Journal of Applied Research in Memory and Cognition xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

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

Short Essay

Evaluating eyewitness identification procedures: ROC analysis and its misconceptions☆

John T. Wixted[a,*], Laura Mickes[b]

[a] Department of Psychology, University of California, San Diego, United States
[b] Department of Psychology, Royal Holloway, University of London, United Kingdom

ARTICLE INFO

Article history:
Received 1 July 2015
Accepted 30 August 2015
Available online xxx

Keywords:
Eyewitness identification
ROC analysis
Discriminability
Bayesian analysis
Filler IDs

ABSTRACT

ROC analysis is a straightforward but non-intuitive way to determine which of two identification procedures better enables a population of eyewitnesses to correctly sort innocent and guilty suspects into their respective categories. This longstanding analytical method, which is superior to using the diagnosticity ratio for identifying the better procedure, is not in any way compromised by the presence of fillers in lineups and is not tied to any particular theory of memory or discrimination (i.e., it is a theory-free methodology). ROC analysis is widely used in other applied fields, such as diagnostic medicine, and this is true even when the medical procedure in question is exactly analogous to a lineup (e.g., a detection-plus-quadrant-localization task in radiology). Bayesian measures offer no replacement for ROC analysis because they pertain to the information value of a particular diagnostic decision, not to the general diagnostic accuracy of an eyewitness identification procedure.

In a mock crime study, the relative diagnostic accuracy of competing eyewitness identification procedures is usually based on an analysis of correct and false identification (ID) rates. The correct ID rate is the proportion of target-present lineups from which the guilty suspect was correctly identified, and the false ID rate is the proportion of target-absent lineups from which the innocent suspect was incorrectly identified. Traditionally, only one correct and false ID rate pair has been computed for each procedure. For example, Table 1 reproduces data from the seminal paper by Lindsay and Wells (1985) comparing simultaneous and sequential lineups. The sequential procedure resulted in a small reduction in the correct ID rate (.58 for simultaneous, .50 for sequential) but resulted in a large reduction in the false ID rate (.43 for simultaneous, .17 for sequential).

Correct and false ID rates like these are typically used to compute a *diagnosticity ratio* (correct ID rate/false ID rate), which, in the original Lindsay and Wells (1985) study, was higher for sequential lineups than for simultaneous lineups. Whether or not subsequent research supports that finding has been a matter of sharp disagreement in the literature (Clark, 2012; Gronlund et al., 2009; McQuiston-Surrett, Malpass, & Tredoux, 2006 or Malpass, Tredoux, & McQuiston-Surrett, 2009), but there is no doubt that the sequential procedure has been thought to be superior to the simultaneous procedure to the extent that it has been thought to achieve a higher diagnosticity ratio than the simultaneous procedure (see, for example, the section entitled "Defining Superiority" in Steblay, Dysart, & Wells, 2011). Because the diagnosticity ratio is based entirely on correct and false suspect ID rates, the putative "sequential superiority effect" has only to do with suspect IDs (line 1 of Table 1) and nothing at all to do with filler IDs (line 2 of Table 1).

The diagnosticity ratio is directly related to the likelihood that an identified suspect is guilty. In fact, when multiplied by the prior odds of guilt, it yields the Bayesian posterior odds of guilt. Thus, the higher the diagnosticity ratio, the more trustworthy a suspect identification is. Intuitively, it seems obvious that a lineup procedure that yields a more trustworthy ID (a higher diagnosticity ratio) is superior to a lineup procedure that yields a less trustworthy ID (a lower diagnosticity ratio). Although the problem with this line of reasoning has been understood for decades in other fields, such as diagnostic medicine, its intuitive appeal is undeniably strong and likely explains why 30% of law enforcement agencies in the U.S. that use photo lineups have now adopted the sequential procedure (Police Executive Research Forum, 2013).

☆ 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.

* Corresponding author. Tel.: +1 858 534 3956.
E-mail address: jwixted@ucsd.edu (J.T. Wixted).

http://dx.doi.org/10.1016/j.jarmac.2015.08.009
2211-3681/

Table 1
Data from Lindsay and Wells (1985).

Response outcome	Simultaneous		Sequential	
	Target-present	Target-absent	Target-present	Target-absent
Suspect ID rate	0.58	0.43	0.50	0.17
Filler ID rate	0.12	0.15	0.02	0.18
No-ID rate	0.30	0.42	0.48	0.65

As recently pointed out in a National Academy of Sciences report on eyewitness identification research (National Research Council, 2014), the superior lineup procedure cannot be determined by measuring the diagnosticity ratio and is instead more accurately assessed using receiver operating characteristic (ROC) analysis. The essential problem with trying to use the diagnosticity ratio is that a lineup procedure cannot be adequately characterized by a single diagnosticity ratio any more than a basketball team can be adequately characterized by the performance of a single player. In other words, there is more than one diagnosticity ratio per eyewitness identification procedure, and they all have to be taken into consideration. That is essentially what ROC analysis does.

Anyone who has ever computed a correct ID rate and a false ID rate from a lineup procedure has already computed the first point on the ROC, which is simply a plot of the correct ID rate vs. the false ID rate. ROC analysis consists of nothing more than computing additional correct and false ID rate pairs beyond the one that is typically computed – often without collecting any additional data. In the original Lindsay and Wells (1985) study, for example, they computed only one pair of correct and false ID rates per lineup procedure (Table 1), but they also collected confidence ratings for suspect IDs using a 7-point scale (1 = low confidence to 7 = high confidence). Nothing more than that is needed to plot an ROC curve (Gronlund, Wixted, & Mickes, 2014).

The easiest way to understand ROC analysis is to begin with the idea that it gives you permission to disregard suspect IDs that are acknowledged by the eyewitness to be untrustworthy (namely, IDs that are made with low confidence). If you disregard low-confidence suspect IDs by treating them as effective non-IDs, then (1) you have adopted a more conservative standard for counting suspect IDs, and (2) you will have fewer correct and false IDs than you did before, so the correct and false ID rates will now both be lower.

Which correct and false ID rate pair should you report? The one that counts all IDs no matter how untrustworthy they are acknowledged by the eyewitness to be, or the one that sets a somewhat higher standard by only counting suspect IDs that were made with more than the lowest level of confidence? The fact that this question can be asked shows that you have a choice, and the answer depends on whether you want your results to generalize to jurisdictions that completely ignore confidence (and therefore treat all suspect IDs as equally trustworthy) or to jurisdictions that discount eyewitness IDs made with extremely low confidence. By reporting both pairs of correct and false ID rates, your findings would generalize to a broader range of real-world jurisdictions. Moreover, reporting both would be reporting a 2-point ROC.

Once you realize that you are not obligated to count IDs made with a confidence rating of 1, it immediately follows that you are also not obligated to count IDs made with a confidence rating of 2. Excluding IDs made with a confidence rating of either 1 or 2 from consideration by treating them as effective non-IDs yields yet another pair of correct and false ID rates (i.e., another ROC point). One can obviously proceed in this manner all the way up the confidence scale. Critically, the diagnosticity ratio increases monotonically as an ever higher confidence standard is applied (see Fig. 1 for an illustration). Although it is easy to imagine that the diagnosticity ratio might not increase as responding becomes more conservative, it invariably does, and this effect is naturally predicted by the classic model of recognition memory, namely, signal-detection theory (Egan, 1958; Wixted & Mickes, 2014).

Which ROC point is the best? In truth, no one point on the ROC is inherently superior to any other without factoring in subjective value judgments. Moreover, one cannot possibly know which individual correct and false ID rate (and, therefore, which diagnosticity ratio) best applies to the real world because they all do. The rightmost ROC point – the one that counts all IDs regardless of confidence – might be the one that is the most relevant early in a police investigation (when even a tentative ID of a suspect might be worth considering) or to police jurisdictions where eyewitness confidence is not assessed at all. The leftmost ROC point – the one that counts only IDs made with the highest level of confidence – might be the one that is the most relevant to cases that make it to a later stage of the investigative process (e.g., to cases that are selected for prosecution) or to police jurisdictions where eyewitness confidence is taken into consideration (e.g., when prioritizing cases for further investigation). A single correct and false ID rate (and its corresponding diagnosticity ratio) cannot compete with the family of correct and false ID rates (and their corresponding diagnosticity ratios) when

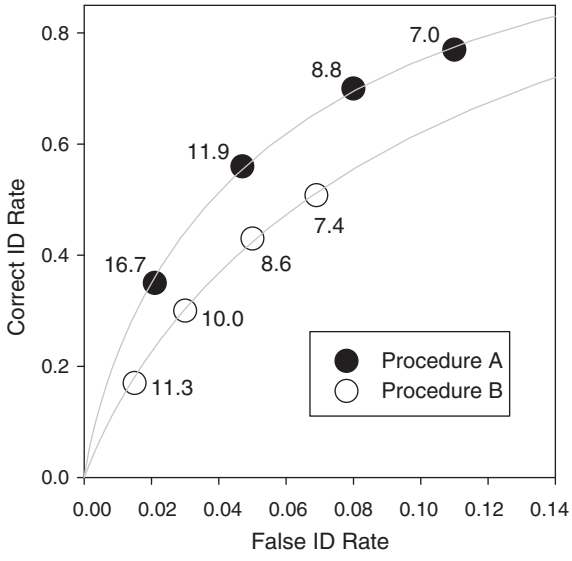

Fig. 1. Illustration of receiver operating characteristic plots for two hypothetical lineup procedures. Each lineup procedure is constrained to yield correct and false ID rates that fall on a curve as responding changes from being very conservative (lower leftmost point of each procedure) to being very liberal (upper rightmost point for each procedure). Values shown next to each data point indicate the diagnosticity ratio (correct ID rate/false ID rate) for that point. In this example, Procedure A is diagnostically superior to Procedure B because for any given false ID rate, Procedure A can achieve a higher correct ID rate. If only a single ROC point is computed for each procedure and are then compared using the diagnosticity ratio (as was done in the vast majority of mock-crime lab studies comparing simultaneous and sequential lineups), the diagnostically inferior lineup procedure could be misconstrued as being the superior procedure (e.g., imagine computing only the rightmost ROC point for each procedure and comparing them using the diagnosticity ratio).

Please cite this article in press as: Wixted, J. T., & Mickes, L. Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition* (2015), http://dx.doi.org/10.1016/j.jarmac.2015.08.009

it comes to generalizing the results of an experiment to the real world.

Because any lineup procedure can achieve a wide range of diagnosticity ratios, it is a mistake to assume that the diagnostically superior procedure is the one that yields the highest diagnosticity ratio based on the singular pair of correct and false ID rates that a researcher decided to focus on. The superior lineup procedure is instead the one that yields the higher ROC (i.e., higher discriminability) because that procedure can be used to achieve a higher correct ID rate and a lower false ID rate than the procedure that yields a lower ROC. The use of the diagnosticity ratio confounds response bias and discriminability (Fig. 1), and this confound is as problematic as other confounds that have plagued the field's search for the most accurate eyewitness identification procedure (e.g., Schacter et al., 2008). Despite its merits (and its computational simplicity), ROC analysis is not very intuitive. Its counterintuitive nature may underlie common criticisms of the procedure that appear to us to be based on misconceptions. We address some of those criticisms next.

1. Criticism #1: ROC analysis ignores filler IDs

Recently, Wells, Yang, and Smalarz (2015) echoed a point we have encountered quite often over the last few years: "The problem is that the ROC approach treats all filler identifications as if they were rejections" (p. 118). However, the question of whether or not it makes sense to count filler IDs is a distraction from the current debate because it is independent of the question of whether or not the diagnostic accuracy of competing lineup formats should be evaluated using the diagnosticity ratio or ROC analysis. As noted earlier, ROC analysis ignores filler IDs to the same extent that a conventional analysis based on the diagnosticity ratio does. Both approaches have been based on correct and false ID rates computed from suspect IDs, and both have ignored filler IDs to the same extent (i.e., necessarily so because the diagnosticity ratio is computed from one ROC point). Researchers are, of course, free to propose some new measure of diagnostic accuracy that counts filler IDs (or to propose a separate analysis that focuses selectively on filler IDs), but the debate we are having now is about the diagnosticity ratio – the measure that is responsible for the substantial real-world impact that eyewitness ID research has had – vs. ROC analysis.

When computing correct and false ID rates, a strong argument can be made that the main focus *should* be placed on consequential suspect IDs, not on comparatively inconsequential filler IDs. Similarly, it has been argued that filler IDs (also known as foil IDs) should be excluded from confidence-accuracy calculations. As pointed out by Penrod and Cutler (1995), "...many erroneous identifications do not result in prosecutions because the police know that the witness incorrectly identified a foil who could not have committed the crime" (p. 821). Presumably for that reason, Wells and Lindsay (1985) once argued that "Eyewitness confidence in foil identifications, although of potential theoretical interest, should not be included in the forensically relevant calculations of confidence–accuracy relationships" (p. 413). Precisely the same argument applies to the method used to evaluate the diagnostic accuracy of lineup procedures, whether that method consists of computing a diagnosticity ratio or performing ROC analysis.

Nevertheless, even if one strongly believes that filler IDs should be included when computing correct and false ID rates, the question still remains as to whether the diagnosticity ratio or ROC analysis identifies the superior lineup procedure. As explained next in connection with another common criticism, either way, ROC analysis is undoubtedly the better way to go.

2. Criticism #2: Eyewitness identification ROCs are fundamentally different than ROCs in diagnostic medicine

ROC analysis is widely used in diagnostic medicine, but one might imagine that trying to use the same approach in eyewitness identification is problematic because lineups (unlike medical tests) have fillers. However, medical science has long used ROC analysis to study diagnostic decision-making using a procedure that is conceptually identical to a lineup. In the relevant medical studies, radiologists have been asked to identify the location of a tumor (if present) in one of four quadrants of an X-ray (Starr, Metz, Lusted, & Goodenough, 1975; Swets & Pickett, 1982) – much like an eyewitness is asked to identify the location of a perpetrator (if present) in one of 6 positions of a lineup. In this detection-plus-quadrant-localization task, the observer is presented with either a "target-present" X-ray (in which a tumor is present in one of the four locations) or a "target-absent" X-ray (in which no tumor is present in any of the four quadrants). In a target-present X-ray, the quadrant containing the tumor is analogous to the guilty suspect in a target-present lineup, and the other three quadrants are analogous to fillers. In a target-absent X-ray, all four quadrants are analogous to fillers. Thus, a target-absent X-ray is like a fair target-absent lineup.

In a task like this, the hit rate plotted on the vertical axis of the ROC can be computed in either of two ways: (1) by giving credit for *all* identifications made from a target-present X-ray (i.e., by counting "guilty suspect" IDs of the quadrant containing the tumor as well as "filler" IDs of the other quadrants), or (2) by giving credit only for "guilty suspect" IDs of the quadrant containing the tumor. The false alarm rate plotted on the horizontal axis is computed by counting all IDs from target-absent X-rays, regardless of the quadrant chosen (i.e., but counting all "filler" IDs).

Using the first approach, the hit and false alarm rates yield a typical-looking ROC in that the curve extends from the origin, where both the hit rate and the false alarm rate equal 0, to the upper right corner of the unit square, where both the hit rate and the false alarm rate equal 1.0. Fig. 2A shows an example using radiology data estimated from Fig. 2D of Starr et al. (1975). Using the second approach (in which only correct "guilty-suspect" IDs are counted from target-present X-rays), the hit and false alarm rates yield a second kind of ROC known as a "location" ROC (LROC). The LROC, which is also shown in Fig. 2A, is a less typical-looking ROC because it does not project to the upper right corner. However, it looks just like a typical lineup ROC. In fact, the LROC plot is directly analogous to lineup ROCs that have recently been used in eyewitness identification research (which only count suspect IDs from target-present lineups when computing the hit rate).

For comparative purposes, Fig. 2B shows ROC and LROC plots using eyewitness lineup data reported by Palmer, Brewer, Weber, and Nagesh (2013). The ROC data in Fig. 2B count all eyewitness IDs from target-present and target-absent lineups (whether to suspects or fillers, which differs from how lineup ROCs have been reported to date), whereas the LROC data only count correct suspect IDs from target-present lineups. What we have referred to as ROCs in the eyewitness identification literature correspond to what others have referred to as LROCs in the radiology literature. The equivalent of an LROC in the basic perception literature goes by several different names, such as *joint-detection-and-identification* ROCs (e.g., Swets, Green, Getty, & Swets, 1978) or 1-of-*m* ROCs (Green, Weber, & Duncan, 1977). The only slight and inconsequential difference between how the lineup LROC is depicted in Fig. 2B and how lineup ROCs have been depicted in our previous work is that the values on the lower *x*-axis in Fig. 2B have not been divided by lineup size to estimate the false (innocent suspect) ID rate. Dividing by lineup size to estimate the false ID rate yields the values shown on the upper *x*-axis in Fig. 2B. The decision to report the target-absent filler ID

Please cite this article in press as: Wixted, J. T., & Mickes, L. Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition* (2015), http://dx.doi.org/10.1016/j.jarmac.2015.08.009

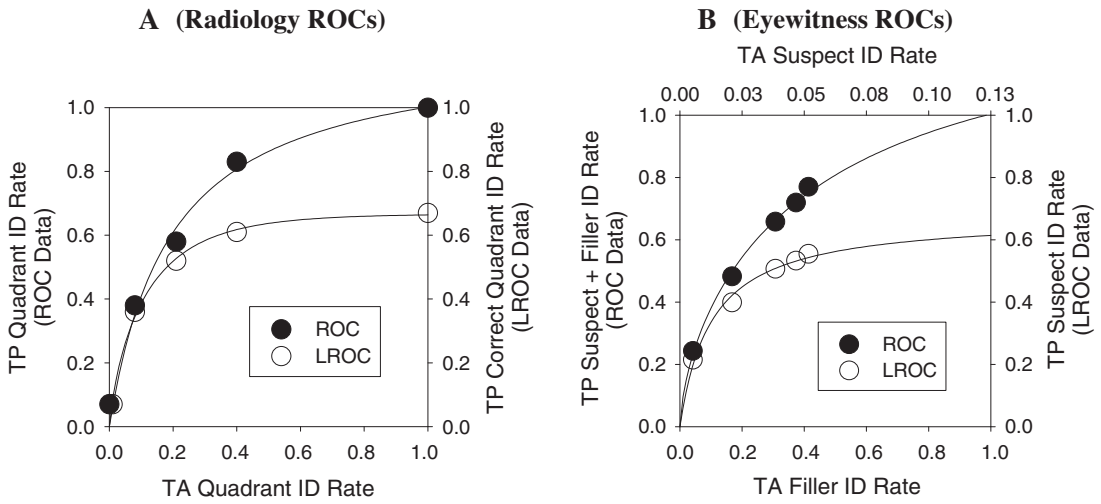

Fig. 2. (A) Detection-plus-identification data estimated from Fig. 2D of Starr et al. (1975). The receiver operating characteristic (ROC) data count all IDs made from tumor-present and tumor-absent X-rays. Thus, it is a plot of the target-present (TP) Quadrant ID rate vs. the target-absent (TA) Quadrant ID Rate. The location receiver operating characteristic (LROC) data differ in that they only count IDs made from the correct quadrant in tumor-present lineups. (B) Detection-plus-identification data from Experiment 1 (combined across conditions) of Palmer et al. (2013). The ROC data count all IDs made from target-present and target-absent lineups. The LROC data differ in that they only count correct (guilty-suspect) IDs made from target-present lineups. The false alarm rate is computed from all TA filler IDs (bottom horizontal axis) but, as is commonly done, can be transformed into an estimated TA (false) suspect ID rate by dividing the axis values by lineup size.

rate or the estimated target-absent suspect ID rate on the lower *x*-axis does not change the ROC data in any way. All that differs is the numbers reported on the lower *x*-axis.

When we compared simultaneous and sequential lineups (Mickes, Flowe, & Wixted, 2012), we reported what might be called LROCs, and we showed the estimated innocent suspect ID rate on the lower *x*-axis by dividing the filler ID rate from target-absent lineups by lineup size. However, we could have just as easily shown the filler ID rate from target-absent lineups on the lower *x*-axis, as in Fig. 2B, and no conclusions would change. In fact, we take this opportunity to reproduce our data with that simple modification in Fig. 3A, with the false ID rate (here called the "TA Suspect ID Rate") that we previously reported on the lower *x*-axis now shown on the upper *x*-axis. Obviously, the results are not affected by which numbers one chooses to report on the lower *x*-axis.

If we had regarded filler IDs (from target-present and target-absent lineups alike) as being important to analyze, we could have gone so far as to plot the data by giving credit for all IDs from target-present lineups (whether to fillers or suspects) as well as counting all IDs from target-absent lineup. In that case, no filler IDs would be ignored. We take this opportunity to do just that in Fig. 3B, where a simultaneous superiority effect is still apparent. Indeed, the whole point of Starr et al. (1975) was to show that one can determine the diagnostically more accurate procedure either way (because you get the same answer either way). We see no reason to count filler IDs from target-present lineups (or from target-absent lineups except in the service of estimating the innocent suspect ID rate) because, in our view, suspect IDs are of overriding importance. Still, one's evaluation of the relative diagnostic accuracy of competing lineup procedures is not affected by whether or not filler IDs are counted. Thus, even if you count filler IDs, ROC analysis is superior to using the diagnosticity ratio to identify the better lineup procedure.

The key point is that the methodology used to compute eyewitness identification ROCs is not new and is not troubled in any way by the presence of fillers. The potential application of LROC analysis to lineups has long been recognized in the medical literature and in the basic perception literature even though it has only recently been implemented by us and others in studies of eyewitness identification. In their classic signal-detection text, Macmillan and Creelman (1991) discuss the LROC (which they refer to as the *identification operating characteristic*, or IOC) and specifically point out that "Among the many possible implementations is eyewitness examination in a police lineup" (p. 251). Thus, the details of how to perform eyewitness ROC analysis (and radiology LROC analysis) were worked out long ago by the leading experts in the field of signal-detection theory.

3. Criticism #3: The diagnosticity ratio is what the legal system wants to know

The diagnosticity ratio has its place, but it has no role to play when it comes to determining which diagnostic procedure is superior to the other. Wells et al. (2015) state that ". . .the question in the real world is: What is the probability that the suspect is the culprit given that he was identified by the witness? That is the question that is of interest to police, prosecutors, judges, and jurors" (Wells et al., 2015, p. 102). True, but it is not the question of interest to policymakers charged with deciding which procedure to use (e.g., whether to use a simultaneous or a sequential lineup procedure).

In Bayesian terms, the probability that the suspect is the culprit given that he was identified by the witness is the *posterior probability* of guilt. In medicine, the analogous value is known as the *positive predictive value* of a test (namely, the probability that a patient with a positive test result actually has the disease). Using Bayes' theorem, positive predictive value is jointly determined by the base rate of guilt and the diagnosticity ratio (also known as the positive likelihood ratio). Thus, for a given base rate, the higher the diagnosticity ratio, the higher the odds that an identified suspect is guilty.

Bayesian measures, such as the posterior probability of guilt, and ROC analysis *address different questions*. The two questions they separately address are:

1. What is the probability that a particular suspect who has been identified from a simultaneous or a sequential lineup is guilty?
2. Which procedure is diagnostically superior, a simultaneous lineup or a sequential lineup?

The first question has to do with the predictive value of a test result for an *individual suspect* who has been identified. This is the question judges and juries who are dealing with an individual suspect care about. The second question has to do with which lineup

Please cite this article in press as: Wixted, J. T., & Mickes, L. Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition* (2015), http://dx.doi.org/10.1016/j.jarmac.2015.08.009

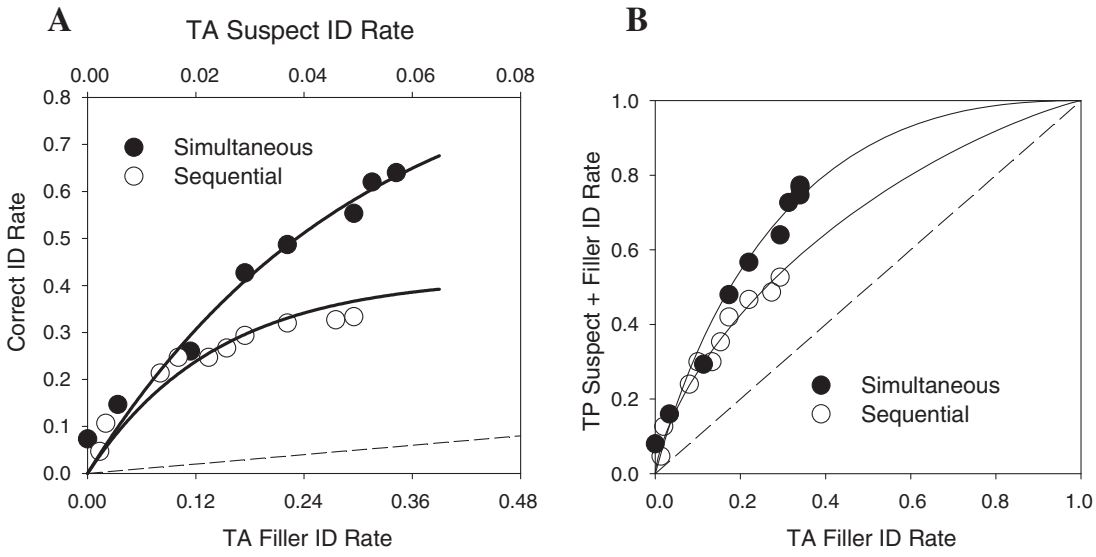

Fig. 3. (A) Simultaneous vs. sequential ROC data reported by Mickes et al. (2012) in the form of what in radiology would be called an LROC. (B) Simultaneous vs. sequential ROC data reported by Mickes et al. (2012) now in the form of what in radiology would be called an ROC because filler IDs are counted from both target-present (TP) and target-absent (TA) lineups.

procedure, when put into general use, better sorts innocent vs. guilty suspects into their proper categories. This is the question policymakers care about, and only ROC analysis can provide the answer.

The distinction between the two questions presented above has been understood in the medical literature for many years. For example, Zweig and Campbell (1993) noted that "Predictive value is more useful for interpreting a given result than for describing test performance" (p. 573). However, the distinction has only recently been addressed in the eyewitness identification literature. Mickes (2015) points out that Question 1 above is a question that usually pertains to estimator variables that affect eyewitness memory. From the perspective of judges and juries, lineup format is an estimator variable. However, from the perspective of policymakers, lineup format is a system variable.

Table 2 provides a summary of the measurements that are relevant when lineup format is construed as an estimator variable or as a system variable. As an estimator variable, the question pertains to the probability (or odds) that an identified suspect is guilty. Only the diagnosticity ratio (when multiplied by the base rate) can provide that information. However, for any lineup procedure, the diagnosticity ratio can be arranged to be low or high (depending on whether a liberal or conservative decision rule is used). Thus, even a diagnostically inferior lineup procedure can yield a high diagnosticity ratio – and a high posterior odds of guilt – if a conservative enough criterion is used. That being the case, a Bayesian analysis (i.e., the diagnosticity ratio multiplied by the base rate of guilt) does not indicate which procedure is diagnostically superior. To ask which procedure is diagnostically superior is to ask about lineup format as a system variable. Only ROC analysis can answer the system-variable question of which lineup procedure enables a population of eyewitnesses to more accurately sort innocent and guilty suspects into their respective categories.

Table 2
Measurements relevant to the analysis of lineup format construed as an estimator variable or as a system variable.

Measurement question	Lineup format as an estimator variable	Lineup format as a system variable
Are base rates relevant?	Yes	No
Is the diagnosticity ratio relevant?	Yes	No
Is Bayes' Theorem relevant?	Yes	No
Is ROC analysis relevant	No	Yes

4. Criticism #4: ROC analysis measures "psychological" discriminability and response bias

Another common misconception was recently expressed by Wells et al. (2015) when they said: "But the idea behind the ROC approach is to examine differences in *psychological discriminability* independently of response bias" (Wells et al., 2015, p. 108, emphasis added). Later in that same paper they said: "In fact, however, it is not clear that the ROC approach is properly controlling for response bias or that it measures discriminability" (Wells et al., p. 118). The idea that the purpose of ROC analysis is to "examine differences in psychological discriminability independently of response bias" needs to be nipped in the bud.

ROC analysis can be used either for applied purposes to measure how well a diagnostic procedure accurately differentiates between two states of the world (no theoretical considerations are involved) or for theoretical purposes to measure discriminability and response bias in the mind of a participant (which clearly depends on theoretical considerations). In diagnostic medicine, and in eyewitness identification, ROC analysis is typically used in the former way. That is, the goal is to identify the more accurate diagnostic procedure regardless of how any theory interprets the results. If one procedure is capable of producing a higher correct ID rate for any given false ID rate than another (i.e., if one procedure yields a higher ROC), it is the superior diagnostic procedure for applied purposes regardless of what any theory might tell you.

In cognitive psychology, ROC analysis is more commonly used to measure theoretical constructs and to test cognitive models of recognition memory. In fact, we have conducted ROC analysis to test theories for years (e.g., Mickes, Wixted, & Wais, 2007). But observable discriminability and unobservable (i.e., theoretical) discriminability are distinct issues and must not be conflated because they are sometimes dissociable. For example, old/new recognition and two-alternative forced-choice (2AFC) recognition yield different observable ROCs (invariably favoring the forced-choice

Please cite this article in press as: Wixted, J. T., & Mickes, L. Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition* (2015), http://dx.doi.org/10.1016/j.jarmac.2015.08.009

procedure). Thus, if one had the choice of using an old/new or a 2AFC forced-choice recognition procedure in an applied situation, it is obvious that the 2AFC procedure would be preferred. Nevertheless, psychological (i.e., theoretical) discriminability is approximately the same for both testing formats (Jang, Wixted, & Huber, 2009).

Similar considerations apply to response bias, which also has both behavioral (theory-free) and psychological (theory-dependent) interpretations. Psychologically, response bias is determined by where the participant's decision criterion is placed with respect to the underlying target and lure distributions. In an eyewitness paradigm, the location of this criterion can be theoretically shifted in the conservative direction by adding admonishments to lineup instructions (e.g., "the person you saw commit the crime may or may not be in the lineup"). Doing so will create a new, more conservative correct and false ID rate pair. However – and this is the key point – one can create the same conservative outcome using methods that *do not affect the participant's psychological criterion at all*. Increasing the number of lineup members has exactly this effect. When computing correct and false ID rates from suspect IDs only, the larger the lineup size, the fewer correct and false IDs there will be because more eyewitnesses will choose fillers as the opportunity to do so increases. This effect is not only empirically observed, it is also naturally predicted by any signal-detection model even when the criterion remains fixed as a function of lineup size. The result of increasing lineup size would still be a more conservative point on the ROC (lower correct and false ID rates) even if the eyewitnesses themselves did not adopt a more conservative decision criterion (i.e., even if "psychological response bias" remains unchanged).

Confusion over the distinction between measured (i.e., behavioral) response bias – which is all the legal system cares about – and psychological response bias (of no interest at all to the legal system) is not limited to the field of eyewitness memory. The same issue has caused confusion in the basic memory literature concerned with the high false alarm rates obtained using the DRM procedure (Wixted & Stretch, 2000) and in the perception literature concerned with the issue of cross-modal priming (Witt, Taylor, Sugovic, & Wixted, 2015).

5. Conclusion

Although we are currently debating the merits of ROC analysis applied to eyewitness identification procedures, it seems important to consider that (1) this analytical methodology was worked out long ago by the most influential signal-detection theorists of our time, (2) it has long been used for medical diagnostic procedures that are conceptually identical to lineup procedures, and (3) it was recently endorsed in preference to the diagnosticity ratio by a National Academy of Sciences committee. Eyewitness ID researchers who are standing in opposition to ROC analysis may have identified a key flaw that all of these other scientists have somehow overlooked, but they might instead be laboring under misconceptions that are standing in the way of seeing what it has to offer.

Conflict of interest statement

The authors declare that they have no conflict of interest.

References

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science*, *7*, 238–259.
Egan, J. P. (1958). *Recognition memory and the operating characteristic. (Tech Note AFCRC-TN-58-51)*. Bloomington, IN: Indiana University, Hearing and Communication Laboratory.
Green, D. M., Weber, D. L., & Duncan, J. E. (1977). Detection and recognition of pure tones in noise. *Journal of the Acoustic Society of America*, *62*, 948–954.
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., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychology Science*, *23*, 3–10.
Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology*, *138*, 291–306.
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.
Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.
Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2009). Public policy and sequential lineups. *Legal and Criminological Psychology*, *14*, 1–12.
McQuiston-Surrett, D., 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.
Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*, 93–102.
Mickes, L., 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., 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.
National Research Council. (2014). *Identifying the Culprit: Assessing Eyewitness Identification*. Washington, DC: The National Academies Press.
Palmer, M., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence–accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*, 55–71.
Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy and Law*, *1*, 817–845.
Police Executive Research Forum. (2013). *A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies*. Retrieved from: http://www.policeforum.org/
Schacter, D. L., Dawes, R., Jacoby, L. L., Kahneman, D., Lempert, R., Roediger, H. L., et al. (2008). . *Law and Human Behavior* (vol. 32).
Starr, S. J., Metz, C. E., Lusted, L. B., & Goodenough, D. J. (1975). Visual detection and localization of radiographic images. *Radiology*, *116*, 538–553.
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.
Swets, J. A., Green, D. M., Getty, D. J., & Swets, J. B. (1978). Signal detection and identification at successive stages of observation. *Perception & Psychophysics*, *23*, 275–289.
Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.
Wells, G. L., & Lindsay, R. C. L. (1985). Methodological notes on the accuracy–confidence relation in eyewitness identifications. *J. Appl. Psychol.*, *70*, 413–419.
Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior*, *39*, 99–122.
Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception*, *44*, 289–300.
Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*, 262–276.
Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, *107*, 368–376.
Zweig, M. H., & Campbell, G. (1993). Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561–577.

Please cite this article in press as: Wixted, J. T., & Mickes, L. Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition* (2015), http://dx.doi.org/10.1016/j.jarmac.2015.08.009