Upper and Lower Bounds of Area Under ROC Curves and Index of Discriminability of Classifier Performance

Shane T. Mueller STMUELLE@INDIANA.EDU
Indiana University, Department of Psychological and Brain Sciences, 1101 East 10th Street, Bloomington, IN 47404

Jun Zhang JUNZ@UMICH.EDU
University of Michigan, Department of Psychology, 530 Church Street Ann Arbor, MI 48109-1043

Abstract

Area under an ROC curve plays an important role in estimating discrimination performance – a well-known theorem by Green (1964) states that ROC area equals the percentage of correct in two-alternative forced-choice setting. When only single data point is available, the upper and lower bound of discrimination performance can be constructed based on the maximum and minimum area of legitimate ROC curves constrained to pass through that data point. This position paper, after reviewing a property of ROC curves parameterized by the likelihood-ratio, presents our recently derived formula of estimating such bounds (Zhang & Mueller, 2005).

1. Introduction

Signal detection theory (Green & Swets, 1966) is commonly used to interpret data from tasks in which stimuli (e.g., tones, medical images, emails) are presented to an operator (experimenter, medical examiner, classification algorithm), who must determine which one of two categories (high or low, malignant or benign, junk or real) the stimulus belongs in. These tasks yield a pair of measures of behavioral performance: the Hit Rate (H), also called "true positive" rate, and the False Alarm Rate (F), also called "false positive" rate. (The other two rates, those of Miss or "false negative" and of Correct Rejection or "true negative", are simply one minus H and F, respectively.) H and F are typically transformed into indices of sensitivity and bias based on assumptions about an underlying statistical model. A curve $c \mapsto (F(c), H(c))$ in the ROC (Receiver-Operating Characteristic) space is a collection of hit and false-alarm rates while the operator/receiver modifies the cutoff criterion c of accepting the input stimulus as belonging to one category versus another; often c is the likelihood ratio of the evidence favoring the two corresponding hypotheses, or a monotonic transformation thereof. In the machine learning context, we map the "operator/receiver" in the SDT sense to a "classification algorithm" or simply an "algorithm", the "stimulus" as an "input instance" or simply "instance" which carries one of the two class labels, and view c as a parameter of the algorithm which biases the output of the algorithm to favor one category or the other; the optimal setting of c is related to the cost structure, i.e., individual payoffs related to correct and incorrect classifications.

A well-known result in SDT is Green's Theorem, which relates the discrimination accuracy performance of an operator to the area under the operator's (i.e., the classification algorithm's) ROC curve. This so-called ROC area is thus a compact measure of how discriminable a classification algorithm is between binary-class inputs. Consequently, the performance of different algorithms can be compared by comparing their respective ROC areas.

Often, algorithms reported in the literature may not contain a tradeoff analysis of the Hit and False Alarm rates produced by varying parameters corresponding to the algorithm's bias. In these cases, the entire ROC curve of an algorithm may not be available — in some cases, only a few or even a single point (called "data point") in the ROC space is available. In this case, performance comparison across different algorithms becomes a question of comparing areas of possible ROC curves constrained to pass through these limited data

Appearing in *Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning*, Pittsburgh, USA, 2006.

points.

In the mathematical psychology community, the problem of estimating area of ROC curves constrained to pass through a single data point is particularly well studied (Norman, 1964; Pollack & Norman, 1964; Pollack & Hsieh, 1969; Grier, 1971; Smith, 1995; Zhang & Mueller, 2005). These estimates of the ROC area do not assume the ROC curves to arise from any specific class of parametric models, and so these estimates are often referred to as a "non-parametric" indices of an operator's discriminability (sensitivity).[1] Typically, the upper and lower bounds of discriminability were obtained by considering the maximal and minimum ROC areas among the class of "admissible" ROC curves satisfying the data constraint. Interestingly, though the basic idea was very simple and advanced over 40 years ago (Pollack & Norman, 1964), the popular formula to calculate this index (Grier, 1971), dubbed A' in psychometrics and cognitive psychology literature, turned out to be erroneous, at least insofar as its commonly understood meaning is concerned; moreover, its purported correction (Smith, 1995), dubbed A'', also contained an error. These formulae incorrectly calculated the upper bound of admissible ROC curves, using either an ROC curve that was not admissible (Pollack & Norman, 1964), or one that was not the maximum for some points (Smith, 1995). Zhang and Mueller (2005) rectified the error and gave the definite answer to the question of non-parametric index of discriminability based on ROC areas.

In this note, we first review the notion of "proper" (or "admissible") ROC curves and prove a lemma basically stating that all ROC curves are proper/admissible when the likelihood functions (for the two hypotheses) used to construct the ROC curve are parameterized by the likelihood ratio (of those hypotheses). We then review Green's Theorem, which related area under an ROC curve to percentage correct in a two-alternative discrimination task. Finally, we present the upper and lower bounds on a 1-point constrained ROC area and reproduce some of the basic arguments underlying their derivation. All technical contents were taken from Zhang and Mueller (2005).

[1]Though no parametric assumption is invoked in the derivation of these indices, the solution itself may correspond to certain models of underlying likelihood process, see MacMillan and Creelman, 1996. In other words, parameter-free here does not imply model-free.

2. Slope of ROC curve and likelihood ratio

Recall that, in the traditional signal detection framework, an ROC curve $u_c \mapsto (F(u_c), H(u_c))$ is parameterized by the cutoff criteria value u_c along the measurement (evidence) axis based on which categorization decision is made. Given underlying signal distribution $f_s(u)$ and noise distribution $f_n(u)$ of measurement value u[2], a criterion-based decision rule, which dictates a "Yes" decision if $u > u_c$ and a "No" decision if $u < u_c$, will give rise to

$$
\begin{aligned}
H(u_c) &= \Pr(\text{Yes}|s) = \Pr(u > u_c|s) = \int_{u_c}^{\infty} f_s(u)du, \\
F(u_c) &= \Pr(\text{No}|s) = \Pr(u > u_c|n) = \int_{u_c}^{\infty} f_n(u)du.
\end{aligned}
\tag{1}
$$

As u_c varies, so do H and F; they trace out the ROC curve. Its slope is

$$
\left.\frac{dH}{dF}\right|_{F=F(u_c),H=H(u_c)} = \frac{H'(u_c)}{F'(u_c)} = \frac{f_s(u_c)}{f_n(u_c)} \equiv l(u_c)\ .
$$

With an abuse of notation, we simply write

$$
\frac{dH(u)}{dF(u)} = l(u)\,. \tag{2}
$$

Note that in the basic setup, the likelihood ratio $l(u)$ as a function of decision criterion u (whose optimal setting depends on the prior odds and the payoff structure) need not be monotonic. Hence, the ROC curve $u \mapsto (F(u), H(u))$ need not be concave. We now introduce the notion of "proper (or admissible) ROC curves".

Definition 2.1. A *proper* (or *admissible*) ROC curve is a piece-wise continuous curve defined on the unit square $[0, 1] \times [0, 1]$ connecting the end points (0,0) and (1,1) with non-increasing slope.

The shape of a proper ROC curve is necessarily concave (downward-bending) connecting (0,0) and (1,1). It necessarily lies above the line $H = F$. Next we provide a sufficient and necessary condition for an ROC curve to be proper/admissible, that is, a concave function bending downward.

Lemma 2.2. An ROC curve is proper if and only if the likelihood ratio $l(u)$ is a non-decreasing function of decision criterion u.

[2]In machine learning applications, "signal" and "noise" simply refer the two category classes of inputs, and "signal distribution" and "noise distribution" are likelihood functions of the two classes.

Proof. Differentiate both sides of (2) with respect to u

$$\frac{dF}{du} \cdot \frac{d}{dF}\left(\frac{dH}{dF}\right) = \frac{dl}{du}.$$

Since, according to (1)

$$\frac{dF}{du} = -f_n(u) < 0,$$

therefore

$$\frac{dl}{du} \geq 0 \Longleftrightarrow \frac{d}{dF}\left(\frac{dH}{dF}\right) \leq 0$$

indicating that the slope of ROC curve is non-increasing, i.e., the ROC curve is proper. $\diamond$

Now it is well known (see Green & Swets, 1966) that a monotone transformation of measurement axis $u \mapsto v = g(u)$ does not change the shape of the ROC curve (since it is just a re-parameterization of the curve), so a proper ROC curve will remain proper after any monotone transformation. On the other hand, when $l(u)$ is not monotonic, one wonders whether there always exists a parameterization of any ROC curve to turn it into a proper one. Proposition 1 below shows that the answer is positive — the parameterization of the two likelihood functions is to use the likelihood ratio itself!

PROPOSITION 2.3. (Slope monotonicity of ROC curves parameterized by likelihood-ratio). The slope of an ROC curve generated from a pair of likelihood functions $(F(l_c), H(l_c))$, when parameterized by the likelihood-ratio l_c as the decision criterion, equals the likelihood-ratio value at each criterion point l_c

$$\frac{dH(l_c)}{dF(l_c)} = l_c. \tag{3}$$

Proof. When likelihood-ratio l_c is used the decision cutoff criterion, the corresponding hit rate (H) and false-alarm rate (F) are

$$\begin{aligned} H(l_c) &= \int_{\{u:l(u)>l_c\}} f_s(u)du, \\ F(l_c) &= \int_{\{u:l(u)>l_c\}} f_n(u)du. \end{aligned}$$

Note that here u is to be understood as (in general) a multi-dimensional vector, and du should be understood accordingly. Writing out $H(l_c + \delta l) - H(l_c) \equiv \delta H(l_c)$ explicitly,

$$\delta H(l_c) = \int_{\{u:l(u)>l_c+\delta l\}} f_s(u)du - \int_{\{u:l(u)>l_c\}} f_s(u)du$$

$$= -\int_{\{u:l_c<l(u)<l_c+\delta l\}} f_s(u)du \simeq -\int_{\{u:l(u)=l_c\}} f_s(u)\,\delta u$$

where the last integral $\int \delta u$ is carried out on the set $\partial \equiv \{u : l(u) = l_c\}$, i.e., across all u's that satisfy $l(u) = l_c$ with given l_c. Similarly,

$$\delta F(l_c) \simeq -\int_{\{u:l(u)=l_c\}} f_n(u)\,\delta u\,.$$

Now, for all $u \in \partial$

$$\frac{f_s(u)}{f_n(u)} = l(u) = l_c$$

is constant, from an elementary theorem on ratios, which says that if $a_i/b_i = c$ for $i \in I$ (where c is a constant and I is an index set), then $(\sum_{i\in I} a_i)/(\sum_{i\in I} b_i) = c$,

$$\frac{\delta H(l_c)}{\delta F(l_c)} = \frac{\int_\partial f_s(u)\,\delta u}{\int_\partial f_n(u)\,\delta u} = \left.\frac{f_s(u)\,\delta u}{f_n(u)\,\delta u}\right|_{u\in\partial} = l_c\,.$$

Taking the limit $\delta l \to 0$ yields (3). $\diamond$.

Proposition 2.3 shows that the slope of ROC curve is always equal the likelihood-ratio value regardless how it is parameterized, i.e., whether the likelihood-ratio is monotonically or non-monotonically related to the evidence u and whether u is uni- or multi-dimensional. The ROC curve is a signature of a criterion-based decision rule, as captured succinctly by the expression

$$\frac{dH(l)}{dF(l)} = l\,.$$

Since $H(l)$ and $F(l)$ give the proportion of hits and false alarms when a decision-maker says "Yes" whenever the likelihood-ratio (of the data) exceeds l, then $\delta H = H(l + \delta l) - H(l)$, $\delta F = F(l + \delta l) - F(l)$ are the amount of hits and false-alarms if he says "Yes" only when the likelihood-ratio falls within the interval $(l, l+\delta l)$. Their ratio is of course simply the likelihood-ratio.

Under the likelihood-ratio parameterization, the signal distribution $f_s(l) = -dH/dl$ and the noise distribution $f_n(l) = -dF/dl$ can be shown to satisfy

$$E_s\{l\} = \int_{l=0}^{l=\infty} l f_s(l)dl \geq 1 = \int_{l=0}^{l=\infty} l f_n(l)dl = E_n\{l\}.$$

The shape of the ROC curve is determined by $H(l)$ or $F(l)$. In fact, its curvature is

$$\kappa = \frac{d}{dl}\left(\frac{dH}{dF}\right) / \left(1 + \left(\frac{dH}{dF}\right)^2\right) = \frac{1}{1+l^2}\,.$$

3. Green's Theorem and area under ROC curves

The above sections studies the likelihood-ratio classifier in a single-instance paradigm — upon receiving an input instance, the likelihood functions in favor of each hypothesis are evaluated and compared with a pre-set criterion to yield a decision of class label. Both prior odds and payoff structure can affect the optimal setting of likelihood ratio criterion l_c by which class label is assigned. On the other hand, in two-alternative force choice paradigms with two two instances, each instance is drawn from one category, and the operator must match them to their proper categories. For example, an auditory signal may be present in one of two temporal intervals, and the operator must determine which interval contains the signal and which contains noise. In this case, the likelihood-ratio classifier, after computing the likelihood-ratios for each of the instances, simply compares the two likelihood-ratio values l_a and l_b, and matches them to the two class labels based on whether $l_a < l_b$ or $l_a > l_b$. It turns out that the performance of the likelihood-ratio classifier under the single-instance paradigm ("detection paradigm") and under the two-instance forced-choice paradigm ("identification paradigm") are related by a theorem first proven by Green (1964).

Proposition 3.1. (Green, 1964). Under the likelihood-ratio classifier, the area under an ROC curve in a single-observation classification paradigm is equal to the overall probability correct in the two-alternative force choice paradigm.

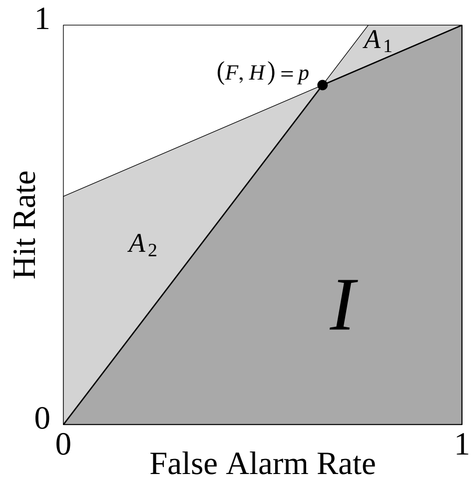
Proof. Following the decision rule of the likelihood-ratio classifier, the percentage of correctly ("PC") matching the two input instances to the two categories is

$$\begin{aligned} \text{PC} &= \int\int_{0\le l_b \le l_a \le \infty} f_s(l_a)\, f_n(l_b)\, dl_a\, dl_b \\ &= \int_0^{\infty} \left(\int_{l_b}^{\infty} f_s(l_a)\, dl_a \right) f_n(l_b)\, dl_b \\ &= \int_{l_b=0}^{l_b=\infty} H(l_b)\, dF(l_b) = \int_{F=0}^{F=1} H dF, \end{aligned}$$

which is the area under the ROC curve $l_c \mapsto (F(l_c), H(l_c))$. ⋄

Green's Theorem (Proposition 3.1) motivates one to use the area under an ROC curve to as a measure of discriminability performance of the operator. When multiple pairs of hit and false alarm rates $(F_i, H_i)_{i=1,2,\cdots}$ (with $F_1 < F_2 < \cdots, H_1 < H_2 < \cdots$) are available, all from the same operator but under manipulation of prior odds and/or payoff structure and

Figure 1. Proper ROC curves through point p must lie within or on the boundaries of the light shaded regions A_1 and A_2. The minimum-area proper ROC curve through p lies on the boundary of region I.



with the constraints

$$0 \le \cdots \le \frac{H_3 - H_2}{F_3 - F_2} \le \frac{H_2 - H_1}{F_2 - F_1} \le \infty,$$

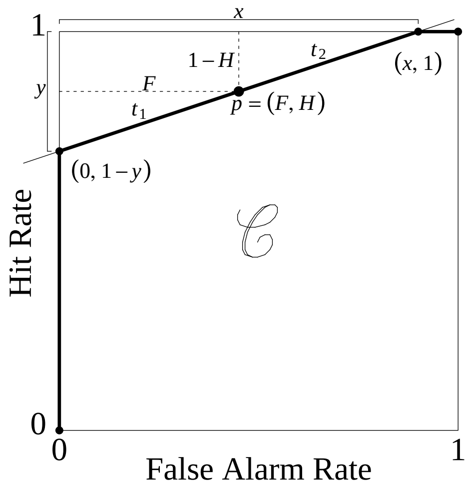
then it is possible to construct proper ROC curves passing through these points, and the bounds for their area can be constructed. The question of finding the areal bounds of ROC curves passing through a single data point has received special attention in the past (since Norman, 1964), because as more data points are added, the uncertain in ROC area (difference between the upper and lower bounds of area measure) decreases. We discuss the bounds of 1-point constrained ROC area in the next sections.

4. ROC curves constrained to pass through a data point

When the data point $p = (F, H)$ is fixed, the non-increasing property of the slope (Corollary 1) immediately leads to the conclusion that all proper ROC curves must fall within or on the bounds of light shaded regions A_1 and A_2 (shown in Figure 1). This observation was first made in Norman (1964). The proper ROC curve with the smallest area lies on the boundary between I and A_1 (to the right of p) and A_2 (to the left of p), whereas the proper ROC curve with the largest area lies within or on the boundaries of A_1 and A_2.

Pollack and Norman (1964) proposed to use the average of the areas $A_1 + I$ and $A_2 + I$ as an index of discriminability (so-called A'), which turns out to equal

Figure 2. Example of a proper ROC curve through p. The ROC curve $\mathcal{C}$, a piecewise linear curve denoted by the dark outline, is formed by following a path from $(0,0)$ to $(0,1-y)$ to $(x,1)$ (along a straight line that passes through $p = (F,H)$) and on to $(1,1)$.



$1/2 + (H-F)(1+H-F)/(4H(1-F))$ (Grier, 1971). However, the A' index was later mistakenly believed to represent the *average* of the maximal and minimum areas of proper ROC curves constrained to pass through $p = (F,H)$. Rewriting

$$\frac{1}{2}((A_1 + I) + (A_2 + I)) = \frac{1}{2}(I + (A_1 + A_2 + I)),$$

the mis-conceptualization probably arose from (incorrectly) taking the area $A_1 + A_2 + I$ to be the maximal area of 1-point constrained proper ROC curves while (correcting) taking the are I to be the minimal area of such ROC curves, see Figure 1. It was Smith (1995) who first pointed out this long, but mistakenly-held belief, and proceeded to derive the true upper bound (maximal area) of proper ROC curves, to be denoted A_+. Smith claimed that, depending on whether p is to the left or right of the negative diagonal $H + F = 1$, A_+ is the larger of $I + A_1$ and $I + A_2$. This conclusion, unfortunately, is still erroneous when p is in the upper left quadrant of ROC space (i.e., $F < .5$ and $H > .5$) — in this region, neither $I + A_1$ nor $I + A_2$ represents the upper bound of all proper ROC curves passing through p.

5. Lower and upper bound of area of 1-point constrained proper ROC curves

The lower bound A_- of the area of all proper ROC curves constrained to pass through a given point $p = (F,H)$ can be derived easily (the area labelled as I in Figure 1):

$$A_- = \frac{1}{2}(1 + H - F).$$

In Zhang and Mueller (2005), the expression was derived for the upper bound A_+ of such ROC area.

Proposition 5.1. (Upper Bound of ROC Area). The areal upper bound A_+ of proper ROC curves constrained to pass through one data point $p = (F,H)$ is

$$A_+ = \begin{cases} 1 - 2H(1-F) & \text{if} \quad F < 0.5 < H\,, \\ \frac{1-F}{2H} & \text{if} \quad F < H < 0.5\,, \\ 1 - \frac{1-H}{2(1-F)} & \text{if} \quad 0.5 < F < H\,. \end{cases}$$

Proof. See Zhang and Mueller (2005). $\diamond$

The ROC curve achieving the maximal area generally consists of three segments (as depicted in Figure 2), with the data point p *bisecting* the middle segment – in other words, $t_1 = t_2$ in Figure 2. When p falls in the $F < H < 0.5$ ($0.5 < F < H$, resp) region, then the vertical (horizontal, resp) segment of the maximal-area ROC curve degenerates to the end point $(0,0)$ ($(1,1)$, resp), corresponding to $y = 1$ ($x = 1$, resp) in Figure 2.

With the upper and lower bounds on ROC area derived, Figure 3 plots the difference between these bounds — that is, the uncertainty in the area of proper ROC curves that can pass through each point. The figure shows that the smallest differences occur along the positive and negative diagonals of ROC space, especially for points close to $(0,1)$ and $(.5,.5)$. The points where there is the greatest difference between the lower and upper bounds of ROC area are near the lines $H = 0$ and $F = 1$. Thus, data observed near these edges of ROC space can be passed through by proper ROC curves with a large variability of underlying areas. Consequently, care should be taken when trying to infer the ROC curve of the observer/algorithm when the only known data point regarding its performance (under a single parameter setting) falls within this region.

By averaging the upper and lower bound $A = (A_+ + A_-)/2$, we can derive the (non-parametric) index of discriminability performance

$$A = \begin{cases} \frac{3}{4} + \frac{H-F}{4} - F(1-H) & \text{if} \quad F \le 0.5 \le H\,; \\ \frac{3}{4} + \frac{H-F}{4} - \frac{F}{4H} & \text{if} \quad F < H < 0.5\,; \\ \frac{3}{4} + \frac{H-F}{4} - \frac{1-H}{4(1-F)} & \text{if} \quad 0.5 < F < H\,. \end{cases}$$

One way to examine A is to plot the "iso-discriminability" curve, i.e, the combinations of F and

Figure 3. Difference between the lower and upper bounds of area of proper ROC curves through every point in ROC space. Lighter regions indicate smaller differences.

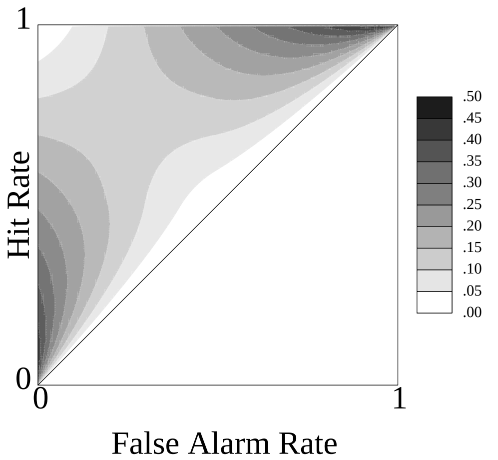
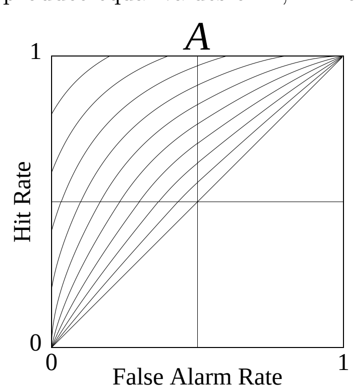


Figure 4. Iso-discriminability contours in ROC space. Each line corresponds to combinations of F and H that produce equal values of A, in increments of 0.05.



H will produce a given value of A. The topography of A in ROC space can be mapped by drawing isopleths for its different constant values. Figure 4 shows these topographic maps for A.

Finally, since the slope of any proper ROC curve is related to the likelihood ratio of the underlying distributions, we can construct an index of decision bias (Zhang & Mueller, 2005), denoted b, as being orthogonal to the slope of the constant-A curve (called b):

$$b = \begin{cases} \frac{5-4H}{1+4F} & \text{if} \quad F \le 0.5 \le H\,; \\ \frac{H^2+H}{H^2+F} & \text{if} \quad F < H < 0.5\,; \\ \frac{(1-F)^2+1-H}{(1-F)^2+1-F} & \text{if} \quad 0.5 < F < H\,. \end{cases}$$

6. Conclusion

We showed that the relationship of ROC slope to likelihood-ratio is a fundamental relation in ROC analysis, as it is invariant with respect to any continuous reparameterization of the stimulus, including non-monotonic mapping of uni-dimensional and multi-dimensional evidence in general. We provided an upper bound for the area of proper ROC curves passing through a data point and, together with the known lower bound, a non-parametric estimate of discriminability as defined by the average of maximal and minimum ROC areas.

References

Green, D. M. (1964). General prediction relating yes-no and forced-choice results. *Journal of the Acoustical Society of America, A*, *36*, 1024.

Green, D. M., & Swets, J. A. (1964). *Signal detection theory and psychophysics*. New York: John Wiley & Sons.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: computing formulas. *Psychological Bulletin*, *75*, 424–429.

Macmillan, N. A., & Creelman, C. D. (1996). Triangles in roc space: History and theory of "non-parametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review*, *3*, 164–170.

Norman, D. A. (1964). A comparison of data obtained with different false-alarm rates. *Psychological Review*, *71*, 243–246.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory* (pp. 171–212).

Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under roc curve and d'_e. *Psychological Bulletin*, *1*, 161–173.

Pollack, I., & Norman, D. A. (1964). Non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*, 125–126.

Smith, W. D. (1995). Clarification of sensitivity measure A'. *Journal of Mathematical Psychology*, *39*, 82–89.

Zhang, J., & Mueller, S. T. (2005). A note on roc analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*, 145–154.