the different prelineup instructions would not change the state of any witness's memory. Thus, on average, the state of memory would be the same across conditions. For the liberal, neutral, and conservative conditions, the HR, FAR, and corresponding PVs (where PV = HR/FAR) in this example are shown in the first three lines of Table 1. There are a few points to make about this example. First, just as no single HR–FAR pair characterizes the performance of the simultaneous lineup procedure, no single measure of PV characterizes the performance of the simultaneous lineup procedure. Second, and this is the most important point of our commentary, as responding becomes more conservative, PV steadily increases. This is not just true of this hypothetical example, it is also true of actual data in diverse fields, including experimental psychology (Stretch & Wixted, 1998) and medicine (Zweig & Campbell, 1993)."

Table 1. Hypothetical Experimental Results Using a Simultaneous Lineup Procedure

Condition	HR	FAR	PV
Simultaneous			
Liberal	0.650	0.500	1.3
Neutral	0.500	0.200	2.5
Conservative	0.300	0.060	5.0
Very conservative	0.100	0.010	10.0
Extremely conservative	0.020	0.001	20.0
Sequential			
Liberal	0.500	0.360	1.4
Neutral	0.340	0.120	2.8
Conservative	0.160	0.040	4.0
Very conservative	0.100	0.016	6.3
Extremely conservative	0.018	0.001	18.0

If maximizing PV is the goal, one might be tempted to induce even more conservative responding. Imagine, for example, adding two even more conservative conditions (using even more extreme instructions geared towards protecting the innocent)—the results might be similar to those shown in the fourth and fifth lines of Table 1.

Here, a possible problem with focusing on PV as a measure of performance and striving to maximize its value becomes apparent. As responding becomes increasingly conservative, PV increases, but the HR and FAR both approach 0. It is not obvious that the condition with the highest PV (which misses 98% of guilty suspects in this example) is the best condition in terms of balancing costs and benefits. This was Clark's (2012) main point about comparing different lineup procedures, but the same issue arises within a single lineup procedure. How liberal or conservative should responding be using a single procedure? There is no way to answer that question without knowing the guilty base rate and without specifying the utilities of the different outcomes (which involve subjective considerations). Thus, this is a question for policymakers, not scientists. However, as illustrated next, scientists do have an essential role to play in terms of characterizing the relative performance of different lineup procedures, but it is a role they have not yet played.

Receiver Operating Characteristic Analysis

Imagine running the same five instructional conditions using the sequential lineup procedure, yielding the results shown in

the bottom half of Table 1. Using hypothetical data like these, how should scientists compare the performance of the simultaneous and sequential lineups? Standard practice in the field would involve comparing a single HR–FAR pair from one procedure with a single pair from another procedure. For example, the simultaneous and sequential procedures might be compared using data from the neutral condition only (for the simultaneous condition, HR = .50, FAR = .20, and PV = 2.5; for the sequential condition, HR = .34, FAR = .12, PV = 2.8). This is the kind of situation that Clark (2012) addressed. His point was that it is not possible to determine which procedure has higher utility, which is certainly true. Others have argued that this pattern might reflect more conservative responding for the sequential procedure (Gronlund, Carlson, Dailey, & Goodsell, 2009), which is also true. Still others have argued that, even if it does reflect more conservative responding, the fact that the sequential procedure also yields higher PV effectively distinguishes between the two procedures. In fact, the higher PV associated with the sequential procedure has been dubbed the "sequential superiority effect" (Steblay, Dysart, & Wells, 2011), and jurisdictions in the United States and abroad have been encouraged to adopt the sequential method largely on this basis. However, a higher PV for the sequential method does not establish its superiority any more than a higher PV associated with the most conservative responding within either procedure establishes that biasing condition as being superior to the other biasing conditions.

Let us examine these data in a different way, by simply plotting the HR and the FAR separately for each lineup procedure (Fig. 1). The entire range of HR–FAR pairs associated with a particular lineup procedure is known as its receiver operating characteristic (ROC). The dashed diagonal line reflects chance performance, and the farther the data points fall above that line, the better the procedure is able to differentiate between innocent and guilty suspects in a lineup. For these hypothetical data, the simultaneous procedure is better able to distinguish innocent from guilty suspects than the sequential procedure.

The ROC can indicate which procedure is diagnostically superior to the other, but PV cannot. ROC data can be obtained from different biasing conditions (as in the examples above), or they can be obtained much more conveniently from confidence ratings, as is routinely done in the fields of experimental psychology and radiology. The lineup procedure that consistently yields ROC data that fall farther from the diagonal line is the diagnostically superior procedure, and it is the procedure that should be recommended by scientists to policymakers. Policymakers would then decide to encourage a liberal, neutral, or conservative response by choosing the appropriate wording for the instructions given to eyewitnesses.

If only a single HR–FAR pair is measured, as is typically done, one can attempt to estimate what the rest of the ROC would look like by computing *d′*. In essence, *d′* is a theoretical measure of the ability of participants to distinguish between innocent and guilty suspects in a lineup regardless of whether responding is liberal or conservative. Clark (2012) included *d′* as a measure of PV, but it is not a measure of PV. Instead, it is a theoretical stand-in for the empirical ROC. However, to compute *d′* from a single HR–FAR pair, one must make detailed theoretical assumptions about lineup memory. The validity of those assumptions is completely unknown (unlike in laboratory studies of list memory), so this is not an ideal solution. A far better solution would be to actually compute the full range of HR–FAR pairs associated with the lineup procedures that are being compared, which is standard practice in other applied fields (e.g., radiology). Plotting the range of HR–FAR pairs in an ROC curve is the only way to convincingly determine which lineup procedure is diagnostically superior. Thus, in our view, the time has come for the field of eyewitness memory to abandon PV and to adopt ROC analysis.

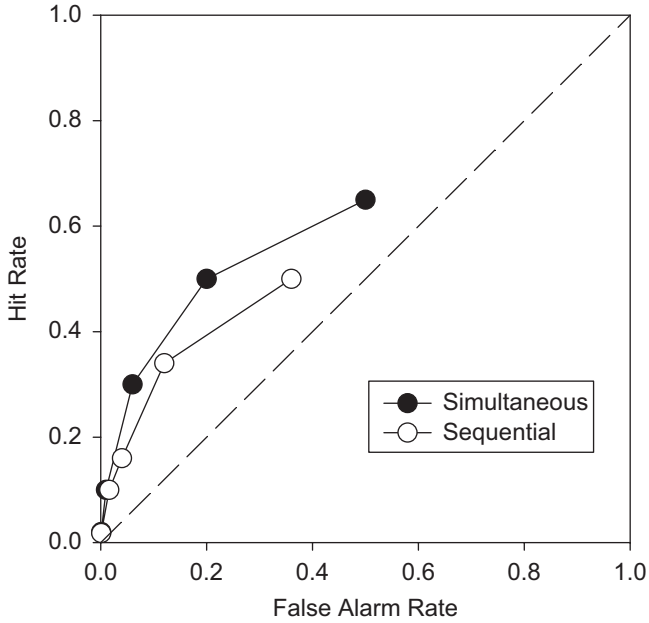

Fig. 1. Receiver operating characteristic curves of hypothetical data from a simultaneous lineup and sequential lineup. The dashed diagonal line represents chance performance. The hypothetical simultaneous lineup yielded a superior ability to discriminate between innocent and guilty suspects in a lineup (i.e., the simultaneous curve falls farther above the diagonal line).

Declaration of Conflicting Interests

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

References

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

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.

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.

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

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