Journal of Applied Research in Memory and Cognition 3 (2014) 54–57

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

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

Eyewitness identification discriminability: ROC analysis versus logistic regression☆

Scott D. Gronlund[a,*], Jeffrey S. Neuschatz[b]

[a] University of Oklahoma, United States
[b] The University of Alabama, Huntsville, United States

ARTICLE INFO

Article history:
Received 14 April 2014
Received in revised form 22 April 2014
Accepted 23 April 2014
Available online 29 May 2014

Keywords:
Eyewitness identification
ROC analysis
Signal detection theory
Logistic regression
Probative value measures

ABSTRACT

To reach conclusions regarding the respective accuracy of two conditions, eyewitness researchers evaluate correct and false identification rates computed across participants. Two approaches typically are employed. One approach relies on ratio-based probative value measures; but Wixted and Mickes (2012) and Gronlund, Wixted, and Mickes (2014) showed that these measures fail to disentangle an assessment of accuracy (i.e., discriminability between guilty and innocent suspects) from response bias (i.e., a willingness to make a response). Our focus is on a second approach, logistic regression analyses of the correct and of the false identification rates. Logistic regression also fails to disentangle discriminability from bias. Therefore, it only can denote the most accurate condition in limited circumstances. The best approach for reaching the proper conclusion regarding which condition is most accurate is to use receiver operator characteristic (ROC) analysis. Simulated ROC data illustrate the problem with a reliance on logistic regression to assess accuracy.

1. Eyewitness identification data: ROC analysis versus logistic regression

A standard eyewitness lineup test includes a target-present and a target-absent lineup. The former contains the guilty suspect and several foils (known innocents); the latter contains a designated innocent suspect and several foils. In most experiments, an eyewitness selects someone from the lineup or indicates that the perpetrator is not present by rejecting the lineup. A correct identification (ID) is made if the witness selects the guilty suspect from the target-present lineup; a false ID is made if the witness selects the innocent suspect from the target-absent lineup. To determine if the performance elicited by condition A (e.g., a sequential lineup) is superior to the performance elicited by condition B (e.g., a simultaneous lineup), the correct and false ID rates typically are analyzed by conducting some form of log-linear analysis (e.g., logistic regression) or by computing a measure of probative value (and usually both).

The goal of this paper is to show that logistic regression is a problematic analytic tool because it fails to disentangle an assessment of accuracy (i.e., discriminability) from the contribution of response bias. Consequently, it often will not allow a researcher to determine which condition results in the best performance. We begin with an example that makes clear the distinction between discriminability and response bias. Signal-detection theory addresses this issue in basic recognition memory research, but because only one observation typically is collected in an eyewitness experiment, signal-detection based measures of discriminability and response bias cannot be computed on a per-participant basis. Therefore, researchers jointly consider correct and false ID rates computed across participants as probative value measures, and statistically, researchers perform logistic regression analyses on the overall correct and false ID rates. As we shall see, both these analytic methods are problematic.

2. Discriminability, response bias, and signal detection theory

Assume that there are two versions of an exam. In Exam A, each correct response is awarded +1 and each error −1. In Exam B, each correct response is awarded +1 and each error −10. If I randomly

☆ This work was supported by the National Science Foundation (NSF) grant SES-1060902 to SDG and SES-1060921 to JSN. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the NSF. We thank Curt Carlson and John Wixted for their comments on a draft of this manuscript.

* Corresponding author at: University of Oklahoma, Department of Psychology, 455 W. Lindsey Street, Norman, OK 73019-2007, United States. Tel.: +1 4053254553.
E-mail address: sgronlund@ou.edu (S.D. Gronlund).

http://dx.doi.org/10.1016/j.jarmac.2014.04.008
2211-3681/

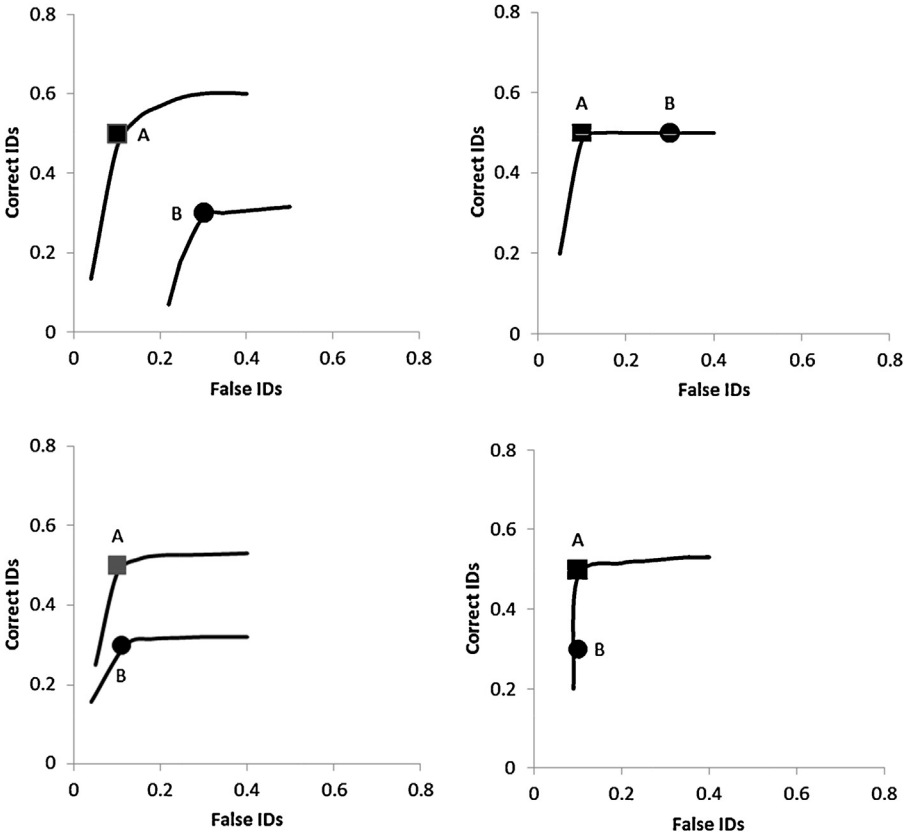

Fig. 1. Possible ROC curves through the sample data points in Table 1. The top left-hand panel depicts ROC curves that pass through the correct and false ID rates from row 1 of Table 1; the top right-hand panel corresponds to row 2 of Table 1; the bottom row depicts two possible results for row 3 of Table 1.

assign students to the two versions of the exam, it would be unfair to assign grades (which reflect course knowledge) based on the number of questions answered correctly because those students taking Exam B would be more cautious when responding, withholding some responses due to the high cost of a wrong answer. This results in fewer correct answers because these students would not risk making an error. The difference in payoffs between the two exams, however, affects only the students' willingness to respond (response bias), not their course knowledge (i.e., discriminability, the ability to distinguish correct answers from foils). Note also the corresponding role that confidence plays in the answers that are proffered. Exam B students will only answer those questions for which they are highly confident whereas exam A students will be highly confident in some answers but will answer other questions despite being less than certain.

The confounding of discriminability and response bias arises from the occurrence of 'success by chance,' coupled with the fact that a participant sets a subjective criterion for what degree of match is sufficient to warrant endorsing an item as 'old' (previously studied). For example, a student with a very liberal criterion might correctly endorse 90% of all previously studied items as 'old' (a hit). But different conclusions are warranted if that same student endorses 90% of unstudied items as 'old' versus endorsing only 30% of unstudied items as 'old' (false alarms). In recognition memory, the need to disentangle discriminability from response bias has long been known (e.g., Banks, 1970; Egan, 1958).

The primary solution to this problem in the recognition memory literature involves the application of signal detection theory (e.g., Macmillan & Creelman, 2005). Signal detection theory provides a means of separately estimating, from a hit (correct ID) and false alarm (akin to a false ID) rate, an index of discriminability (d') and an index of response bias (i.e., a willingness to make a response, e.g., β). Signal detection analyses have been applied to eyewitness data in a couple of instances. Meissner, Tredoux, Parker, and MacLin (2005) computed non-parametric signal detection quantities, counting any choice from a target-absent lineup as a false alarm.[1] Palmer and Brewer (2012) utilized a compound signal detection model (Duncan, 2006), fitting the model to a set of simultaneous and sequential lineup data and finding that sequential lineup presentation resulted in a more conservative response bias but no discriminability advantage. Clark (2012) used d' meta-analytically. But computing d' (and related measures) relies on underlying assumptions (e.g., normal evidence distributions), which usually are not met in an eyewitness experiment. In order to avoid violating assumptions associated with d', researchers have utilized probative values to support reasoning vis-à-vis which condition is superior, and logistic regression to make statistical assessments of the difference between conditions.

3. Reasoning about probative values

There are several probative value measures based on the ratio of correct (C) to false (F) ID rates (e.g., diagnosticity = C/F; conditional probability = $C/(C+F)$). If C/F for condition A is .6/.2, it is interpreted to mean that an ID of a guilty suspect is three times more likely than an ID of an innocent suspect. But recently, Wixted and Mickes (2012, 2014) (see also Clark, Erickson, & Breneman, 2011) showed that ratio-based measures of probative value are

[1] Many experiments designate an innocent suspect in a target-absent lineup and count only the choice of that innocent suspect as a false alarm (a false ID). Alternatively, one can assume a fair lineup and divide the number of false alarms to any individual in a lineup by the number of individuals in the lineup.

misleading measures of lineup performance because they confound discriminability with response bias. However, constructing ROC curves for each condition can separate the contributions of discriminability from response bias (see Gronlund et al., 2014). ROC curves are a stalwart of the signal detection approach, are assumption-free, and have facilitated the analysis of diagnostic systems in many different domains (Swets, Dawes, & Monahan, 2000). To understand the problem with logistic regression applied to eyewitness data, we need to describe ROC analysis in greater detail.

An ROC curve plots correct IDs versus false IDs at various levels of response bias. In an eyewitness experiment, these different levels of bias reflect different levels of confidence expressed by different eyewitnesses. For example, one eyewitness might select the guilty suspect and report 90% confidence in that decision, while another eyewitness might select the innocent suspect and report 60% confidence. The far left point on an ROC curve reflects the proportion of correct IDs and the proportion of false IDs reported with the highest level of confidence (see Fig. 1 for examples). The next point to the right on an ROC curve reflects the proportion of correct and false IDs reported with the highest and next highest levels of confidence, and so on. The far right hand point of an ROC curve reflects the proportion of correct and false IDs reported with any level of confidence; these are the values researchers use to compute the probative value for a particular condition.

The condition that elicits superior discriminability is the one with the ROC curve furthest from the chance diagonal. The statistical evaluation of ROC curves involves comparing the areas under the respective curves, with the condition exhibiting superior discriminability reflected by the largest area under the ROC curve. Lineup ROCs are constructed using only suspect IDs (perpetrator and innocent suspect); foil IDs are excluded, just as they are from probative value calculations, because they involve the identification of known innocents. However, the exclusion of foil IDs means that the resulting lineup ROCs are truncated, because as the response bias becomes more liberal the increased likelihood of choosing results in more foil and suspect choosing, not just more suspect choosing. Consequently, a partial area under the lineup ROCs (pAUC) must be computed. That is, rather than computing the area under an ROC curve as the false ID ranges from 0 to 1, researchers compute the pAUC by restricting the range of the false IDs (see Gronlund et al., 2014, for a tutorial).

Now that we have explained how ROC analysis can disentangle discriminability from response bias, we turn to an examination of logistic regression applied to eyewitness lineup data. We will argue that ROC analysis is necessary and sufficient for determining differences between conditions. ROC analysis answers the forensically relevant question researchers want answered: Which condition is better? Logistic regression answers that question in only restricted circumstances, and in other circumstances it can provide a misleading answer.

4. Logistic regression

Logistic regression is used when a researcher seeks to predict a categorical outcome (e.g., a correct ID or not) as a function of one or more predictor variables (e.g., weapon presence or absence, simultaneous or sequential lineup). For example, one might conclude that the odds of a correct ID are significantly greater when a weapon is absent than when a weapon is present. But logistic regression does not directly reveal to researchers what they want to know, which is whether discriminability in one condition (e.g., when a weapon is absent) is better than in another condition (e.g., when a weapon is present). To make that determination one must simultaneously consider the correct and false ID rates. But in logistic regression, correct and false identifications are analyzed separately.

Table 1
Sample data for three possible outcomes of a logistic regression.

Possible outcomes	Correct IDs	False IDs	Diagnosticity
1	A > B (.5 > .3)	B > A (.3 > .1)	A: 5.0 B: 1.0
2	A = B (.5 = .5)	B > A (.3 > .1)	A: 5.0 B: 1.7
3	A > B (.5 > .3)	B = A (.1 = .1)	A: 5.0 B: 3.0

Note: Diagnosticity is defined as correct ID rate/false ID rate.

Consider the three examples illustrated in Table 1. In the table, A > B denotes that the results of a logistic regression reveal that the correct ID rate for condition A is significantly greater than the correct ID rate for condition B. Beside that are given possible data values that correspond to that pattern. These values represent the correct and false ID rates achieved in these conditions, collapsed over all levels of response confidence. Row 1 depicts that condition A is better on both counts: condition A correct IDs are greater than condition B correct IDs and condition A false IDs are less than condition B false IDs. But in rows 2 and 3, a significant difference in correct or false IDs is paired with a nonsignificant difference in the converse response. Nevertheless, some researchers might conclude that condition A results in significantly better performance in all three of these situations because condition A always has an equivalent or greater number of correct IDs than condition B, but never has more false IDs. A probative value measure like diagnosticity (far right column) also supports the superiority of condition A. However, this conclusion is not always warranted, as the following simulated ROC data reveal.

Fig. 1 traces possible ROC curves through the sample data points in Table 1. The top left-hand panel in Fig. 1 depicts possible ROC curves that pass through the sample correct and false ID rates from row 1 of Table 1. This graph depicts ROC curves for which performance in condition A is superior to the performance in condition B. The top right-hand panel in Fig. 1 corresponds to row 2 of Table 1. It depicts a single ROC curve passing through the correct and false ID rates for both conditions; this indicates equivalent discriminability between these two conditions. The bottom row of Fig. 1 depicts two possible results for row 3 of Table 1. The left-hand graph, perhaps the more likely outcome, depicts superior performance for condition A. But the right-hand graph, like the top right-hand panel in Fig. 1, also depicts equivalent discriminability between conditions A and B.

It is only when the correct IDs for condition A are significantly greater and the false IDs are significantly less, that the conclusion reached by logistic regression and the conclusion reached by ROC analysis agree. The other two cases potentially signal differences in response bias between conditions A and B, not discriminability differences.

But the interpretive problems for logistic regression grow if we assume that the correct and false ID rates reflected in the simulated ROC curves do not reflect the maximum possible number of suspect IDs. For example, this might happen if participants viewing a sequential lineup are unwilling to respond at low levels of confidence due to the perceived difficulty of the task, reject a very good-matching foil early in the lineup sequence, or uncertainties regarding how many more lineup members are to follow. But given the proper reassurance, some participants would have produced additional correct and false IDs at lower levels of confidence. Fig. 2 shows the same data from row 2 of Table 1. However in this case, because more suspect IDs are possible in condition B beyond the values reported in Table 1, the ROC curve that passes through condition B asymptotes at a higher level than the ROC curve for condition A. Furthermore, if a sufficiently high value (β_2, as opposed to β_1) is selected to compute pAUC, ROC analysis might reveal that condition B actually results in greater discriminability than condition

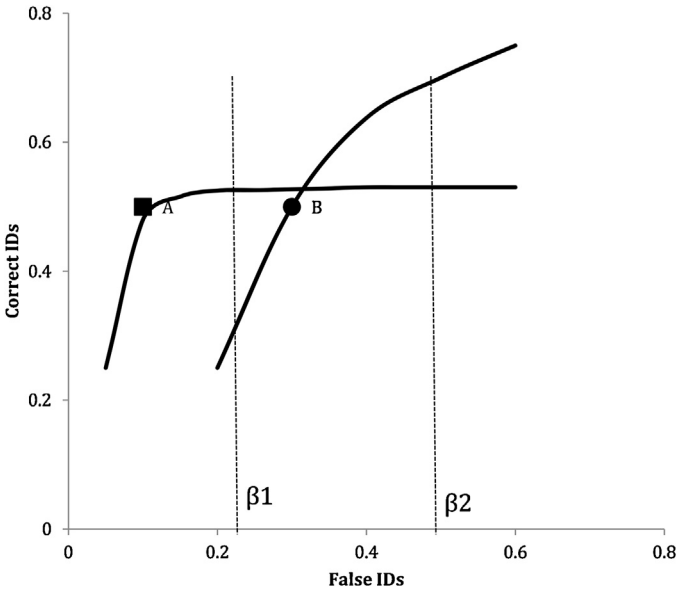

Fig. 2. Possible ROC curves through the sample data points in row 2, Table 1. The ROC curve that passes through condition B asymptotes at a higher level than the ROC curve for condition A. If pAUC is computed using β_2 rather than β_1, discriminability for condition B may be greater.

A. Something similar can happen even given the data in row 1 of Table 1, although such an outcome is unlikely.

In sum, if the goal is to make a determination of which condition results in the best performance (results in the best discriminability), utilizing logistic regression to separately analyze correct and false ID rates collapsed over all levels of confidence is open to alternative interpretations, interpretations that are resolved by ROC analysis. Does logistic regression have any role to play in the analysis of eyewitness identification data?

Logistic regression can be used in conjunction with ROC analysis (e.g., Andersen, Carlson, Carlson, & Gronlund, 2014). Once ROC analysis reveals a discriminability difference, logistic regression can determine whether the discriminability difference is due to a change in correct IDs or a change in false IDs (or possibly both). Carlson and Carlson (in this volume) used ROC analysis to show that the presence of a weapon harmed discriminability if no distinctive feature (i.e., no sticker on the perpetrator's face) was present, and found that the discriminability difference was the result of a change in false IDs and not a change in correct IDs. But as the bottom right-hand panel in Fig. 1 (row 3 in Table 1) illustrates, Carlson and Carlson could not be certain about the discriminability deficit that arises from the presence of a weapon without having first conducted an ROC analysis.

In general, caution must be exercised when interpreting comparisons of correct (or false) ID rates across conditions, because logistic regression can mislead a researcher about a purported difference in correct (or false) ID rates. For example, what if the correct ID rate for condition A is .6 and the correct ID rate for condition B is .4 (assume $p < .05$, according to a logistic regression)? It would be incorrect to interpret that difference without considering the respective response biases of conditions A and B. If condition A induces more liberal responding, the greater correct ID rate in condition A likely reflects an increased willingness to make a selection from the lineup and is not indicative of superior discriminability. From the perspective of logistic regression, the significant correct ID difference remains, but reaching the correct conclusion must incorporate concurrent consideration of the response biases in the respective conditions.

5. Conclusions

To date, most researchers have tried to reach conclusions about discriminability using measures (ratio-based probative value) or analytic techniques (logistic regression) that are confounded by differences in response bias across conditions. The best approach for reaching the proper conclusion regarding which condition results in the best performance is to disentangle discriminability from response bias by utilizing ROC analysis. Although ROC analysis can be costly and time-consuming due to the large number of observations needed to construct ROC curves for each condition, online data collection can ease its use. Moreover, it is worth the effort, both because the issues being investigated have public policy implications, but also because logistic regression (or probative value) can direct us towards the wrong conclusions, and the wrong policy.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Andersen, S. M., Carlson, C. A., Carlson, M., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences*, *60*, 36–40.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *14*, 81–99.

Carlson, C. A., & Carlson, M. A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis. *Journal of Applied Research in Memory and Cognition* (in this volume).

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

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior*, *35*, 364–380.

Duncan, M. J. (2006). *A signal detection model of compound decision tasks (Technical Report No. TR2006-256)*. Toronto, ON: Defence Research and Development Canada.

Egan, J. P. (1958). *Recognition memory and the operating characteristic (Tech. Note AFCRC-TN-58-51)*. Bloomington: Indiana University, Hearing and Communication Laboratory.

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

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

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

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

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

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

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