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/

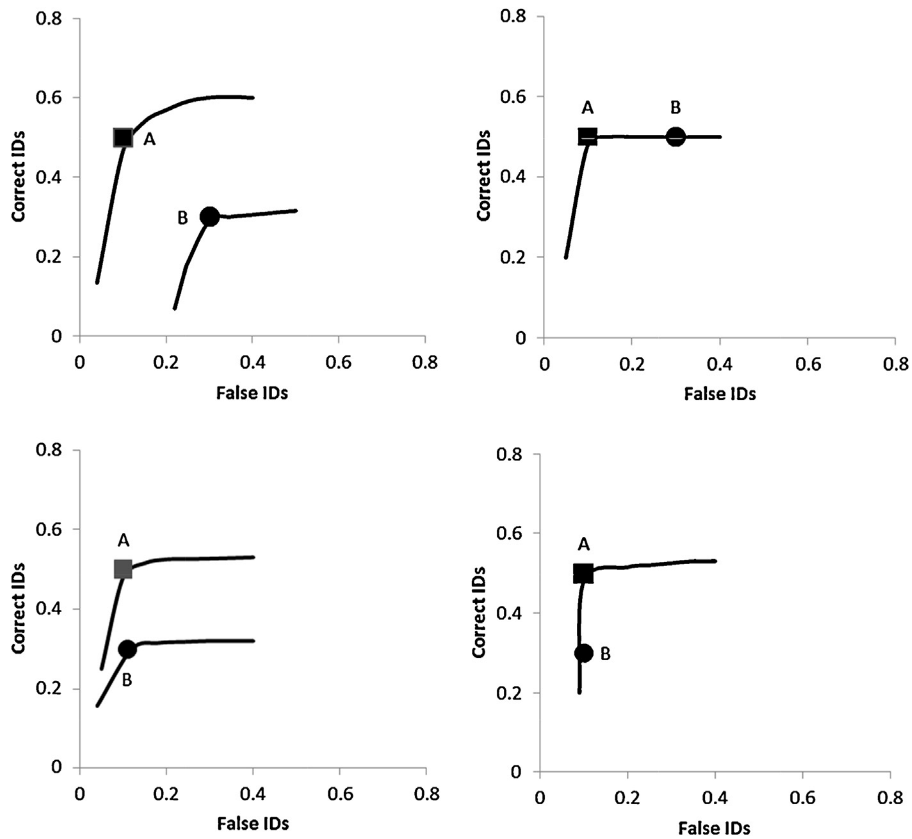

Fig. 1. Possible ROC curves through the sample data points in Table 1. The top left-hand panel depicts ROC curves that pass through the correct and false ID rates from row 1 of Table 1; the top right-hand panel corresponds to row 2 of Table 1; the bottom row depicts two possible results for row 3 of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