Missing the information needed to perform ROC analysis?

Then compute *d′*, not the diagnosticity ratio

Laura Mickes[1], Molly B. Moreland[2], Steven E. Clark[2], and John T. Wixted[3]

[1]Royal Holloway, University of London

[2]University of California, Riverside

[3]University of California, San Diego

Author Note

Laura Mickes, Department of Psychology, Royal Holloway, University of London. Molly Moreland, Department of Psychology, University of California, Riverside. Steven E. Clark, Department of Psychology, University of California, Riverside. John T. Wixted, Department of Psychology, University of California, San Diego.

This work was supported in part by the National Science Foundation SES-1155248 to John T. Wixted and Laura Mickes and SES 1061183 to Steven E. Clark. The content is solely the responsibility of the authors and does not necessarily reflect the views of the National Science Foundation.

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

Abstract

Recently, the argument has been made that receiver operating characteristic (ROC) analysis should be used to compare the diagnostic performance of different lineup procedures. However, a drawback to ROC analysis is that is requires multiple correct and false ID rates for each lineup procedure being compared. When only a single pair of correct and false ID rates is measured, what dependent measure should be used? Here, we contrast the use of d' with the diagnosticity ratio using the data reported by Carlson and Carlson (in press) and other previously reported data. Despite being based on a theory that was developed for list memory procedures, we show that, in practice, d' performs surprisingly well for lineup procedures. Moreover, d' far outperforms the diagnosticity ratio. We recommend that d' – not the diagnosticity ratio – be used as a dependent measure to compare the diagnostic performance of different lineup procedures.

Keywords: Receiver operating characteristic analysis, d', diagnosticity ratio, eyewitness identification, lineups

In a typical recognition memory experiment, the participant's task is to discriminate between "old" items that were previously encountered and "new" items that were not. Common examples include discriminating between words that were presented on a list and words that were not, or discriminating between guilty suspects who appeared in a mock-crime video and innocent suspects who did not. For eyewitness identification experiments, performance for guilty suspects (old items) and innocent suspects (new items) is characterized by the correct ID rate (the proportion of guilty suspects who are correctly identified), and the false ID rate (the proportion of innocent suspects who are incorrectly identified). What is the best way to combine these two measures to gauge overall recognition performance? Although the possibilities are limitless, this question usually boils down to a choice between some kind of ratio measure (e.g., correct ID rate / false ID rate) vs. some kind of difference score (e.g., correct ID rate minus false ID rate). The choice depends on one's goal, so the first issue to consider is what that goal should be.

In many previous eyewitness identification experiments, it has been assumed that the goal should be to estimate the posterior odds of guilt because, once a suspect is identified, what a court of law really wants to know is how likely it is that the identified suspect is guilty. That is precisely the kind of information that the *diagnosticity ratio* – correct ID rate / false ID rate – provides. If Lineup Procedure 1 yields a higher diagnosticity ratio than Lineup Procedure 2, then a suspect identified using Procedure 1 is more likely to be guilty than a suspect identified using Procedure 2. An alternative (and arguably far more important) goal is to characterize the ability of eyewitnesses to differentiate between innocent and guilty suspects, and that ability is usually measured using a difference score.

To understand why it is more important to measure the ability to tell the difference between guilty and innocent suspects than it is to measure the posterior odds of guilt, it is important to first appreciate the fact that there is nothing special about the singular correct and false ID rate pair obtained in any particular experimental condition. The performance of a given lineup procedure is characterized by an entire family of correct and false ID rate pairs, not by a single correct and false ID rate pair. After we illustrate that point, we will return to the question of what to do when all you have is a single pair of correct and false ID rates.

Imagine an experiment designed to investigate how well eyewitnesses perform when a simultaneous lineup is used to test their memory. If the instructions do not underscore the fact that the guilty suspect may not be in the lineup, the correct and false ID rates might be relatively high, such as correct ID rate = .50 and false ID rate = .10 (diagnosticity ratio = 5). However, using instructions that explicitly state that the guilty suspect may or may not be in the lineup, more conservative responding would likely result (Clark, 2005) and the correct and false ID rates might decrease to .42 and .07, respectively (diagnosticity ratio = 6). Instructions designed to induce even more conservative responding (e.g., telling the participant that false IDs are known to be a problem and that one should be wary of making any ID at all) might result in still lower correct and false ID rates of .32 and .04, respectively (diagnosticity ratio = 8). Which of those three correct and false ID rate pairs (and their corresponding diagnosticity ratios) characterizes the performance of the simultaneous lineup procedure? Considered in isolation, none of them do; instead, performance is characterized by the entire family of correct and false ID rate pairs as the tendency to make an ID varies across a wide range. A different family of correct and false ID rate pairs (and a different family of diagnosticity ratios) would characterize the performance of the sequential lineup procedure.

As illustrated using hypothetical data in Figure 1A, the family of points for each procedure constitutes the Receiver Operating Characteristic (ROC). The farther the points bow away from the diagonal line of chance performance, the better participants are at discriminating guilty suspects from innocent suspects. The degree to which the points bow away from the line of chance performance is measured by the partial area under the curve (pAUC), as illustrated in Figure 1B and Figure 1C. Note that when the target-absent lineup contains a designated innocent suspect and a fair lineup is used, the maximum false ID rate is 1/ *n*, where *n* is the lineup size. This is the false ID rate that would result if every witness who was presented with a target-absent lineup made an ID. Because the maximum false ID rate is less than 1, the rightmost extent of the area under the curve is correspondingly limited, hence the term "partial" AUC. In practice, measured pAUC values often seem curiously small (e.g., 0.05), and Figure 2 illustrates why. The reason why they are small is that pAUC values represent an area measure expressed as a proportion of the unit square ROC, with both axes ranging from 0 to 1 (Figure 2B). The fact that pAUC values are typically small does not limit their effectiveness in quantifying recognition memory performance associated with a lineup procedure.

The procedure that yields the higher pAUC is the objectively superior procedure (e.g., Procedure 1 in Figure 1A), and this is the critical point. It is objectively superior because it can be used to achieve a higher correct ID rate while, at the same time, achieving a lower false ID rate than the alternative procedure. For example, as you move to the left along the ROC associated with Procedure 2 in Figure 1A, choose the single ROC point that seems to you to represent the best tradeoff between the gain associated with a lower false ID rate and the cost associated with a lower correct ID rate. Next, consider the fact that the closest point above it and to the left on the ROC associated with Procedure 1 has both a higher correct ID rate and a lower

false ID rate. This is true of any ROC point that you might choose for Procedure 2. Hence, Procedure 1 is the objectively superior procedure. These considerations show why using ROC analysis to measure the pAUC is always the best approach to use when comparing the level of performance supported by different lineup procedures. However, many experiments report only a single correct and false ID rate pair for each condition, and what to do under those circumstances is the question of interest here.

If measuring the pAUC (based on a family of correct and false ID rate pairs) is the real goal, the question of what to do when only a single pair of correct and false ID rate pairs is measured has a simple answer: one should combine the correct and false ID rates in such a way as to provide the best approximation to the pAUC. This is where the role of theory usually comes into play. That is, theory does the work of inferring from a single pair of correct and false ID rates what the rest of the ROC would probably look like. The theory does that by providing the appropriate equation to use when trying to measure overall recognition memory performance from a single pair of correct and false ID rates. An accurate theory will provide an equation that makes a correct inference about the full ROC from that single pair, making the job of running an experiment easier. In experiments that use list memory designs, the theory that most often serves that role is signal detection theory (illustrated in Figure 3). This is the theory that gives rise to the formula needed to compute d' from a single pair of correct and false ID rates. The d' formula is not a ratio but is instead a difference score: d' = z(correct ID rate) – z(false ID rate)[1] (Macmillan & Creelman, 2005). If the theory on which the equation for d' is based is accurate, then a condition with a higher d' yield would also yield a higher pAUC (i.e., a higher ability to distinguish between old and new items) than a condition with a lower d'. Many experiments that use list-memory designs in experimental psychology do not use ROC analysis but instead collect

a single correct and false ID rate pair from each condition and then rely on signal-detection theory to compute a *d′* score for each condition. In essence, the theory saves the experimenter the work of actually performing ROC analysis.

Why not use the same approach when memory is tested using a lineup? That is, why not simply compute a *d′* score? Although much evidence suggests that the signal-detection model shown in Figure 3 usually provides a reasonable approximation to the truth when a list-memory design is used, that theory does not automatically apply when a lineup design is used. In other words, signal-detection theory is specifically written for an old/new recognition procedure, not for a lineup recognition procedure (Wixted & Mickes, 2014). Thus, one cannot automatically assume that computing *d′* from the standard formula from the correct and false ID rates obtained from a lineup will serve as an adequate proxy for the lineup pAUC. Then again, in practice, *d′* might work reasonably well despite being based on a theory that applies to a different recognition memory procedure. The three hypothetical correct and false ID rates presented earlier were chosen to illustrate how this might work. Although the three points yielded 3 different diagnosticity ratios, they all yield approximately the same value when *d′* = z(correct ID rate) – z(false ID rate) is computed:

$$d' = z(.50) - z(.10) = 1.28$$

$$d' = z(.42) - z(.07) = 1.27$$

$$d' = z(.32) - z(.04) = 1.28$$

Thus, in this hypothetical example, had the experimenter collected only one pair of correct and false ID rates, it would not matter very much which set of instructions had been used. The same answer would be obtained in each case. That is how it works when a theory provides a good

approximation to the entire ROC (and, therefore, a good approximation of the pAUC) from a single correct and false ID rate pair.

In practice, does d' work for lineups despite being based on a theory developed for a list-memory design? The answer, somewhat surprisingly, is that it does work well – much better than relying on a probative value measure like the diagnosticity ratio. This can be shown in two ways. First, we can examine experiments that reported ROC data and then compute d' from each of the multiple pairs of correct and false ID rates they reported. In this case, d' ought to remain constant because in ROC analysis, discrimination is held constant across all correct and false ID rate pairs (as in the example above). All that varies is how liberal or conservative responding is (i.e., all that varies is response bias). Second, we can use the Carlson and Carlson (in press) data to examine the relationship between d' and pAUC across conditions in which discriminability instead varies over a wide range. Carlson and Carlson reported correct ID rates, false ID rates and pAUC values across 12 different experimental conditions, which makes it possible to compute both d' and the diagnosticity ratio for each condition (based on the overall correct and false ID rates) to see how well they correlate with the corresponding pAUC scores.

The kind of data needed to perform the first type of analysis (i.e., ROC data with discriminability held constant across different levels of bias) were reported in a study by Brewer and Wells (2006). They used the simultaneous lineup procedure, and participants made confidence judgments using a 100-point confidence scale, with ratings of 100% indicating absolute certainty that the identified individual was the perpetrator and ratings of 1% indicating only slight confidence that the identified individual was the perpetrator. These data were previously used to explain how to perform ROC analysis using confidence ratings (Mickes et al., 2012) and to make the point that the diagnosticity ratio does not remain constant across different

levels of bias and so could not possibly be used to estimate pAUC (because there is only one pAUC, yet many different diagnosticity ratios, per ROC). Here, we again make use of those same data to show that, by contrast, *d′* does remain relatively constant across different levels of bias (so *d′* can be used to estimate pAUC).

Figure 4A shows the ROCs computed from the "Thief Lineups" and "Waiter Lineups" conditions reported in Table 9 of Brewer and Wells (2006; see Mickes, Flowe, & Wixted, 2012, for details). The correct ID and false ID rate pair plotted at the lower left of each ROC was computed by treating suspect identifications as correct IDs or false IDs only if they were made with a confidence of 90% or higher (anything less was treated as a non-ID). This point corresponds to the most conservative decision criterion. The remaining points on the ROC were computed by using ever lower (i.e., increasingly liberal) cutoff values on the confidence scale. For example, the next pair of correct and false ID rates was computed by treating as correct IDs or false IDs only those identifications made with a confidence rating of 70% or higher; the next point was based on identifications made with a confidence rating of 50% or higher, and so on. It is obvious from the figure that discriminability (i.e., pAUC) was higher in the Waiter condition than in the Thief condition. Thus, ideally, any measure of recognition memory performance computed from any single correct and false ID rate pair from each condition would reflect that fact.

The data in Figure 4B show the diagnosticity ratio values and *d′* values associated with each correct and false ID rate pair generated by the different confidence criteria in the Thief and Waiter conditions. Again, more conservative responding is represented by higher levels of confidence used to compute the correct and false ID rates. As responding becomes more conservative (i.e., as you move to the right on the x-axis), the diagnosticity ratio increases

dramatically. By contrast, *d′* remains essentially constant, as it should. Thus, instead of performing ROC analysis, one could have done almost as well by collecting a single pair of correct and false ID rates in each condition (one pair for the Thief condition and one pair for the Waiter condition) and using *d′* as the dependent measure. The *d′* from any singular correct and false ID rate pair from the Waiter ROC could be compared to the *d′* from any singular correct and false ID rate pair from the Thief ROC, and the correct answer would result (i.e., discriminability would be judged to be higher in the Waiter Condition). This is important because in an experiment that does not collect confidence ratings, one cannot be sure where on the ROC the singular correct and false ID rate pairs from a given condition actually falls. Using *d′*, it would not matter much – the Waiter condition would be judged superior to the Thief condition regardless. By contrast, using the diagnosticity ratio as the dependent measure would be problematic because it conflates discriminability with response bias. For example, if one condition happened to yield more conservative responding than the other, the more conservative condition could be mistakenly judged to be superior because of its higher diagnosticity ratio. But that condition could actually be inferior in terms of discriminating innocent suspects from guilty suspects (i.e., that condition could be associated with a lower *d′* and a lower pAUC). These considerations may explain why, in the past, the sequential procedure (which tends to induce conservative responding) has sometimes been judged to be diagnostically superior to the simultaneous procedure.

Using the data reported by Brewer and Wells (2006), it is clear that *d′* provides the right answer as to which of the two conditions yielded higher discriminability. The data reported by Carlson and Carlson (in press) can be used to perform a similar test across many more conditions. For practical purposes, this is the key test because experimenters who collect a single

pair of correct and false ID rates per condition usually want to know if one condition supports better performance than the other. The two conditions reported by Brewer et al. (2006) are encouraging in that the one with the higher pAUC was also associated with the higher *d′*, but does that story hold up when a much larger range of conditions are used? The Carlson and Carlson data are unique in helping to answer that question because they ran 12 different conditions[2]. Figure 5A shows a plot of *d′* (computed from the overall correct and false ID rates) vs. pAUC across the 12 conditions, and Figure 5B shows a plot of the diagnosticity ratio (again, computed from the overall correct and false ID rates) vs. pAUC across the same 12 conditions. Obviously, *d′* does a good job of estimating pAUC – much better than the diagnosticity ratio does. *d′* accounts for 84% of the variance in pAUC scores across conditions (r = .92), whereas the diagnosticity ratio accounts for only 50% of the variance (r = .71). Figure 5B shows a possible outlier in the diagnosticity ratio graph, but the results favor *d′* even when that condition is removed from both plots (87% of the variance accounted for using *d′* vs. 65% of the variance accounted for using the diagnosticity ratio). The fact that the diagnosticity ratio is positively correlated with pAUC makes sense because, theoretically, that measure should be sensitive to both discriminability and response bias (Wixted & Mickes, 2014). But that is its main problem. What is needed is a measure that does not change when all that changes is response bias. The diagnosticity ratio is clearly inadequate in that respect, whereas *d′* performs very well.

One question that remains concerns the extent to which statistical conclusions based on *d′* correspond to statistical conclusions based on pAUC. Eyewitness identification experiments present a unique challenge because such comparisons involve only two *d′* scores, one for each experimental condition, rather than a distribution of *d′* scores for each condition. However, a method for comparing two *d′* scores is described by Gourevitch and Galanter (1967). To evaluate

the correspondence in statistical inference tests we compared *G* statistics from Gourevitch and Galanter with ROC-based *D* statistics from Robin et al. (2011) for simultaneous-sequential lineup comparisons from Carlson and Carlson (in press), Gronlund et al. (2012), and Mickes et al. (2012). The results in Table 1 show a strong correspondence between *G* based on statistical comparisons of *d′* and *D* based on statistical comparisons of pAUC. The correlation between *D* and *G*, $r_{D,G}$ is .95. Thus, whether *d′* or pAUC is used, the statistical conclusions will often be similar.

Clark (2012) reviewed the pre-ROC simultaneous vs. sequential empirical literature using *d′* as the dependent measure and found that the two procedures yielded essentially identical scores on average (cf. Palmer & Brewer, 2012). This outcome seems inconsistent with a series of recent ROC analyses – including the new ROC analysis by Carlson and Carlson (in press) – that consistently show a statistically significant simultaneous advantage. One possible explanation for the inconsistency is provided by McQuiston, Malpass, and Tredoux (2006), who found that studies from the Lindsay lab did not report balancing suspect position across early and late sequential lineup positions (unlike many other labs) and consistently obtained an unusually strong advantage for the sequential procedure compared to other labs. Whether or not this explains the apparent discrepancy between *d′*-based analyses and more recent ROC-based analyses remains to be seen.

The take-home message is simply this: when only a single pair of correct and false ID rates is collected, *d′* should be computed, not the diagnosticity ratio. It would always be better to perform the full ROC analysis because even in list memory designs, ROC analysis shows that conclusions based on the theoretical *d′* measure are sometimes wrong (see Dougal & Rotello, 2007, for an example). On those occasions when ROC analysis and *d′* disagree, conclusions must

be based on the theory-free ROC analysis. Nevertheless, the analyses we have presented here indicate that the standard, often-used statistic for list memory experiments seems appropriate for eyewitness lineup experiments as well. Generally speaking, for the evaluation of two eyewitness identification procedures, it seems reasonable to compute d' from a single pair of correct and false identification rates and interpret the results based on that measure, but it is a mistake to compute the diagnosticity ratio from a single pair of correct and false identification rates and interpret the results based on that measure.

References

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

Carlson, C. A. & Carlson, M. A. (in press). An Evaluation of Perpetrator Distinctiveness, Weapon Presence, and Lineup Presentation using ROC Analysis. *Journal of Applied Research in Memory and Cognition.*

Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior, 29,* 395-424.

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

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review, 14*, 423-429.

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.

Gourevitch, V., & Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika*, 32, 25-33.

Macmillan N. A. & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

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

Mickes, L., Flowe, H. D. & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361-376.

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

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.

Wixted, J. T. & Mickes, L. (in press). A Signal-Detection-Based Diagnostic-Feature Model of Eyewitness Identification. *Psychological Review.*

Footnote

1. Using Excel, d' = normsinv(correct ID rate) – normsinv(false ID rate). Using MATLAB, d' = norminv(correct ID rate) – norminv(false ID rate). Using R, d' = qnorm(correct ID rate) – qnorm (false ID rate).

2. pAUC values are sensitive to the specified false ID rate range. Thus, this range needs to be equated when comparing the pAUC values across conditions. Fortunately, Carlson and Carlson (in press) did just that for all of the pAUC values reported in their Table 3. That is, the same false ID rate range was used for all 12 conditions.

Figure Captions

Figure 1. **A**. Receiver Operating Characteristic (ROC) plots for datasets from two hypothetical lineup procedures (the same data are potted in panels A, B and C). **B**. Illustration of the partial area under the curve (pAUC) for Procedure 1. The shaded region shows the false ID rate cutoff (at the rightmost point on that ROC curve). **C**. Illustration of the pAUC for Procedure 2. The shaded region shows that same false ID rate cutoff, which necessarily extends past the rightmost point on that ROC curve. Given the cutoff used for Procedure 1, this same cutoff would be used in pAUC analyses two compare the two procedures[2]. The dashed line represents chance performance.

Figure 2. An illustration of why partial area under the curve (pAUC) values are small. **A**. An illustration of the pAUC for Procedure 1 using a truncated range for the false ID rate axis (which ranges from 0 to 0.10). **B**. When the shaded area is shown on the full unit square ROC (with both axes ranging from 0 to 1), it becomes clear that the pAUC represents less than 5% of the entire area. Thus, pAUC < .05. The dashed line represents chance performance.

Figure 3. An illustration of the standard signal-detection model.

Figure 4. **A**. Receiver Operating Characteristic (ROC) plots the data presented in Table 9 of Brewer and Wells (2006). The dashed line represents chance performance. **B**. The diagnosticity ratio (left vertical axis) and d' (right vertical axis) for different criterion levels of confidence in the Thief and Waiter conditions.

Figure 5. Scatterplot of d' vs. pAUC across the 12 conditions from Carlson and Carlson (in press) in Figure 5A. Scatterplot of the diagnosticity ratio vs. pAUC across the same 12 conditions in Figure 5B.

Table 1

Statistical Inference Based on *D* from pAUC analysis and *G* from *d'* analysis

	D	*p*	*G*	*p*
Carlson & Carlson (in press)	1.97	0.048	1.89	0.058
Gronlund et al. (2012)				
Suspect in position 2	2.96	0.003	3.61	0.0003
Suspect in position 5	1.34	0.181	0.98	0.328
Mickes et al. (2012)				
Experiment 1a	2.02	0.043	2.53	0.011
Experiment 1b	0.70	0.484	1.17	0.242
Experiment 2	0.40	0.688	0.34	0.737

Figure 1

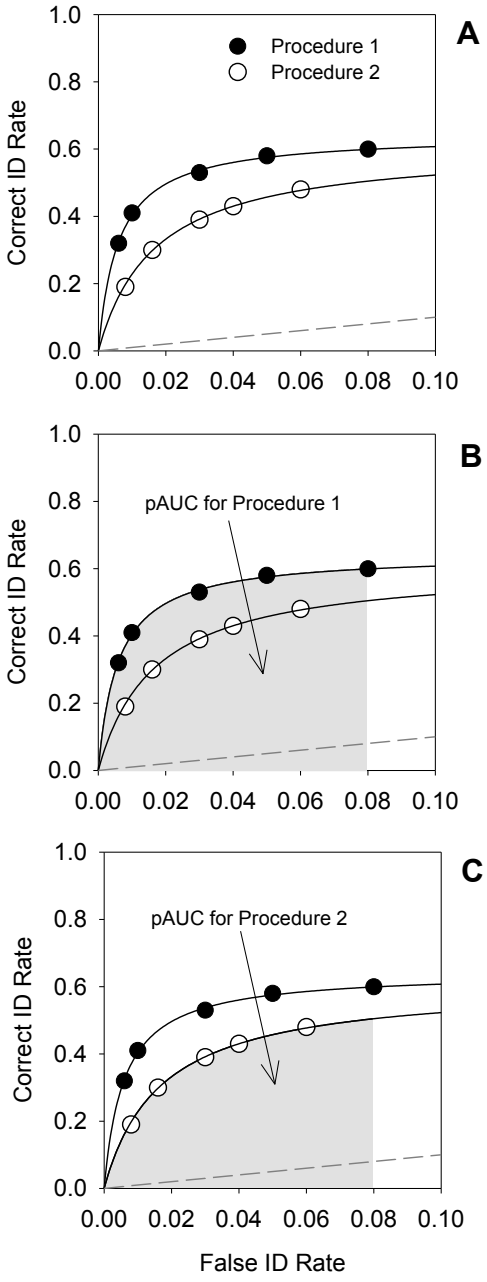

A
Procedure 1
Procedure 2
Correct ID Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.00 0.02 0.04 0.06 0.08 0.10
B
pAUC for Procedure 1
C
pAUC for Procedure 2
False ID Rate

Figure 2

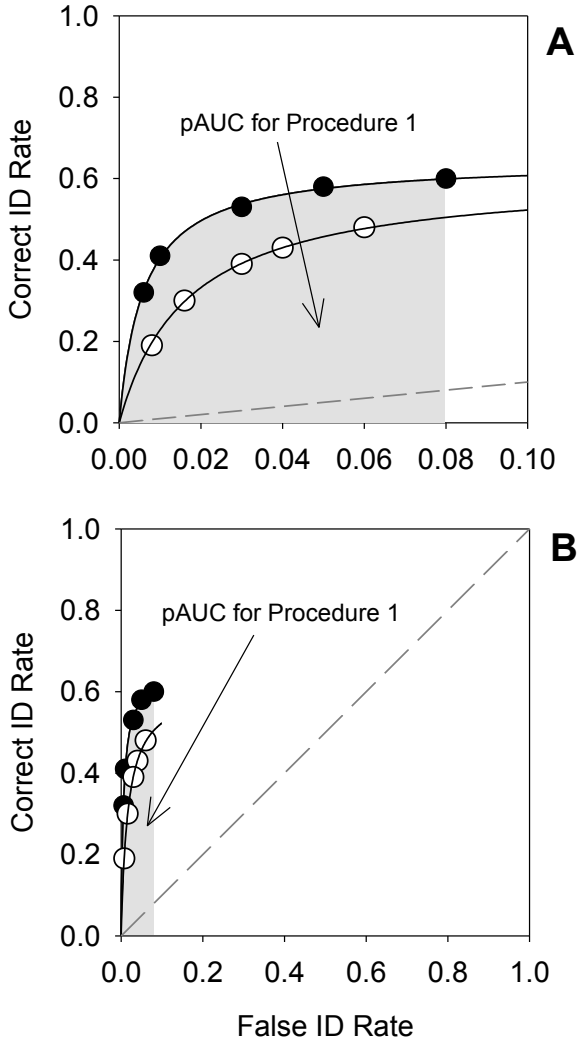

A
pAUC for Procedure 1
Correct ID Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.02
0.04
0.06
0.08
0.10
B
pAUC for Procedure 1
Correct ID Rate
False ID Rate
0.0
0.2
0.4
0.6
0.8
1.0

Figure 3

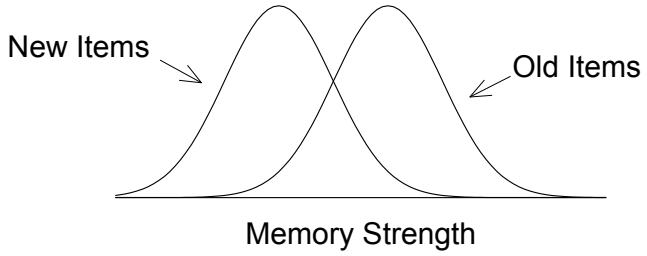


Figure 4

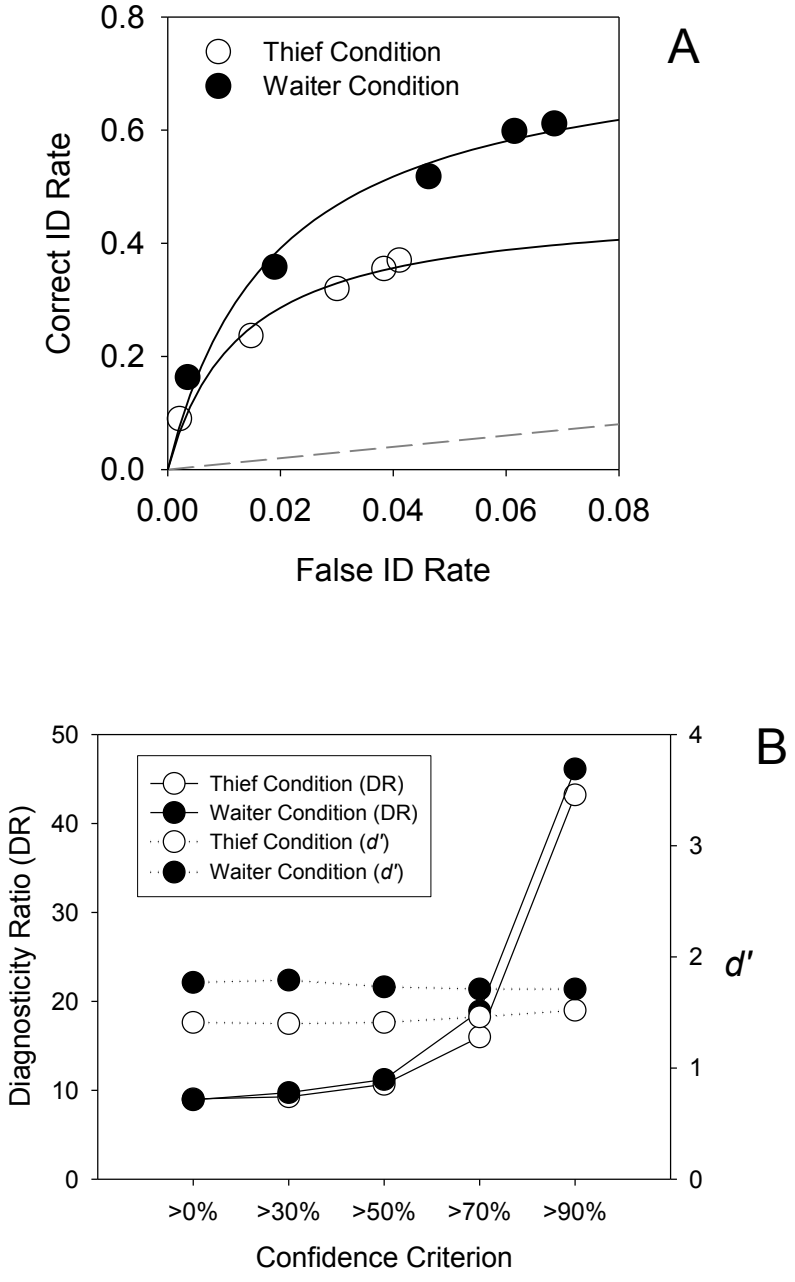
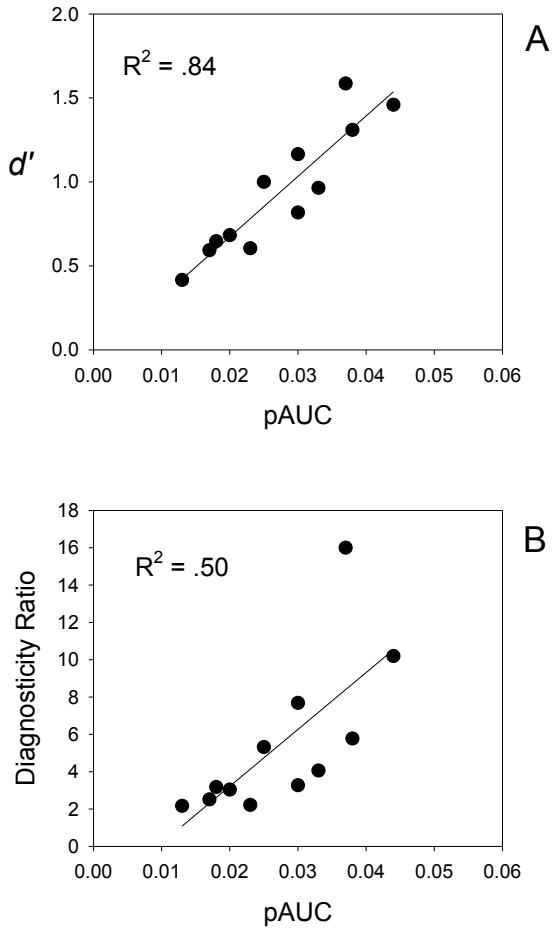

A
Thief Condition
Waiter Condition
Correct ID Rate
False ID Rate
0.0
0.2
0.4
0.6
0.8
0.00
0.02
0.04
0.06
0.08
B
Thief Condition (DR)
Waiter Condition (DR)
Thief Condition (d′)
Waiter Condition (d′)
Diagnosticity Ratio (DR)
d′
0
10
20
30
40
50
0
1
2
3
4
>0%
>30%
>50%
>70%
>90%
Confidence Criterion

Figure 5


A
R^2 = .84
d′
2.0
1.5
1.0
0.5
0.0
0.00
0.01
0.02
0.03
0.04
0.05
0.06
pAUC
B
R^2 = .50
Diagnosticity Ratio
18
16
14
12
10
8
6
4
2
0
0.00
0.01
0.02
0.03
0.04
0.05
0.06
pAUC