ROC Analysis in Theory and Practice

John T. Wixted[1], Laura Mickes[2], Stacy A. Wetmore[2], Scott D. Gronlund[3] & Jeffrey S. Neuschatz[4]

[1]University of California, San Diego, United States

[2]Royal Holloway, University of London, England

[3]University of Oklahoma, United States

[4]The University of Alabama, Huntsville, United States

Author Note

This work was supported in part by the National Science Foundation [SES-1456571] to John T. Wixted and the Economic and Social Research Council [ES/L012642/1] to Laura Mickes and John T. Wixted. The content is solely the responsibility of the authors and does not necessarily reflect the views of the National Science Foundation or the Economic and Social Research Council.

Correspondence concerning this article should be addressed to John T. Wixted (jwixted@ucsd.edu).

Abstract

Lampinen (2016) suggested that proponents of ROC analysis may prefer that approach to the diagnosticity ratio because they are under the impression that it provides a theoretical measure of underlying discriminability (*d'*). In truth, we and others prefer ROC analysis for applied purposes because it provides an *atheoretical* measure of empirical discriminability (namely, partial area-under-the-curve, or pAUC). The issue of underlying theoretical discriminability only arises when theoreticians seek to explain why one eyewitness identification procedure yields a higher pAUC than another. Lampinen (2016) also argued that favoring the procedure that yields a higher pAUC can lead to an irrational decision outcome. However, his argument depends on needlessly restricting which points from two ROCs can be compared. As a general rule, the maximum-utility point will fall somewhere on the higher ROC, underscoring the need for ROC analysis. Thus, Lampinen's (2016) arguments against the usefulness of ROC analysis are unfounded.

Keywords: Eyewitness Identification; Signal Detection Theory; ROC Analysis; Discriminability

ROC Analysis in Theory and Practice

In recent work, we and others have used receiver operating characteristic (ROC) analysis to measure how well eyewitnesses can distinguish between innocent and guilty suspects using simultaneous lineups, sequential lineups, or showups (Colloff, Wade, & Strange, 2016; Seale-Carlisle & Mickes, 2016; Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes, Flowe, & Wixted, 2012; Neuschatz, Wetmore, Key, Cash, Gronlund, & Goodsell, 2016). The preferred dependent measure is partial area under the ROC curve (pAUC) because its interpretation is unambiguous and is not dependent on any theory. Except under rare conditions (such as when ROCs from two conditions cross over), the diagnostic procedure that yields the highest pAUC is the one that eyewitnesses can use to most accurately classify innocent suspects and guilty suspects into their true categories. Stated differently, the procedure that yields a higher pAUC can be used to achieve both a higher correct ID rate and a lower false ID rate than a competing procedure. The fact that pAUC provides that kind of objective, theory-free information about the ability of eyewitnesses to discriminate the innocent from the guilty is why we and others judge it to be an improvement over a measure like the diagnosticity ratio (National Research Council, 2014; Rotello, Heit & Dube, 2014).

Contrary to this view, Lampinen (2016) argued that ROC analysis is not the superior methodology it has been made out to be. For example, he considered the possibility that proponents of ROC analysis favor it because they are under the impression that pAUC provides a pure measure of underlying (i.e., theoretical) discriminability.[1] Using a simulation, he presented evidence against that idea by showing that a theoretical measure of underlying discriminability (*d'*) can be held constant across two conditions that differ in pAUC. However, Lampinen is

[1] Specifically, Lampinen (2016) speculated that "One possibility is that ROC analyses provide a pure measure of underlying discriminability and thus, the proponents argue, should be the preferred method of measuring lineup performance" (p. 31).

wrong to suggest that pAUC has been judged to be superior to the diagnosticity ratio because it is thought to provide a pure measure of underlying discriminability. As noted above, pAUC has been judged to be superior to the diagnosticity ratio because it provides an objective measure of empirical discriminability, not because it provides a theoretical measure of underlying discriminability. Still, we consider his argument in some detail because it has been interpreted as providing "…strong additional evidence that ROC analyses on lineups are not measures of discriminability" (Wells, Smalarz, & Smith, 2015, p. 316).

Separately, Lampinen argued that ROC analysis, when used for applied purposes, may not properly take into account the costs of decision-making errors (e.g., the cost of false IDs) and can therefore result in an irrational preference for the procedure that ROC analysis deems to be superior (namely, the one that yields a higher pAUC). However, Lampinen erred in this analysis by proposing that there are restrictions on the points from two ROCs that can be legitimately compared to each other. No such restrictions exist, and once that fact is appreciated, it becomes clear that utility can be maximized by using the procedure that yields the highest pAUC. In the two main sections that follow, we critically evaluate the theoretical and applied arguments that were advanced by Lampinen in his effort to discredit ROC analysis.

Theoretical considerations

Does pAUC provide a pure measure of underlying discriminability? We have never suggested that it does, and there is no way that it could. Underlying discriminability (*d'*) is a theoretical concept that does not exist independent of a specific quantitative model, one that specifies the mathematical shapes of the underlying memory strength distributions as well as the decision rule that eyewitnesses are assumed to use. By contrast, empirical discriminability (pAUC) is measured without reference to any theory. It is simply a measure of the area under the empirically obtained ROC points.

For a given set of ROC data, there is only one estimate of pAUC, and this is the objective measure that policymakers should care about. By contrast, there is no end to the number of different theoretical *d′* estimates that could be obtained from those same ROC data. In fact, the number of different theoretical *d′* estimates that could be obtained from a given set of ROC data is limited only by the number of different signal detection models one is willing to entertain. That being the case, pAUC cannot invariably serve as a pure substitute for *d′*. This is not to suggest that the two measures are likely to be unrelated. Because pAUC and *d′* are both computed from the same set of ROC data, it seems reasonable to suppose that they will often be correlated with each other, and indeed they are (Mickes, Moreland, Clark & Wixted, 2014). However, they are not mathematically identical measures, so it must be the case that they will sometimes disagree. As described next, Lampinen's simulations simply confirmed that uncontroversial fact.

Lampinen specified one particular quantitative signal detection model characterized by the following specific (and, in our view, reasonable) assumptions: (1) underlying memory strength was represented by Gaussian distributions; (2) the decision rule consisted of basing a decision solely on the strength of the most familiar face in the lineup (without regard for the other faces in the lineup); and (3) the decision criteria were assumed to be fixed across participants (with the implicit assumption that any difference in criterion variability that might exist across conditions was sufficiently small that it could be safely ignored). Using that particular model, he generated simulated ROC data from showups and lineups that were equated in terms of underlying *d′* as defined by that model. Even so, the two procedures yielded different simulated empirical ROCs and (therefore) different pAUCs. The fact that the two measures can disagree in this way proves only that the mathematics underlying the computation of pAUC and *d'* are not identical, which must be true given that one measure (pAUC) reflects area under the ROC points and the other measure (underlying *d'*) is derived from a specific signal detection model and would likely differ if another signal detection model were employed. However, Lampinen instead took this result to mean that "pAUC analyses do not provide a valid way of comparing identification procedures in terms of the effect they have on underlying memory discriminability" (p. 26).

Although a simulation study like the one that Lampinen performed does not support that conclusion, it does underscore a potential problem for any theory that has been advanced to explain an observed difference in pAUC across two conditions in terms of a corresponding difference in underlying discriminability. To illustrate the potential problem, we follow Lampinen (see his section on "Underlying Discriminability," p. 24) and consider this issue in relation to our own diagnostic feature detection theory (Wixted & Mickes, 2014). Here is a brief summary of how we have previously interpreted pAUC differences in terms of their applied implications and, separately, in terms of our theory:

1. In 2012, an unexplained empirical phenomenon was observed. Specifically, ROC studies reported that pAUC for simultaneous lineups was greater than pAUC for sequential lineups (Gronlund et al., 2012; Mickes et al., 2012). Other work showed that, in addition, pAUC for simultaneous lineups was greater than pAUC for showups (Gronlund et al., 2012). These pAUC results were taken to mean that simultaneous lineups can be used to achieve both a higher correct ID rate and a lower false ID rate

(i.e., they can be used to achieve higher empirical discriminability) than alternative procedures. Note that the results were *not* interpreted to mean that pAUC provides a pure measure of underlying discriminability.

2. In 2014, the diagnostic feature-detection theory was advanced to explain those empirical pAUC results (Wixted & Mickes, 2014). This theory assumes that simultaneous lineups help witnesses to discount non-diagnostic facial features, and it predicts (1) that simultaneous lineups will yield higher discriminability than both sequential lineups and showups and (2) that fair lineups will yield higher discriminability than unfair lineups. The diagnostic feature-detection theory does claim that the observed pAUC differences across conditions reflect corresponding differences in underlying discriminability. However, it does not also claim that pAUC provides a pure measure of underlying discriminability, which is the claim that Lampinen sought to disprove.

Although Lampinen's simulation study addressed a claim that, to our knowledge, no one has made, it nevertheless does raise the following legitimate concern: Conditions that differ in terms of pAUC (e.g., simultaneous and sequential lineups), which the diagnostic feature-detection theory explains in terms of a corresponding difference in underlying discriminability, may not actually differ in terms of underlying *d′* as measured by a specific signal detection model. After all, Lampinen's simulations show that, in one particular circumstance, at least, a difference in pAUC can arise even when underlying *d′* is equated across conditions. If the same is true of actual data, not just those simulated data, then a model like ours, which relies on a difference in underlying discriminability to explain a difference in pAUC, might be in error.

Note that the diagnostic feature-detection theory is not itself a quantitative model that can be used to measure underlying *d'*. It is instead a conceptual theory about the discounting of shared facial features under certain conditions. Still, the theory does assume that if one specified a viable signal detection model and used it to quantify underlying discriminability across conditions, then the observed differences in pAUC would be reflected in corresponding differences in underlying *d'*.

How can that assumption be tested? There is only one way to do so that we know of, and it is not the approach that Lampinen used. Step 1 is to specify a quantitative signal detection model. As noted above, this is an essential step because the concept of underlying discriminability exists *only in relation to a specific quantitative model*, and it is the proposed quantitative model that makes it possible to measure *d'*. Step 2 is to compute both pAUC and, using that proposed quantitative measurement model, *d'* from the ROC data that a theory like the diagnostic feature-detection theory purports to explain.

Critically, one obtains the model's estimate of underlying *d'* for each of two conditions being compared by fitting the specified signal detection model to the empirical ROC data from each condition using an optimization procedure – not by conducting a simulation. Separately, one also obtains an atheoretical estimate of pAUC from the same two conditions using pROC software (Gronlund, Mickes & Wixted, 2014; Robin et al., 2011). To the extent that pAUC and *d'* agree about which condition is associated with higher discriminability, it increases confidence in the validity of the assumption that the observed pAUC differences reflect corresponding differences in underlying discriminability (as assumed by the diagnostic feature-detection theory). To the extent that they disagree, it reduces confidence in that assumption – although it could instead mean that the specific signal detection model that was used to compute *d′* is wrong.

The model that Lampinen used for his simulations is a reasonable candidate for Step 1. Therefore, to estimate underlying *d′* from conditions that differ in pAUC, we used his signal detection model and fit it to empirical ROC data obtained from several studies that compared (1) showups vs. lineups, (2) fair lineups vs. unfair lineups, (3) simultaneous vs. sequential lineups, and (4) simultaneous lineups vs. 9-member sequential video lineups (i.e., US vs. UK lineups). In each of these studies, a significant pAUC difference between conditions has been reported – a difference that the diagnostic feature-detection theory explains in terms of a difference in underlying discriminability. The key question is whether, using Lampinen's own measurement model, a difference in underlying discriminability (*d′*) is also observed between those conditions. The details of the model-fitting procedure are presented in the appendix. Keep in mind that the issue being addressed here has to do with the validity of a theory that has been proposed to explain pAUC differences between, for example, simultaneous and sequential lineups. It has nothing at all to do with which procedure is, empirically, diagnostically superior to the other. The pAUC measure already provides that information.

Do the two measures (*d′* and pAUC) agree? Indeed they do. As shown in Table 1, for each of the 7 comparisons taken from 5 published studies performed by multiple independent labs, the theory-based estimate of underlying discriminability, *d′*, agrees with the theory-free measure, pAUC. Note that the pAUC results from all of these comparisons are consistent with the predictions of the diagnostic feature-detection theory. Thus, if the signal detection model used by Lampinen in his simulations is assumed to provide valid estimates of *d′*, then the observed differences in pAUC reflect differences in underlying discriminability, as assumed by the diagnostic feature-detection theory.

Although these findings support the interpretation provided by the diagnostic feature-detection theory, they do not deny what Lampinen showed with his simulation, which is that it is possible to identify a circumstance where the two measures disagree. That can certainly happen because pAUC and *d′* are not mathematically identical measures. Lampinen's mistake was to conclude that because the two measures *can* disagree, then any effort to explain a difference in pAUC in terms of underlying discriminability is invalid. The results summarized in Table 1 show that such a conclusion is unwarranted.

Our own conclusion that pAUC and *d′* generally agree is based on extant data, but there is no way to know what future studies will show. Moreover, the agreement between pAUC and *d′* shown in Table 1 might change if a different signal detection model were used to estimate underlying *d′*. Our point is simply that the way to test the assumption that differences in pAUC reflect differences in underlying discriminability is to follow the steps illustrated here, including the indispensable step of actually fitting the specified signal detection measurement model to empirical ROC data instead of performing a simulation showing that a hypothetical circumstance can be found where the two measures disagree. A simulation like that only serves to test the claim that the two measures always agree (i.e., that one measure is a pure substitute for the other), which is a claim that no one has made.

Again, these considerations, while relevant to theory development, have no bearing on the applied question of which eyewitness identification procedure is diagnostically superior in actual practice. Instead, these considerations are of concern to theoreticians only. For applied purposes, pAUC is the only measure that matters because (rare exceptions notwithstanding, such as crossover ROCs) the procedure that yields a higher pAUC can be used to achieve both a higher correct ID rate and a lower false ID rate than a competing procedure.

Applied Considerations

In addition to raising the theoretical issues discussed above, Lampinen also argued against the applied value of ROC analysis. In so doing, he introduced a pivotal idea that we have not previously encountered and that seems entirely incorrect to us. The crux of his argument is that certain considerations constrain the comparison of a point from one ROC to a point on a different ROC. In Lampinen's view, the failure to appreciate those constraints can create the impression that the procedure associated with a higher pAUC is diagnostically superior to a procedure associated with a lower pAUC even when it is actually the other way around. He illustrated this argument using "utility" (i.e., overall value) computed for the hypothetical ROC points shown in his Figure 6. We consider this argument after first describing the putative constraint on which his conclusions depend.

A Constraint on ROC Analysis. Unlike the pAUC measure, which is based on all of the ROC points for a given identification procedure, utility is computed separately for each point on the ROC. Lampinen's argument is that a measure associated with a particular point on the ROC for one procedure (Procedure A) can be legitimately compared only to the corresponding point on the ROC for another procedure (Procedure B). Corresponding points are those that are based on IDs made with the same level of confidence. For example, both the rightmost point on the ROC for Procedure A and the rightmost point on the ROC for Procedure B are computed using all suspect IDs regardless of the level of confidence (these are the overall correct and false ID rates). In Lampinen's view, because they were computed in the same way, a measure associated with those two points (e.g., their utility measures) can be legitimately compared to each other but not to any other ROC point. Similarly, the next point to the left on each ROC is computed using IDs that were made with all but the lowest level of confidence, so they, too, can be legitimately compared to each other but not to any other point (and so on). According to Lampinen, not constraining comparisons across two ROCs in this way "…is not a reasonable or scientifically valid way to compare two conditions" (p. 28). The basis for that claim was not explained, and no source was cited. We believe that no source could be cited because this idea is incorrect, has no precedent, and loses sight of the primary advantage of ROC analysis. Every point on the ROC is achievable (if not, the points could not be plotted in the first place). Therefore, any point on one ROC can be legitimately compared to any point on the other ROC when the goal is to determine the most valued (achievable) outcome. In other words, there is no constraint on which of two ROC points can be legitimately compared to each other. Moreover, as we explain next, Lampinen's claim that a procedure associated with a lower pAUC can have greater utility than a procedure associated with a higher pAUC is dependent on the presumed necessity of this (in truth, inappropriate) constraint.

Maximizing Utility. Consider the hypothetical data from the two eyewitness identification procedures (Procedure A and Procedure B) shown in Figure 1, which were reproduced from Lampinen's Figure 6. Obviously, Procedure A yields a higher pAUC than Procedure B, but Lampinen made the case that once utility is computed, Procedure B might be diagnostically superior nonetheless.

A utility analysis involves taking into consideration subjective values, so we will illustrate his argument as well as our rebuttal by assigning specific subjective values to the four possible decision outcomes.[2] The four possible outcomes are as follows: a correct suspect ID,

[2] In this example, we focus solely on whether or not a suspect ID was made from a target-present or target-absent lineup because suspect IDs are clearly consequential, whereas filler IDs are much less consequential (i.e., their

also known as a hit (the guilty suspect was correctly identified from a target-present lineup), a miss (the suspect was not identified from a target-present lineup because the lineup was rejected or a filler was incorrectly identified), a correct rejection (the innocent suspect was not identified from a target-absent lineup because the lineup was rejected or a filler was incorrectly identified), and a false ID, also known as a false alarm (the suspect was incorrectly identified from a target-absent lineup). Using arbitrary units, shows hypothetical subjective values that might be assigned to each decision outcome. In this example, the cost of a false alarm (-10) is ten times the cost of a miss (-1). The benefit of a hit (+5) is five times the benefit of a correct rejection (+1). Note that this is just an example. The point we make here would apply even if different non-zero values were used for the positive and negative outcomes so long as correct decisions (hits and correct rejections) are assigned higher value than incorrect decisions (misses and false alarms).

Assuming equal base rates of target-present and target-absent lineups for the sake of simplicity, the overall utility of each ROC point can be calculated once the subjective values associated with the different decision outcomes are specified. The equation for computing utility is as follows (see Equation 1.14 in Green & Swets, 1966, p. 22, for the full equation that allows for unequal base rates):

$$\text{Utility} = HR{*}V_{HR} + M{*}V_{M} + CR{*}V_{CR} + FA{*}V_{FA}$$

where HR is the hit rate, M is the miss rate (equal to $1 - HR$), CR is the correct rejection rate, FA is the false alarm rate (equal to $1 - CR$), and, for this concrete example, $V_{HR} = +5$, $V_M = -1$, $V_{CR} = +1$, and $V_{FA} = -10$.

The utility values shown next to each ROC point in Figure 2 were computed using this equation. Thus, for example, the correct and false ID rates for the rightmost ROC point for

utility is assumed to be negligible by comparison). However, as noted by Mickes et al. (2012), to the extent that one attaches utility to filler IDs, those considerations could easily be factored in as well (but they would not likely affect any conclusions for fair lineups).

Procedure A are .95 and .35, respectively. For that point, *HR* = .95, *M* = 1 - .95 = .05, *CR* = 1 - .35 = .65, and *FA* = .35. Using the decision-outcome values shown in Table 1, the utility of that ROC point is equal to .95 * (+5) + .05 * (-1) + .65 * (+1) + .35 * (-10) = 1.81. The same approach was followed to compute the utility values for each ROC point in Figure 2.

Note that the utility of the rightmost ROC point in Figure 2 is higher for Procedure B (2.96) than it is for Procedure A (1.81), just as in Lampinen's example in which he argued that, under conditions like these, it might be a mistake to prefer Procedure A (because of its higher ROC) to Procedure B. In fact, he went so far as to imply that it would be irrational to prefer Procedure A under these circumstances by pointing out that: "A basic tenant [sic] of rational choice theory, is that rational choices are associated with selecting the alternative with greatest utility (von Neumann & Morgenstern, 1944). But as this example shows, ROC analyses will sometimes lead to selecting the alternative with the lowest expected utility" (p. 29).

When will ROC analysis lead to selecting the alternative with the lowest expected utility? It will do so when an analytical constraint is unnecessarily imposed according to which a decision about which procedure is diagnostically superior is limited to a comparison between a single point on one ROC (such as the rightmost point) and the corresponding point on the other ROC, as if the other points do not exist or are otherwise unachievable. Imposing that constraint is the only way to make an argument that the procedure associated with the lower ROC should be preferred to the procedure associated with the higher ROC. However, imposing that constraint is itself inconsistent with the rational goal of maximizing utility because doing so results in selecting an outcome with lower utility than can otherwise be achieved. We agree with Lampinen that it can be considered irrational to choose the alternative with lower expected utility. However, look again at the utility values shown next to each ROC point in Figure 1. The highest utility is found on the higher ROC (that point is circled), as will almost always be true. The fact that the maximum-utility point will generally be found on the higher ROC is precisely why ROC analysis is essential for determining the diagnostically superior procedure.

The key consideration is that each point on both ROCs represents an achievable outcome, and each point on both ROCs is associated with a specific utility. It therefore follows that the optimal decision strategy is to choose the achievable point that maximizes utility, without worrying about any other point on either ROC. Because the maximum-utility point will be found on the higher ROC (except in unusual circumstances, such as when ROCs cross over), maximizing utility requires first identifying and then making use of the procedure that yields a higher ROC (i.e., the one with the higher pAUC).

Conclusion

Lampinen erred with regard to his main point about the relationship between theory and ROC analysis. His mistake was to assume that proponents of ROC analysis favor that approach because they believe that pAUC provides a pure measure of underlying discriminability. Instead, proponents of ROC analysis favor that approach because pAUC (a measure computed from empirical ROC data) indicates which procedure can achieve both a higher correct ID rate and a lower false ID rate than a competing procedure.

In an effort to explain certain findings based on the pAUC measure (e.g., simultaneous lineups yield a higher pAUC than sequential lineups), Wixted and Mickes (2014) proposed the diagnostic feature-detection theory. This theory is based on several assumptions, one of which is that differences in pAUC across conditions reflect differences in underlying discriminability across conditions. Lampinen showed that underlying d' can remain constant across two conditions that nevertheless yield different pAUCs, raising the possibility that the assumed correspondence between pAUC and underlying discriminability is incorrect. However, if Lampinen had fit the same signal detection model that he used for his simulations to the relevant ROC data, he would have discovered what we report here in Table 1: pAUC and *d′* agree with respect to the ROC data that the diagnostic feature-detection theory purports to explain. It did not have to turn out that way, and future results may tell a different story, but the available evidence supports the assumption that differences in pAUC reflect differences in underlying discriminability (which is not to say that pAUC provides a pure measure of discriminability in every conceivable circumstance).

Lampinen also erred in his diagnostic utility analysis of ROC data comparing different lineup procedures. His key mistake was to suggest that it makes sense to compare ROC points only from two different procedures that are equated in terms of expressed confidence. In truth, when used for applied purposes, the goal is to find the ROC point that maximizes utility without worrying about the other ROC points. Rare exceptions notwithstanding, the ROC point that maximizes utility will be found on the higher ROC. That is precisely why ROC analysis should be performed when comparing eyewitness identification procedures (i.e., to find out which procedure yields the higher ROC).

References

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

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups don't just make witnesses more willing to choose the suspect, they also make them more likely to confuse innocent and guilty suspects. *Psychological Science, 27,* 1227-1239.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (A reprint, with corrections of the original 1966 ed.). Huntington, NY: Robert E. Krieger Publishing Co.

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

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.

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

Gronlund, S. D., Wixted, J. T. & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science, 23*, 3-10.

Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory and Cognition, 5,* 21-33.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver Operating Characteristic Analysis of Eyewitness Memory: Comparing the Diagnostic Accuracy of Simultaneous vs. Sequential Lineups*. Journal of Experimental Psychology: Applied, 18,* 361-376.

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

National Research Council (2014). *Identifying the Culprit: Assessing Eyewitness Identification*. Washington, DC: The National Academies Press.

Neuschatz, J. S., Wetmore, S. A., Key, K. N., Cash, D. K., Gronlund, S. D., & Goodsell, C. A. (2016). A comprehensive evaluation of showups. In B. Bornstein, & M. K. Miller (Eds.), *Advances in psychology and law* (pp. 43-69). Switzerland: Springer International Publishing.

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.

Seale-Carlisle, T. M. & Mickes, L. (2016). US lineups outperform UK lineups. *Royal Society Open Science*. DOI: 10.1098/rsos.160300

Wells, G. L., Smalarz, L., Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition.*

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A. & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition, 4,* 8-14.

Wixted, J. T. & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121,* 262-276.

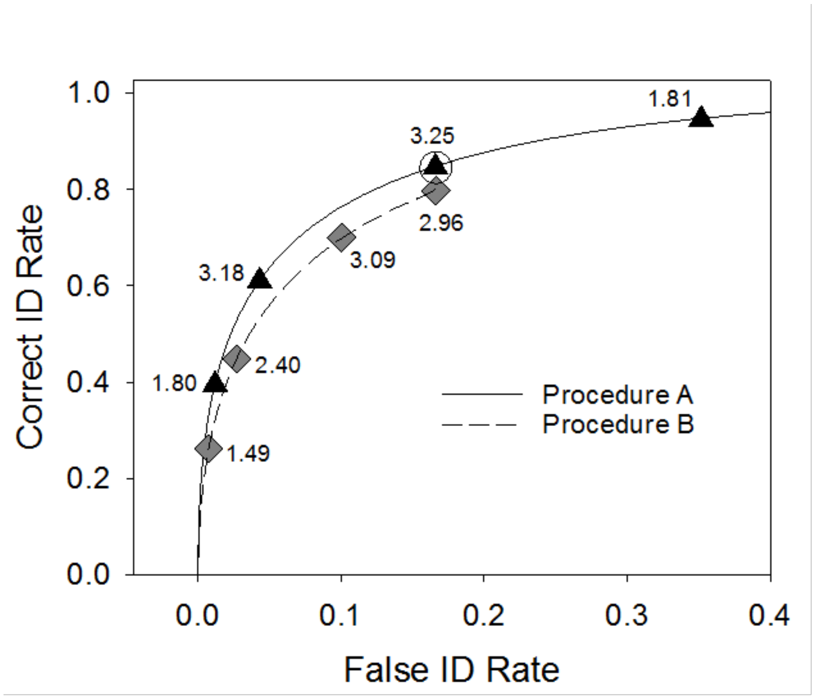
Table 1. *A comparison of pAUC and d' estimated from the same ROC data from different conditions in five recent studies.*

Study	Procedure	pAUC	*d'*
Wetmore et al. (2015)	Lineup	0.09	1.33
	Showup	0.04	0.72
	p-value	<.001	<.001
Wetmore et al. (2015)	Fair Lineup	0.06	1.68
	Biased Lineup	0.03	1.19
	p-value	<.01	<.001
Colloff et al. (2016)	Fair Lineup	0.016	0.83
	Biased Lineup	0.008	0.54
	p-value	<.001	<.001
Mickes et al. (2012)	Simultaneous	0.13	1.85
	Sequential	0.09	1.32
	p-value	<.05	<.001
Dobolyi & Dodson (2013) [2 repetitions]	Simultaneous	0.034	2.05
	Sequential	0.028	1.81
	p-value	<.001	0.032
Dobolyi & Dodson (2013) [4 repetitions]	Simultaneous	0.044	2.52
	Sequential	0.037	2.18
	p-value	<.001	0.007
Seale-Carlisle & Mickes (2016)	US	0.17	1.07
	UK	0.10	0.41
	p-value	0.006	<.001

Table 2. Example of possible subjective decision outcome values assigned to the 4 decision outcomes with respect to suspects in lineups or showups.

		Decision Outcome	
		Suspect ID	No ID or Filler ID
Lineup Status	Target Present	+5	-1
	Target Absent	-10	+1

Figure 1. Hypothetical ROC data from two conditions (Procedure A and Procedure B). The data were estimated from Lampinen's Figure 6. The values adjacent to each ROC points are its utility computed using the subjective decision outcome values in Table 2. The encircled point is the point with the highest utility.



Appendix

The diagnostic feature detection theory explains empirical trends in pAUC data by assuming that underlying discriminability differs in the same way as pAUC does across conditions. There is only one way to find out if that assumption is valid, and that is to compute both pAUC and *d′* from the relevant data to see how often they actually do lead to different conclusions. Any conclusion about whether or not they do agree depends on the validity of the model that is used to estimate *d′*. We estimate underlying *d′* here by fitting the signal-detection model that Lampinen used to perform his simulations to actual empirical data from studies comparing (1) showups vs. lineups, (2) fair lineups vs. unfair lineups, (3) simultaneous vs. sequential lineups, and (4) simultaneous lineups vs. 9-member PROMAT sequential video lineups (US vs. UK, respectively). The results of the analyses described below are presented in Table 1 of the main article.

The showup vs. lineup data come from Wetmore et al. (2015), as do the data for one of the fair vs. biased lineup comparisons. In both cases, we analyzed the data collapsed across their delay manipulation (immediate vs. delay) and used the innocentweak target-absent data. The innocentstrong data were not used for either comparison because, even in the "fair" innocentstrong lineup condition, the innocent suspect was misidentified from a target-absent lineup much more often than a filler was (i.e., that condition was not actually fair in that the innocent suspect stood out from the fillers).

The other fair vs. biased lineup results shown in Table 1 of the main article came from Colloff et al. (2016). Because they reported both pAUC and *d′* values, we simply reproduced in our Table 1 the values they reported in their article and supplemental material (i.e., we did not perform any fits to their data). In reporting the results from their fair lineup data, we collapsed the relevant measures (pAUC and *d′*) across their 3 (very similar) fair conditions.

The simultaneous vs. sequential data come from Experiment 1a of Mickes et al. (2012, see their Figure 6A) and from Dobolyi and Dodson (2013, see their Figure 1). Dobolyi and Dodson (2013) had two learning conditions (2 repetitions and 4 repetitions of the target stimuli during encoding), so their data provide two opportunities to compare pAUC and *d′*. Finally, the US vs. UK data come from Seale-Carlisle and Mickes (2016, see their Figure 1).

As noted above, the *d′* values were estimated by fitting the signal-detection process model that Lampinen used for his simulations to the ROC data from these studies. To estimate *d′*, an equal-variance version of the model was fit to the data by minimizing the chi-square goodness-of-fit statistic. Separately, pAUC values were computed using the statistical package pROC (Robin et al., 2011). This program estimates pAUC over a partial range without resorting to any theoretical model. Thus, for each condition, we computed both pAUC and *d′* from the ROC data.

According to the simple signal-detection model that Lampinen used, memory strength values for fillers, innocent suspects and guilty suspects are distributed according to Gaussian distributions with means of μ_{Filler}, $\mu_{Innocent}$, and μ_{Guilty}, respectively. A 6-member target-present lineup is conceptualized as 5 random draws from the Filler distribution and 1 random draw from the Guilty distribution; a 6-member target-absent lineup is conceptualized as 5 random draws from the Filler distribution and 1 random draw from the Innocent distribution. If a fair target-absent lineup is used, as is true of most of our fits, then $\mu_{Filler} = \mu_{Innocent}$, in which case the model reduces to a 2-distribution model (and a target-absent lineup is conceptualized as 6 random draws from the Filler distribution).

Of primary interest is the ability of eyewitnesses to collectively discriminate between innocent and guilty suspects, and that ability is represented by the distance between the means of the $\mu_{Innocent}$ and μ_{Guilty} distributions. That distance is the d' measure of interest, and it is the value reported in Table 1. To estimate d', we fit this model to the following data sets:

1. Showup vs. lineup: Wetmore et al. (2015, Perpetrator and Innocentweak conditions in their Tables 1 and 2, collapsed across delay)
2. Fair vs. biased lineups: Wetmore et al. (2015, again, Perpetrator and Innocentweak conditions in their Tables 1 and 2, collapsed across delay)
3. Simultaneous vs. Sequential: Experiment 1a of Mickes et al. (2012, Figure 6A)
4. Simultaneous vs. Sequential: Dobolyi and Dodson (2013, separately for their 2-repetition and 4-repetition conditions)
5. US vs. UK: Seale-Carlisle & Mickes (2016)

To fit the model, we first collapsed some data sets to a 3-point scale. For the Wetmore et al. (2015) data, this was accomplished by combining confidence ratings of 6 and 7 (high confidence), 3, 4 and 5 (medium confidence), and 1 and 2 (low confidence). For the Mickes et al. (2012) data, this was accomplished by combining confidence ratings of 90 to 100 (high confidence), 70-80 (medium confidence), and 0-60 (low confidence). We collapsed the data in this manner to keep the number of parameters to be estimated reasonably low.

With $\mu_{Innocent}$ always set to 0 as a reference point and the standard deviations for all three distributions always set to 1 for the sake of simplicity, the basic model has $n_c + 1$ parameters, where n_c is the number of confidence levels. Thus, for example, if the confidence data are collapsed to a 3-point scale (low, medium and high), there would be 4 parameters to estimate from the ROC data of a given condition (μ_{Guilty}, c_{Low}, c_{Medium}, and c_{High}), where μ_{Guilty}, is the mean of the guilty suspect distribution, and c_{Low}, c_{Medium}, and c_{High} are the decision criteria. If the lineup is fair, then $\mu_{Filler} = \mu_{Innocent} = 0$, so no additional parameter is needed to estimate those means. Note that for the equal-variance model we used, μ_{Guilty}, is the standardized difference between μ_{Guilty} and $\mu_{Innocent}$ (i.e., $\mu_{Guilty} = d'$). If the lineup is unfair, then $\mu_{Filler} \neq \mu_{Innocent}$, and another parameter (μ_{Filler}) has to be estimated. Because $\mu_{Innocent} = 0$, for an unfair lineup, the estimated value of μ_{Filler} will be negative (i.e., the fillers will seem less like the guilty perpetrator than the innocent suspect does).

For a given fit, the degrees of freedom equal the degrees of freedom in the data minus the number of parameters estimated from the data. As an example, consider the fair vs. biased fit of the Wetmore et al. (2015) data. For the target-present lineup data, there are 6 degrees of freedom: 3 levels of confidence for suspect IDs plus 3 levels of confidence for filler IDs (the observed number of no IDs is fully constrained once those other 6 values are known and so does not add another degree of freedom). Similarly, for target-absent lineup data involving an innocent suspect, there are 6 degrees of freedom: 3 levels of confidence for suspect IDs plus 3 levels of confidence for filler IDs. Thus, there are 12 degrees of freedom in the biased data. In a fair target-absent lineup, the innocent suspect can be treated as another filler (which is the approached we used), so the target-absent lineup provides only 3 degrees of freedom (namely, the observed number of filler IDs, counting IDs to the innocent suspect, across 3 levels of confidence). The total degrees of freedom in the biased and fair data added together comes to 12 + 9 = 21. A total of 9 parameters are estimated when the model is fit to the ROC data from these two conditions (μ_{Guilty}, c_{Low}, c_{Medium}, and c_{High} for both conditions, yielding 8 estimated parameters in all, plus an estimate of μ_{Filler} for the biased condition). Therefore, when the full model is simultaneously fit to the fair and biased data, the degrees of freedom associated with that fit is equal to 21 degrees of freedom in the data – 9 estimated parameters = 12 remaining degrees of freedom.

The parameters were estimated by adjusting them until the chi-square goodness-of-fit statistic comparing observed and predicted observations was minimized. For each of the *d′* comparisons, the model was fit twice: once allowing *d′* to differ across conditions and once again constraining *d′* to be equal across conditions (i.e., full vs. reduced models, respectively). The other parameters were allowed to differ across conditions. The question of interest was whether the chi-square goodness-of-fit statistic would be significantly worse for the constrained fits, in which case one would conclude that *d′* differed across conditions.

Tables A1 through A6 show the fits of the full model (allowing *d′* to differ across conditions) and constrained model (requiring that *d′* be equal across conditions) for the showup vs. lineup, fair vs. biased lineups, simultaneous vs. sequential lineups, and US vs. UK lineups. As shown in Table 1 of the main article, in each case, the pAUC difference across conditions is significant. Is the same true of *d′*? The answer is given by the difference in the chi-square goodness of fit statistics between the two fits, which is itself a chi-square with 1 degree of freedom. In all cases, the chi-square was significant:

1. Showup vs. lineup (Wetmore et al., 2015): $\chi^2_{\text{Reduced}} - \chi^2_{\text{Full}} = 61.9$, $p < .001$.

2. Fair vs. unfair lineups (Wetmore et al., 2015): $\chi^2_{\text{Reduced}} - \chi^2_{\text{Full}} = 10.9$, $p < .001$.

3. Simultaneous vs. Sequential (Mickes et al., 2012): $\chi^2_{\text{Reduced}} - \chi^2_{\text{Full}} = 12.6$, $p < .001$.

4. Simultaneous vs. Sequential (Dobolyi & Dodson, 2013, 2-rep): $\chi^2_{\text{Reduced}} - \chi^2_{\text{Full}} = 4.6$, $p = .032$.

5. Simultaneous vs. Sequential (Dobolyi & Dodson, 2013, 4-rep): $\chi^2_{\text{Reduced}} - \chi^2_{\text{Full}} = 7.2$, $p = .007$.

6. US vs. UK (Seale-Carlisle & Mickes, 2016): $\chi^2_{\text{Reduced}} - \chi^2_{\text{Full}} = 62.7$, $p < .001$.

These findings indicate that the results accord with conclusions based on pAUC and that they also accord with the predictions about underlying *d′* made by the diagnostic feature-detection theory. Still, as noted earlier, any conclusion about whether or not pAUC and *d′* agree depends on the validity of the model that is used to estimate *d′*. It would be easy to specify a different signal detection model that holds *d′* constant across all of the conditions compared in Table 1 despite the pAUC difference that is observed in each case. One could, for example, hypothesize that criterion variability (which reduces pAUC even without changing underlying *d′*) is greater in each of the conditions in which pAUC is reduced compared to the other condition. We know of no principled reason for proposing such a model, but the point is that it could explain the pAUC results without assuming a *d′* difference across conditions. No simulation is needed to prove it. Nevertheless, using Lampinen's (reasonable) model, *d′* and pAUC agree.

Table A1. *Showup vs. lineup fits to data (collapsed across retention interval and excluding Innocentstrong) from Wetmore et al. (2015).*

Full Model

estimate	Showup	Lineup
μ_{Guilty} *(d')*	0.47	1.69
c_{Low}	0.23	1.10
c_{Medium}	0.47	1.51
c_{High}	1.30	2.45
χ^2	20.9	
df	7	
p	0.004	

Reduced Model

estimate	Showup	Lineup
μ_{Guilty} *(d')*	1.19	1.19
c_{Low}	0.51	1.01
c_{Medium}	0.74	1.40
c_{High}	1.58	2.26
χ^2	82.8	
df	8	
p	<.001	

Table A2. *Fair vs. Biased lineup to data (collapsed across retention interval and excluding Innocentstrong) from Wetmore et al. (2015).*

Full Model

estimate	Fair	Biased
μ_{Guilty} (d')	1.69	1.15
$\mu_{Innocent}$	0.00	0.00
$\mu_{Filler\text{-}TA}$	0.00	-1.01
c_{Low}	1.10	1.39
c_{Medium}	1.51	1.70
c_{High}	2.45	2.46
χ^2	33.7	
df	12	
p	0.001	

Reduced Model

estimate	Fair	Biased
μ_{Guilty} (d')	1.48	1.48
$\mu_{Innocent}$	0.00	0.00
$\mu_{Filler\text{-}TA}$	0.00	-0.89
c_{Low}	1.47	1.73
c_{Medium}	2.38	2.50
c_{High}	1.89	1.76
χ^2	44.6	
df	13	
p	<.001	

Table A3. *Simultaneous vs. Sequential lineup fits for data from Experiment 1a of Mickes et al. (2012).*

Full Model

estimate	SIM	SEQ
μ_{Guilty} *(d')*	1.87	1.33
c_{Low}	1.49	1.59
c_{Medium}	2.00	2.09
c_{High}	2.78	2.64
χ^2	18.7	
df	10	
p	0.044	

Reduced Model

estimate	SIM	SEQ
μ_{Guilty} *(d')*	1.60	1.60
c_{Low}	1.42	1.65
c_{Medium}	1.92	2.17
c_{High}	2.69	2.73
χ^2	31.3	
df	11	
p	<.001	

Table A4. *Simultaneous vs. Sequential lineup fits for data from Dobolyi and Dodson (2013, 2-repetition condition).*

Full Model

estimate	SIM	SEQ
μ_{Guilty} (d')	2.05	1.81
c_1	1.46	1.52
c_2	1.49	1.54
$c3$	1.68	1.63
$c4$	1.98	1.86
$c5$	2.29	2.20
$c6$	2.83	2.73
χ^2	33.1	
df	22	
p	0.061	

Reduced Model

estimate	SIM	SEQ
μ_{Guilty} (d')	1.92	1.92
c_1	1.45	1.52
c_2	1.49	1.54
$c3$	1.66	1.63
$c4$	1.95	1.87
$c5$	2.25	2.23
$c6$	2.76	2.79
χ^2	37.7	
df	23	
p	0.028	

Table A5. *Simultaneous vs. Sequential lineup fits for data from Dobolyi and Dodson (2013, 4-repetition condition).*

Full Model

estimate	SIM	SEQ
μ_{Guilty} *(d')*	2.52	2.18
c_1	1.54	1.59
c_2	1.56	1.61
c3	1.72	1.67
c4	2.01	1.87
c5	2.37	2.25
c6	2.80	2.81
χ^2	35.3	
df	22	
p	0.036	

Reduced Model

estimate	SIM	SEQ
μ_{Guilty} *(d')*	2.34	2.34
c_1	1.53	1.62
c_2	1.55	1.65
c3	1.69	1.71
c4	1.96	1.92
c5	2.29	2.32
c6	2.69	2.91
χ^2	42.5	
df	23	
p	0.008	

Table A6. *US vs.UK lineup fits for data from Seale-Carlisle & Mickes (2016).*

Full Model

estimate	US	UK
μ_{Guilty} *(d')*	1.07	0.41
c_{Low}	1.19	0.95
c_{Medium}	1.85	1.53
c_{High}	2.53	2.16
χ^2	21.3	
df	10	
p	0.019	

Reduced Model

estimate	US	UK
μ_{Guilty} *(d')*	0.78	0.78
c_{Low}	1.13	1.00
c_{Medium}	1.78	1.59
c_{High}	2.45	2.21
χ^2	83.9	
df	11	
p	<.001	