Signal Detection Theory

Peter R Killeen

Arizona State University Tempe

Signal Detection Theory

Peter R. Killeen

Arizona State University

Killeen, P. R. (2014). Signal Detection Theory. In H. L. Miller (Ed.), *Encyclopedia of Theory in Psychology*. Thousand Oaks, CA: Sage.

Killeen@asu.edu

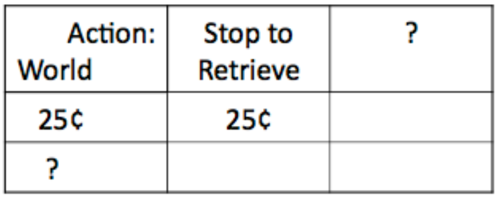
Signal Detection Theory

All decisions in life involve at least two factors: What we know and what we want. When both of these are clear, decisions almost make themselves; when both are ambiguous, decisions are wisely delayed to allow time for data collection and value clarification. But often we find ourselves in the middle ground, of useful but imperfect information about the state of the world, and good, if imperfect, understanding of our desires. How then do we decide? This is the subject of decision theory, of which signal detection theory (SDT) is a special case. Decision theory is normative: it tells us what to do. Signal detection theory is descriptive: it attempts to interpret behavior in terms of the knowledge (signal strength) and values (bias) that most likely motivated it.

Decision Theory

Decision theory is a fundamentally important approach to the world, and to every decision that we make in it. This is true even though we are typically unable to assign the numbers necessary to formally instrument it. It reminds us of key issues about which we need to be informed. A simplistic illustration is useful. We believe we see a 25-cent piece on the ground; do we stop our walk for a closer look? The outcomes concern money, and we prefer more of that to less. The rows of Table 1 give the true state of the world, and the columns the proposed actions. If there *is* a quarter on the ground and we stop to retrieve it, the value of the payoff to us is 25¢. What do we place as the header of the right column? “Move on”? Or “Slow down for a better look”? How we frame a problem will often foreordain its answer.

Table 1:



Action: World	Stop to Retrieve	?
25¢	25¢	
?		

What goes in the second row header? 0¢? But this question could not have arisen if there were 0¢--nothing--on the ground. A more likely alternative is bottle cap, or a piece of foil from a cigarette wrapper, or a nickel. A quick review of these possibilities is accomplished almost without breaking stride. By providing likely alternative stimuli for which you have idealized visual templates, you effectively improve the signal strength of the stimulus. Perhaps a better set of row headers is “Something of value, possibly a coin” vs. “Nothing of value”.

The decision theory framework gives us many lessons of value in life. It reminds us that there are always (at least) 4 cells to the table: We should always consider just what the alternative states of the world may be, and what might happen if we take no action, or a different action. It is at home within the framework of Bayesian updating of information, meaning that we should have at least some sense of the base rates of the phenomena. How likely is it that a 25¢ piece should be on that spot of ground? How likely a cigarette wrapper? Finally, it is utility that really matters—not cost or dollar value, but value to you. A white elephant may cost thousands to purchase, yet have negative utility for you. Know what matters. Whenever the decision-theoretic framework brings these considerations to mind, the ensuing decisions will be enlightened. Once the table is laid out, the actual numbers for the cells are frosting, often not worth calculating beyond order-of-magnitude estimates.

Signal Detection Theory (SDT)

Signal detection theory (SDT) operates the decision table in reverse. It applies to situations in which we have estimates of the probability of an individual taking actions *a* or *b*, in the context of stimuli A and B. From these we may fill in the table with the probabilities, and infer how clear—how detectable—the target stimulus was, and how biased the individual was to act or to withhold action, a bias that reflects the relative utilities of the outcomes to the individual.

As with many instruments that have evolved with time, its name is something of a misnomer. SDT was developed to inform sonar and radar operators how to respond if they thought that they had detected an enemy ship. When instruments are pushed to their limits, spurious signals--due perhaps to birds, or inversion layers, or electronic malfunction--could look like the target to be detected. SDT is really *Signal Discrimination Theory*. In most cases the receiver is trying to discriminate between two signals, one of the target, and an alternative, which might be well-defined: “Look, up in the sky—is that a bird? Is it a plane? Is it Superman?” or poorly defined. In the latter case it is simply called *noise*. When inserted by an experimenter, it is called the *foil.* A very pragmatic modern use for SDT is in screening airline passengers for contraband. The inspectors are periodically tested with planted foils.

As with decision theory, the primary strength of SDT is its ability to separate knowledge from value, detectability from bias. Consider the following experiment. One thousand trials were conducted in which stimulus A was present on 400, and absent (~A) on 600. Table 2 gives the table of joint frequencies of object and action (saying *a* or *not a*) that were recorded.

Table 2:

Action:	*a*	*Not a*	
World			Total
A	300	100	400
~A	250	350	600
Total:	550	450	1000

We may compute the accuracies in terms of relative frequencies. The probability of making a correct response in the presence of A—$p(a|\text{A})$, a *Hit*—was 300/(300+100) = 0.75; when A was absent the probability of a "correct rejection", $p(\textit{not } a|\sim\text{A})$, was 350/600 = 0.583. A mistaken "a" when A was absent—$p(a|\sim\text{A})$, a *False Alarm*—had probability 250/600 = 0.417. Finally, the probability of blinking—of a *Miss, p*(*not a*|A)—is 100/400 = 0.25. It is helpful if the novice now pencils in *H, M, FA* and *CR* in the cells of Table 2. Notice that the Miss and Correct Rejection rates are the complements of the Hit and False Alarm rates, so if we know the first two probabilities, the second two add no new information, and may be ignored. In the medical literature, Hit rate and Correct Rejection rate are emphasized, where they are called the *sensitivity* and the *specificity* of a test. In the discussion below correct rejection rates are used where they are more intuitive than False Alarm rates, remembering that $p(\text{CR}) = 1 - p(\text{FA})$, just as $p(\text{H}) = 1 - p(\text{M})$.

All of these probabilities are conditionalized on the presence or absence of the target by dividing by the sums in the rightmost column. It would often be more appropriate to conditionalize on the response, by dividing by the column totals. This would be the case, for instance, if we knew the basic statistics on a particular TSA agent scanning luggage, and had to evaluate the probability of a new stimulus being the target given her positive response to it: *p*(A|*a*). But for historical reasons this is seldom done.

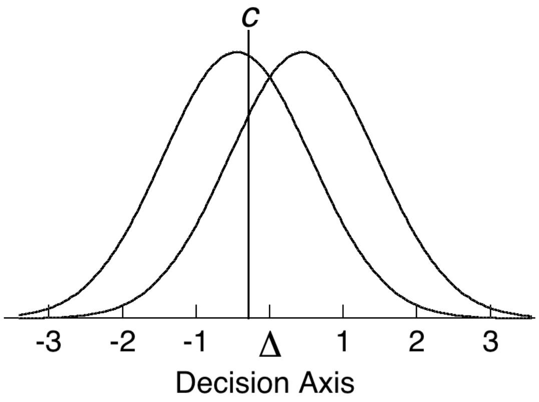


Figure 1. The underlying representation implicit in SDT, here corresponding to the data in Table 1. The distribution on the right represents signal A, and that on the left signal B, or noise (~A). Instances of the signals from one trial to the next can occur anywhere along the axis, with the likelihood of each position given by the height of the curves. An observer may set a criterion *c* indicated by the vertical line, and respond to all signals that fall to the right *a*, and those that fall to the left *not a*, or *b*. The area to the right of *c* under the right distribution constitutes Hits, (H), and to the left of it under the left distribution Correct Rejections (CR). The unbiased observer sets her criterion at 0. Then, like weights on a balance, *p*(H) = *p*(CR).

We have reduced the 4 joint frequencies of Table 2 to two estimates of conditional probabilities, but still do not have the information we want. Getting it requires some assumptions about how decisions are made. Figure 1 shows the standard icon of SDT, derived from the work of Thurstone. (A prominent psychologist said of this icon that, once learned, it is impossible to think of a discrimination problem without seeing it. It is a meme.) Thurstone argued that the same object does not always give rise to the same perception, but rather to a spread of stimulation—a coin may lay heads or tails, be clean or dirty, seen at dawn or dusk. It is this spread of possibilities that the spread of these distributions reflects. Based on the theory of errors, the many small independent factors that add to create an overall perception will tend toward a normal distribution. The distribution on the right is that of the signal, A. That on the left is the foil, ~A. They are arrayed over what is called the *decision axis*. On any single occasion, the location of the signal is at one of the points covered by the right distribution, with the likelihood of being at a particular location given by the height of the curve. The same is true for the foil on trials without a signal. On any trial the location of the target will probably fall toward the right of the side of the axis. A savvy observer will set a mark on that axis, called the *criterion*, and respond *a* whenever the sensation falls to the right of it, and respond *not a* otherwise. Then the probability of a hit is the area under the A distribution to the right of the criterion—here 0.75. But sometimes with signal absent, the foil will itself exceed the criterion. The probability of a false alarm is the area under the tail of the noise distribution to the right of the criterion—here 0.417. The novice should color in these areas, and identify the two remaining areas. Those who have take a course in statistical inference will recognize that FA correspond to Type 1 errors, and M to Type II errors.

With these two bits of information, hit rate (H) and false alarm rate (FA), we can infer how far apart the distributions must be (signal strength, or detectability), and where the criterion is located (bias, determined by the relative value we place on hits and correct rejections). But there is more information in Figure 1 than given by those two bits, and just as we reduced the "degrees of freedom" in Table 2 by converting to conditional probabilities, we must make assumptions to reduce the degrees of freedom in the model. Both distributions have means and variances; we stipulate both variances to be 1.0, and place the origin of the decision axis, 0, halfway between the means of the two distributions. Then we are down to two unknowns: the distance between the signal and noise distributions, *d'*, and the location of the criterion, called *c*. The distance equals the sum of the *z*-scores of the two correct response probabilities, H and CR: $d' = z(\text{H}) + z(\text{CR}) = 0.67 + 0.21 = 0.88$. Bias tells us how much the individual on the average favors Hits over Correct Rejections, and equals half the differences of their *z*-scores: $c = (z(\text{H}) - z(\text{CR}))/2 = -0.23$. If the observer had set her criterion where the distributions cross, those *z*-scores would be equal, and $c = 0$. These two most fundamental measures of SDT, the index of discriminability *d'* and the index of bias, *c*, are thus derived from the sum and difference of *z*-scores of the two ways of being correct.

We may think of the decision axis as a balance beam. The farther the distributions are from the fulcrum, the more clearly and forcefully they speak; and the farther the criterion is from their center of gravity at 0, the more weight that is given to negative or positive evidence. The observed negative value of *c* in our example shows that the observer is not unbiased; she is more likely to say *a* than *not a* in the presence of an ambiguous signal. This negative criterion is called a "liberal bias". Bias is not necessarily bad. If the signal were a calcium deposit that might signal cancer, the radiologist should have a liberal criterion to order additional tests. But if those additional tests were invasive and not reliable, she should move her criterion in a more conservative direction. Exactly where the criterion should be placed is a matter of costs and benefits of the four cells of Table 1. Although those are often evaluated intuitively by experts, their license is not unlimited; the values of informed clients should also play a role in criterion setting. As may be inferred from Figure 1, when the distributions are close together, small shifts in the criterion can have a large effect on decisions. When they are far apart, the same shifts in the criterion may have negligible effect. That is why, in the case of contested decisions, it is important to have as great a separation in the distributions—as great a *d'*—as possible.

One of the great advantages of SDT is that it gives an authoritative answer to the question: How can we evaluate the informative value of a test in a way that is independent of the criterion of the expert? *d'* gives one good answer, but there is a more general one. To understand that, we must study the second icon of SDT, the ROC: the Receiver (or Relative) Operating Characteristic.

The ROC

Reconsider Figure 1, with the criterion all the way to the left--an extremely liberal bias. Such a criterion yields a test that is 100% effective in detecting a target when it is present. Great, you might say—but too many tests have been accepted based on enthusiasm for such half-information. Unfortunately the test is also 100% **in**effective in saying when the target is absent—its false alarm rate is 100%. It is as though the radiologist recommended surgery for everyone who walked in the door. Now sweep the criterion all the way to the right, taking note of hit rates and false alarm rates as you go, until it is much too conservative, with no false alarms, but at the cost of no hits. It is an extremely risk averse position, with no chances taken because of the cost or fear of failure.

It is difficult to see the relationship between H and FA without graphing it. That is done in Figure 2, which shows the ROC curve that results from the above experiment. The area to the right of the criterion for the A distribution (H) is plotted as a function of the area to the right of the criterion for the ~A distribution (FA). The dashed lines shows where the criterion drawn in Figure 1 has left the tradeoff between success and failure.

From this figure it is clear at a glance how changes in criterion that affect hit rate will also affect false-alarm rate. If detectability—*d'*—is 0, the ROC falls along the diagonal: one can achieve a hit rate of 75%, at the cost of a false alarm rate of 75%; guessing at chance levels all the way along that line. If *d'* is large, say 3 or 4, the ROC will rise quickly from 0 toward the top, then smoothly curve over to the right. If the observer is perceptive but has confused the response, saying *a* when *b* is the correct and vice versa, the resulting ROC will curve below the diagonal, as though the bow just flopped down. If we use *z*(FA) and *z*(H) as the axes, the ROC becomes a straight line.

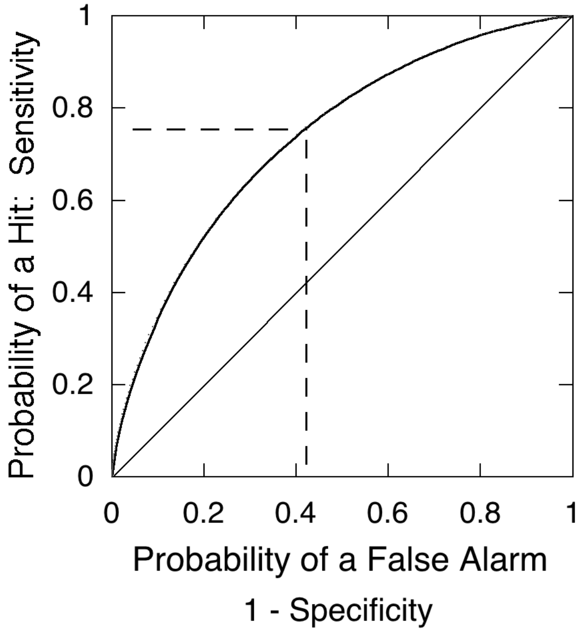


Figure 2. The Receiver Operating Characteristic (ROC) based on the data in Table 2. This figure plots the evolving values of *p*(H) as a function of the values of *p*(FA) as the criterion in Figure 1 is swept from the left of its frame to the right. The placement of *c* in Figure 1 gives the point of intersection of the two dashed lines.

The value of an ROC curve is that it manifests the costs in FA for any desired level of H: Simply draw the horizontal at H until it intersects the ROC, then drop the vertical to predict the FA. This procedure remains valid even if the underlying distributions are not the normal ones we see in Figure 1. If, for instance, the distributions are exponential decays starting at 0, with the noise distribution starting high and decaying quickly, and the signal starting lower and decaying slowly, an ROC like that in Figure 3 results. Here, and wherever the ROC is not symmetric, *d'* is not so useful an index of test quality.

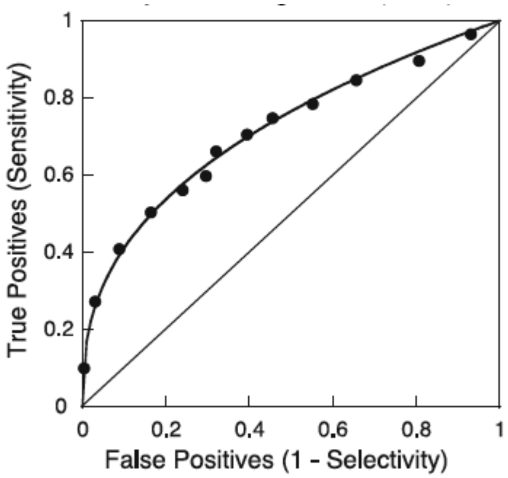


Figure 3. The probability of correctly diagnosing Attention Deficit Hyperactivity Disorder where it exists, as function of the probability of a misdiagnosis. The data are kindly provided by Anneke Meyer, from a study by Meyer and Sagvolden (2006). The test is the sum of two measures of fine motor skills, administered to 264 children from 7 ethnic groups in South Africa, and 264 controls. The ROC used the rating scale technique with 12 equally-spaced criteria on motor proficiency.

The AUC

An index that is robust over the shape of the ROC is the A*rea Under the Curve*, AUC. The total area in Figures 2 and 3 is 1. The area under the diagonal is ½. As the test becomes more accurate, the AUC increases from ½ to 1. In the case of Figure 2, the AUC for the ROC is N(*d'*/√2) = 0.73, where N(z) is the cumulative normal distribution. When the underlying distributions are exponential, the ROC is a power function such as shown in Figure 3, $p(\mathrm{H}) = p(\mathrm{FA})^{1/k}$, where k is a measure of signal strength like *d'*. The AUC for a power ROC is $k/(1 + k)$. k may be estimated as $k = \ln(\mathrm{FA})/\ln(\mathrm{H})$, where *ln* is the natural logarithm. The AUC for Figure 3 is 0.75.

We seldom have direct knowledge of the underlying distributions. It is therefore reassuring that AUCs, even when calculated based on one or a few data points or under different assumptions about underlying distributions, are typically quite similar. Thus, the power-ROC estimate of the AUC in Figure 2 is 0.75, close to the estimate from *d'* of 0.73. Of course, having additional pairs of H and FA will greatly refine our measures, and help decide whether the ROCs are symmetric, as in Figure 2, or asymmetric as in Figure 3. How to achieve those is discussed below. Asymmetric ROCs are often modeled under the assumption of normal underlying distributions with different variances.

2AFC. The experiments treated so far are called *single stimulus, yes/no,* or *go/no-go*. In another standard situation the observer is confronted with two stimuli, A and B, and asked to decide which is the target. They may occur on the top and bottom of a computer screen simultaneously, or as the first or second of two successive presentations of auditory stimuli. Sometimes more than one alternative is presented, as in police eye-witness lineups. With two stimuli, the experiment is called a *2 alternative forced-choice* (2AFC). Intuitively we expect accuracy to be higher in 2AFC than in yes/no (Y/N) designs for at least two reasons. 2AFC reduces bias, as the target is present on every trial, just in a different place or order. Also the individual has more information clustered in space or time, and can judge the difference between the stimuli, not each independently. The difference gives an advantage similar to pair-comparison measures in statistical inference, and for similar reasons. In the case of an unbiased observer in a Y/N experiment, the signal is a distance *d'*/2 away from the criterion (see Figure 1). In 2AFC the signal is the difference in the means of A and B samples, on the average equal to *d'*, thus giving 2AFC a 2-to-1 advantage. Undermining this a bit is the presence of noise in both observations. If the underlying distributions are equal-variance, then the standard-deviation of the difference is $\sigma_{A-B} = \sigma\sqrt{2(1+\rho)}$, where ρ is the correlation between the samples of A and B. If the variables are uncorrelated ρ = 0, and the 2AFC offers a $2/\sqrt{2} = \sqrt{2}$ advantage in *d'*. If some of the noise is due to variability in the criterion, ρ will be negative, giving larger improvements in *d'*. Whatever the underlying distributions, when ρ = 0, the *area theorem* predicts that the probability of being correct in a 2AFC for an unbiased observer equals the AUC in the equivalent Y/N ROC: p_C = AUC.

Along with the above advantages for AUC, we close with two additional ones. It is intuitively clearer to say of a test: "A criterion-free measure of this test gives it 92% accuracy." than it is to say "The *d'* for this test is 2.0". The second advantage is that AUC = p_{rep}, where p_{rep} is the probability that another experiment of equal power will return a result in the same direction (e.g., that if one spam filter calls an email *spam*, that another, equally powerful one, will also; Irwin, 2009).

The Rating Scale ROC

A single point based on one pair (H, FA) is a thin reed onto which to place so much of the weight of the preceding paragraphs. Why not collect more data with different payoffs (Table 1), to bias the criterion from liberal to conservative, and thus dot out the true shape of the ROC? This has been done, but it takes patience. Because of sampling variability, many trials must be conducted for just a single decent point; multiply that by the number of points you want, and divide by the patience of the observer, to estimate the work required. There must be an easier way!

There is. Most of the time that an observer responds Y or N they could say much more. Not all responses are uttered with equal confidence. Individuals with a conservative criterion are presumably more confident that when they say Y, the target is really there, as they have chosen to call ambiguous stimuli N rather than risk a false alarm. If we let the observer tell us how confident they are—say on a 3-point scale of *positive, not sure, just guessing*—we arrive at a 6-point scale: *N, and I'm positive*, … etc, through *Y, and I'm positive*. Now imagine we conducted the experiment on 6 separate sessions, asking observers to adopt one of those criteria—on the first day, for instance, asking them to say N only if they are positive, and Y to all other stimuli. Signal detection theorists assume that the data collected in these two ways will be similar, but with an almost six-fold savings in labor in the first, rating scale design. Another advantage of rating scales is that the observers are less likely to become bored, and thus have declining levels of *d'* over the sessions; or to become expert, and thus have increasing levels of *d'* over the sessions.

To create a ROC from these data, envisage them as resulting from not just one criterion, as shown in Figure 1, but 6 of them corresponding to each confidence rating, 3 for N and 3 for Y. Call these criteria c_1 through c_6. Array them in a 6-column version of Table 2. To place the first point on the ROC, divide the number of responses in c_1 in the presence of A by that row total. This gives the conditional probability of N in the presence of A (a miss): $M_1 = 1 - H_1$. Compute the conditional probability of a response in c1 given the absence of A (a correct rejection), and call it $CR_1 = 1\text{-}FA_1$. That gives the first, most conservative point on the ROC. The next criterion separates the responses in the first *and* second categories from all the rest. Sum the responses in c_1 and c_2 when A was present, dividing once again by that row total to derive their conditional probability, M_2. Sum the responses in c_1 and c_2 when A was absent, and derive their conditional probability, CR_2. Continue the process. The resulting pairs of H_i and FA_i are the locus of the ROC.

A practical example. Figure 3 shows a rating scale ROC on an issue of interest to all citizens: Given the imperfection of clinical tests, when should we assert that an individual has a particular condition? All personality traits, from the most normal through strange to the craziest, are distributed on continua, much as the signals in Figure 1. All that we have learned in this chapter tells us that categorical decisions involve considerations of both detectability and bias. It is to everyone's advantage to have tests with large *d'*s, large AUCs. Just where we set the criterion for saying “not normal” depends on the costs and benefits from a table such as Table 1. Is the condition debilitating with safe and inexpensive remedies available? Then the criterion should be liberal. Is the condition merely a quirk, with only invasive and dangerous remedies available? Then the criterion should be very conservative indeed. These are some of the considerations that exercise the committees who determine criteria for categorization in the Diagnostic and Statistical Manual of Mental Disorders (the DSM).

One of the categories of the DSM is Attention Deficit Hyperactivity Disorder (ADHD). It would be valuable to have biometric measures to augment the parent and teacher reports that drive this categorization. This is especially important for different cultures where languages and expectations of children differ. As a step toward this goal, Meyer and Sagvolden (2006) collected several measures of fine motor skills among 7 ethnically diverse tribes in South Africa, along with their scores on a standardized measure of ADHD. I added the scores on two of their measures, and set a dozen criteria ranging from the lowest scores to the highest, calculating $p(H_i)$ and $p(FA_i)$ for each criterion as described above. I relied on the standard division into A (ADHD) and ~A given by the standard test to define signal present vs. absent. The ROC in Figure 3 resulted, giving a criterion free accuracy (AUC) of 75%. By itself this is not high enough for clinical use (motor disturbances could occur for many reasons), but along with other measures could improve categorization accuracy. Or it could be used as a quick pre-screening test (in such uses, with a liberal criterion) to save resources for better analysis or treatment of those affected.

It is in such clinical diagnoses that the importance of decision theory is most obvious. There are many—often rancorous—debates about the proper approach to ADHD, ones that often conflate *d'* with *c*. It is not necessary, however, to deny the existence of ADHD in order to argue against medication, as the evaluation of costs and benefits in Table 1 will honestly differ among teacher, parent and child. Nor is it necessary to believe in ADHD as a discrete disability in order to recommend medication or behavioral therapy, as those will be of some benefit (and cost!) no matter on which side of an (always somewhat arbitrary) criterion line a child's scores fall. Decision theory and its offspring SDT provide important frameworks for this discussion. Table 1 reminds us that the alternative to a diagnosis like ADHD is not necessarily complete normality (rare as that is!). Nor should the courses of action be restricted to binary alternatives; a range of strategies should be evaluated. It reminds participants to clarify costs and benefits, and to educe their values, and those of other parties with standing. Finally SDT and its icons remind us of the importance of improving signal strength—AUC—as accurate diagnoses help bridle the role played by criterion placement; it reminds us that most signals reside on a continuum with many shades of gray. And it partitions responsibilities: specifying and improving signal strength is a technical undertaking, often best left to experts; clarifying options, outcomes and values is a more deliberative political process involving clear discussions among all interested parties.

Bibliography

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley and Sons, Inc. The ur-text of SDT.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. An invaluable reference book, the bible of SDT.

McNichol, D. (2005). *A primer of signal detection theory*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Excellent undergraduate text.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY: Oxford University Press. "Alas, 'elementary' is not synonymous with 'easy'" p.vii. But really not that hard either for an interested graduate student.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286. In this and other articles, Thurstone planted the meme of Figure 1.

Meyer, A., & Sagvolden, T. (2006). Fine motor skills in South African children with symptoms of ADHD: influence of subtype, gender, age, and hand dominance. *Behavioral and Brain Functions, 2*, 33. The study that provided the data for Figure 3.

Irwin, R. J. (2009). Equivalence of the statistics for replicability and area under the ROC curve. *British Journal of Mathematical and Statistical Psychology, 62*, 485-487. Connects SDT to the statistics of replicability.

Killeen, P. R., & Taylor, T. J. (2004). Symmetric receiver operating characteristics. *Journal of Mathematical Psychology, 48*, 432-434. Gives the conditions required of underlying distributions to yield symmetric ROCs.

Killeen, P. R. (2003). The yins/yangs of science. *Behavior and Philosophy, 31*, 251-258. Applies SDT to love, truth, and politics.

Thanks to Ray Nickerson