Behavior Research Methods, Instruments, & Computers
1999, 31 (1), 137-149

Calculation of signal detection theory measures

HAROLD STANISLAW
California State University, Stanislaus, Turlock, California
and
NATASHA TODOROV
Macquarie University, Sydney, New South Wales, Australia

Signal detection theory (SDT) may be applied to any area of psychology in which two different types of stimuli must be discriminated. We describe several of these areas and the advantages that can be realized through the application of SDT. Three of the most popular tasks used to study discriminability are then discussed, together with the measures that SDT prescribes for quantifying performance in these tasks. Mathematical formulae for the measures are presented, as are methods for calculating the measures with lookup tables, computer software specifically developed for SDT applications, and general purpose computer software (including spreadsheets and statistical analysis software).

Signal detection theory (SDT) is widely accepted by psychologists; the *Social Sciences Citation Index* cites over 2,000 references to an influential book by Green and Swets (1966) that describes SDT and its application to psychology. Even so, fewer than half of the studies to which SDT is applicable actually make use of the theory (Stanislaw & Todorov, 1992). One possible reason for this apparent underutilization of SDT is that relevant textbooks rarely describe the methods needed to implement the theory. A typical example is Goldstein's (1996) popular perception textbook, which concludes a nine-page description of SDT with the statement that measures prescribed by SDT "can be calculated . . . by means of a mathematical procedure we will not discuss here" (p. 594).

The failure of many authors to describe SDT's methods may have been acceptable when lengthy, specialized tables were required to implement the theory. Today, however, readily available computer software makes an SDT analysis no more difficult than a *t* test. The present paper attempts to demonstrate this and to render SDT available to a larger audience than currently seems to be the case.

We begin with a brief overview of SDT, including a description of its performance measures. We then present the formulae needed to calculate these measures. Next, we describe different methods for calculating SDT measures. Finally, we provide sample calculations so that readers can verify their understanding and implementation of the techniques.

We are indebted to James Thomas, Neil Macmillan, John Swets, Douglas Creelman, Scott Maxwell, Mark Frank, Helena Kadlec, and an anonymous reviewer for providing insightful comments on earlier versions of this manuscript. We also thank Mack Goldsmith for testing some of our spreadsheet commands. Correspondence concerning this article should be addressed to H. Stanislaw, Department of Psychology, California State University, Stanislaus, 801 West Monte Vista Avenue, Turlock, CA 95382 (e-mail: hstanisl@toto.csustan.edu).

OVERVIEW OF SIGNAL DETECTION THEORY

Proper application of SDT requires an understanding of the theory and the measures it prescribes. We present an overview of SDT here; for more extensive discussions, see Green and Swets (1966) or Macmillan and Creelman (1991). Readers who are already familiar with SDT may wish to skip this section.

SDT can be applied whenever two possible stimulus types must be discriminated. Psychologists first applied the theory in studies of perception, where subjects discriminated between *signals* (stimuli) and *noise* (no stimuli). The signal and noise labels remain, but SDT has since been applied in many other areas. Examples (and their corresponding signal and noise stimuli) include recognition memory (old and new items), lie detection (lies and truths), personnel selection (desirable and undesirable applicants), jury decision making (guilty and innocent defendants), medical diagnosis (diseased and well patients), industrial inspection (unacceptable and acceptable items), and information retrieval (relevant and irrelevant information; see also Hutchinson, 1981; Swets, 1973; and the extensive bibliographies compiled by Swets, 1988b, pp. 685–742). Performance in each of these areas may be studied with a variety of tasks. We deal here with three of the most popular: *yes/no* tasks, *rating* tasks, and *forced-choice* tasks.

Yes/No Tasks

A yes/no task involves *signal trials*, which present one or more signals, and *noise trials*, which present one or more noise stimuli. For example, yes/no tasks in auditory perception may present a tone during signal trials and nothing at all during noise trials, whereas yes/no tasks for memory may present *old* (previously studied) words during signal trials and *new* (distractor) words during noise trials. After each trial, the subjects indicate whether a sig-

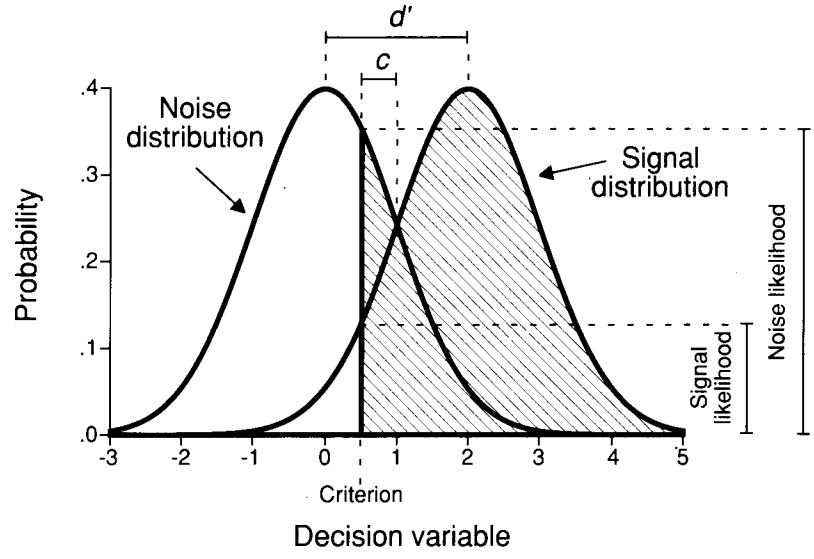

Figure 1. Distribution of the decision variable across noise and signal trials, showing *d'*, *c*, and the likelihoods on which β is based.

nal was presented (i.e., whether a tone was presented, or whether the word was previously studied).

According to SDT, the subjects in a yes/no task base their response on the value that a *decision variable* achieves during each trial. If the decision variable is sufficiently high during a given trial, the subject responds *yes* (a signal was presented); otherwise, the subject responds *no* (no signal was presented). The value that defines *sufficiently high* is called the *criterion*.

For example, consider a psychologist attempting to determine whether or not a child has attention deficit hyperactivity disorder (ADHD). The psychologist might administer the Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1983) and diagnose the child as having ADHD if the resulting score is 10 or higher (Rey, Morris-Yates, & Stanislaw, 1992). In this case, the decision variable is the CBCL score, and the criterion is set at 10.

In the ADHD example, the decision variable is readily observed. However, most of the tasks studied by psychologists involve decision variables that are available only to the subject performing the task. For example, the decision variable may be the apparent loudness experienced during each trial in an auditory perception study, the feeling of familiarity associated with each stimulus item in a memory study, or the apparent guilt of each defendant in a study of jury decision making. In each of these cases, the subjects compare the decision variable (which only they can observe) to the criterion they have adopted. A *yes* response is made only if the auditory stimulus seems sufficiently loud, the stimulus item seems sufficiently familiar, or the defendant seems sufficiently guilty.

On signal trials, *yes* responses are correct and are termed *hits*. On noise trials, *yes* responses are incorrect and are termed *false alarms*. The *hit rate* (the probability of responding *yes* on signal trials) and the *false-alarm rate* (the probability of responding *yes* on noise trials) fully describe performance on a yes/no task.

If the subject is using an appropriate decision variable, and if the subject is capable of distinguishing between signals and noise, the decision variable will be affected by the stimuli that are presented. For example, previously studied words in a memory study should, on average, seem more familiar than distractors. However, some previously studied words will seem more familiar than others. Distractors will also vary in their familiarity. Furthermore, factors such as neural noise and fluctuations in attention may affect the decision variable, even if the stimulus is held constant. Thus, the decision variable will have a range of different values across signal trials and a range of different values across noise trials. (For more examples of this, see McNicol, 1972, pp. 11–14.)

The distribution of values realized by the decision variable across signal trials is the *signal distribution*, whereas the corresponding distribution for noise trials is the *noise distribution*. The hit rate equals the proportion of the signal distribution that exceeds the criterion, whereas the false-alarm rate equals the proportion of the noise distribution that exceeds the criterion. This is illustrated in Figure 1, where the decision variable (measured in arbitrary units) has a mean of 0 and a standard deviation of 1 on noise trials. On signal trials, the mean is higher ($M = 2$), but the standard deviation is unchanged. A *yes* response is made for trials in which the decision variable exceeds 0.5; these trials lie in the shaded region of the two distributions. The shaded region of the noise distribution constitutes 30.85% of the entire noise distribution, so the false-alarm rate is .3085. By contrast, 93.32% of the signal distribution is shaded, so the hit rate is .9332.

If the criterion is set to an even lower, or more *liberal*, value (i.e., moved to the far left in Figure 1), it will almost

always be exceeded on signal trials. This will produce mostly *yes* responses and a high hit rate. However, the criterion will also be exceeded on most noise trials, resulting in a high proportion of *yes* responses on noise trials (i.e., a high false-alarm rate). Thus, a liberal criterion biases the subject toward responding *yes*, regardless of the stimulus. By contrast, a high, or *conservative*, value for the criterion biases the subject toward responding *no*, because the criterion will rarely be exceeded on signal or noise trials. This will result in a low false-alarm rate, but also a low hit rate. The only way to increase the hit rate while reducing the false-alarm rate is to reduce the overlap between the signal and the noise distributions.

Clearly, the hit and false-alarm rates reflect two factors: *response bias* (the general tendency to respond *yes* or *no*, as determined by the location of the criterion) and the degree of overlap between the signal and the noise distributions. The latter factor is usually called *sensitivity*, reflecting the perceptual origins of SDT: When an auditory signal is presented, the decision variable will have a greater value (the stimulus will sound louder) in listeners with more sensitive hearing. The major contribution of SDT to psychology is the separation of response bias and sensitivity. This point is so critical that we illustrate it with five examples from widely disparate areas of study.

Our first example is drawn from perception, where higher thresholds have been reported for swear words than for neutral stimuli (Naylor & Lawshe, 1958). A somewhat Freudian interpretation of this finding is that it reflects a change in sensitivity that provides perceptual "protection" against negative stimuli (Erdelyi, 1974). However, a false alarm is more embarrassing for swear words than for neutral stimuli. Furthermore, subjects do not expect to encounter swear words in a study and are, therefore, cautious about reporting them. Thus, different apparent thresholds for negative than for neutral stimuli may stem from different response biases, as well as from different levels of sensitivity. In order to determine which explanation is correct, sensitivity and response bias must be measured separately.

A second example involves memory studies. Hypnosis sometimes improves recall, but it also increases the number of intrusions (false alarms). It is important, therefore, to determine whether hypnosis actually improves memory, or whether demand characteristics cause hypnotized subjects to report more memories about which they are uncertain (Klatzky & Erdelyi, 1985). The former explanation implies an increase in sensitivity, whereas the latter implies that hypnosis affects response bias. Again, it is important to measure sensitivity and response bias separately.

Our third example involves a problem that sometimes arises when comparing the efficacy of two tests used to diagnose the same mental disorder. One test may have a higher hit rate than the other, but a higher false-alarm rate as well. This problem typically arises because the tests use different criteria for determining when the disorder is actually present. SDT can solve this problem by determining the sensitivity of each test in a metric that is independent of the criterion (Rey et al., 1992).

A fourth example concerns jury decision making. SDT analyses (Thomas & Hogue, 1976) have shown that jury instructions regarding the definition of *reasonable doubt* affect response bias (the willingness to convict) rather than sensitivity (the ability to distinguish guilty from innocent defendants). Response bias also varies with the severity of the case: Civil cases and criminal cases with relatively lenient sentences require less evidence to draw a conviction than criminal cases with severe penalties.

Our final example is drawn from industry, where quality-control inspectors often detect fewer faulty items as their work shift progresses. This declining hit rate usually results from a change in response bias (Davies & Parasuraman, 1982, pp. 60–99), which has led to remedies that would fail if declining sensitivity were to blame (Craig, 1985). SDT also successfully predicts how repeated inspections improve performance (Stanislaw, 1995).

Medical diagnosis illustrates the problems that existed before SDT was developed. In his presidential address to the Radiological Society of North America, Garland (1949) noted that different radiologists sometimes classified the same X ray differently. This problem was considered both disturbing and mysterious until it was discovered that different response biases were partly to blame. Subsequently, radiologists were instructed first to examine all images, using a liberal criterion, and then to reexamine positive images, using a conservative criterion.

Sensitivity and response bias are confounded by most performance measures, including the hit rate, the false-alarm rate, the hit rate "corrected" by subtracting the false-alarm rate, and the proportion of correct responses in a yes/no task. Thus, if (for example) the hit rate varies between two different conditions, it is not clear whether the conditions differ in sensitivity, response bias, or both.

One solution to this problem involves noting that sensitivity is related to the distance between the mean of the signal distribution and the mean of the noise distribution (i.e., the distance between the peaks of the two distributions in Figure 1). As this distance increases, the overlap between the two distributions decreases. Overlap also decreases if the means maintain a constant separation but the standard deviations decrease. Thus, sensitivity can be quantified by using the hit and false-alarm rates to determine the distance between the means, relative to their standard deviations.

One measure that attempts to do this is d', which measures the distance between the signal and the noise means in standard deviation units. The use of standard deviates often makes it difficult to interpret particular values of d'. However, a value of 0 indicates an inability to distinguish signals from noise, whereas larger values indicate a correspondingly greater ability to distinguish signals from noise. The maximum possible value of d' is $+\infty$, which signifies perfect performance. Negative values of d' can arise through sampling error or response confusion (responding

yes when intending to respond *no*, and vice versa); the minimum possible value is $-\infty$. d' has a value of 2.00 in Figure 1, as the distance between the means is twice as large as the standard deviations of the two distributions.

SDT states that d' is unaffected by response bias (i.e., is a pure measure of sensitivity) if two assumptions are met regarding the decision variable: (1) The signal and noise distributions are both normal, and (2) the signal and noise distributions have the same standard deviation. We call these the *d' assumptions*. The assumptions cannot actually be tested in yes/no tasks; rating tasks are required for this purpose. However, for some yes/no tasks, the d' assumptions may not be tenable; the assumption regarding the equality of the signal and the noise standard deviations is particularly suspect (Swets, 1986).

If either assumption is violated, d' will vary with response bias, even if the amount of overlap between the signal and the noise distributions remains constant. Because of this, some researchers prefer to use *nonparametric* measures of sensitivity. These measures may also be used when d' cannot be calculated. (This problem is discussed in more detail below.) Several nonparametric measures of sensitivity have been proposed (e.g., Nelson, 1984, 1986; W. D. Smith, 1995), but the most popular is A'. This measure was devised by Pollack and Norman (1964); a complete history is provided by Macmillan and Creelman (1996) and W. D. Smith. A' typically ranges from .5, which indicates that signals cannot be distinguished from noise, to 1, which corresponds to perfect performance. Values less than .5 may arise from sampling error or response confusion; the minimum possible value is 0. In Figure 1, A' has a value of .89.

Response bias in a yes/no task is often quantified with β. Use of this measure assumes that responses are based on a likelihood ratio. Suppose the decision variable achieves a value of x on a given trial. The numerator for the ratio is the likelihood of obtaining x on a signal trial (i.e., the height of the signal distribution at x), whereas the denominator for the ratio is the likelihood of obtaining x on a noise trial (i.e., the height of the noise distribution at x). Subjects respond *yes* ifz the likelihood ratio (or a variable monotonically related to it) exceeds β, and *no* otherwise.

When subjects favor neither the *yes* response nor the *no* response, $\beta = 1$. Values less than 1 signify a bias toward responding *yes*, whereas values of β greater than 1 signify a bias toward the *no* response. In Figure 1, the signal likelihood is .1295 at the criterion, whereas the noise likelihood is .3521 at the criterion. Thus, β equals .1295 ÷ .3521, or .37. This implies that the subjects will respond *yes* on any trial in which the height of signal distribution at x, divided by the height of the noise distribution at x, exceeds 0.37.

Because β is based on a ratio, the natural logarithm of β is often analyzed in place of β itself (McNicol, 1972, pp. 62–63). Negative values of $\ln(\beta)$ signify bias in favor of *yes* responses, whereas positive values of $\ln(\beta)$ signify bias in favor of *no* responses. A value of 0 signifies that no response bias exists. In Figure 1, the natural logarithm of β equals -1.00.

Historically, β has been the most popular measure of response bias. However, many authors now recommend measuring response bias with c (Banks, 1970; Macmillan & Creelman, 1990; Snodgrass & Corwin, 1988). This measure assumes that subjects respond *yes* when the decision variable exceeds the criterion and *no* otherwise; responses are based directly on the decision variable, which some researchers regard as more plausible than assuming that responses are based on a likelihood ratio (Richardson, 1994). Another advantage of c is that it is unaffected by changes in d', whereas β is not (Ingham, 1970; Macmillan, 1993; McNicol, 1972, pp. 63–64).

c is defined as the distance between the criterion and the *neutral point*, where neither response is favored. The neutral point is located where the noise and signal distributions cross over (i.e., where $\beta = 1$). If the criterion is located at this point, c has a value of 0. Deviations from the neutral point are measured in standard deviation units. Negative values of c signify a bias toward responding *yes* (the criterion lies to the left of the neutral point), whereas positive values signify a bias toward the *no* response (the criterion lies to the right of the neutral point). In Figure 1, the neutral point is located 1 standard deviation above the noise mean. The criterion is located 0.50 standard deviations to the left of this point, so c has a value of -0.50.

A popular nonparametric measure of response bias is B''. This measure was devised by Grier (1971) from a similar measure proposed by Hodos (1970). Unfortunately, some researchers have confused the two measures, using the formula for Grier's B'' to compute what is claimed to be Hodos's bias measure (e.g., Macmillan & Creelman, 1996). Both nonparametric bias measures can range from -1 (extreme bias in favor of *yes* responses) to 1 (extreme bias in favor of *no* responses). A value of 0 signifies no response bias. In Figure 1, Grier's B'' has a value of -0.55.

Rating Tasks

Rating tasks are like yes/no tasks, in that they present only one stimulus type during each trial. However, rather than requiring a dichotomous (*yes* or *no*) response, rating tasks require graded responses. For example, subjects in a study of jury decision making may rate each defendant on a scale of 1 (*most certainly guilty*) to 6 (*most certainly innocent*). These ratings can be used to determine points on a *receiver operating characteristic* (ROC) curve, which plots the hit rate as a function of the false-alarm rate for all possible values of the criterion. A typical ROC curve is illustrated in Figure 2.

A rating task with r ratings determines $r - 1$ points on the ROC curve, each corresponding to a different criterion. For example, one criterion distinguishes ratings of "1" from ratings of "2." Another criterion distinguishes ratings of "2" from ratings of "3," and so on. The ROC curve in Figure 2 was generated from a rating task with four possible ratings, resulting in three points (the open squares).

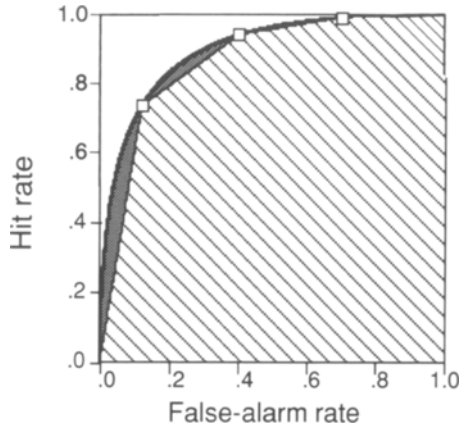

Figure 2. Receiver operating characteristic (ROC) curve for a rating task with four ratings and A_z = .90. Three points on the ROC curve are shown (open squares). The area under the curve, as estimated by linear extrapolation, is indicated by shading; the actual area includes the gray regions.

The remainder of the ROC curve is determined by extrapolation, as will be described below.

Rating tasks are primarily used to measure sensitivity. According to SDT, the area under the ROC curve is a measure of sensitivity unaffected by response bias. The ROC area typically ranges from .5 (signals cannot be distinguished from noise) to 1 (perfect performance). Areas less than .5 may arise from sampling error or response confusion; the minimum possible value is 0. The area under the ROC curve in Figure 2 equals .90; this includes both the large shaded region and the smaller gray regions.

The ROC area can be interpreted as the proportion of times subjects would correctly identify the signal, if signal and noise stimuli were presented simultaneously (Green & Moses, 1966; Green & Swets, 1966, pp. 45–49). Thus, Figure 2 implies that signals would be correctly identified on 90% of the trials in which signal and noise stimuli were presented together.

The ROC area can be estimated quite easily if linear extrapolation is used to connect the points on the ROC curve (for details, see Centor, 1985, and Snodgrass, Levy-Berger, & Haydon, 1985, pp. 449–454). However, this approach generally underestimates sensitivity. For example, linear extrapolation measures only the shaded region in Figure 2; the gray regions are excluded. A common remedy is to assume that the decision variable is normally distributed and to use this information to fit a curvilinear function to the ROC curve. The ROC area can then be estimated from parameters of the curvilinear function. When this procedure is used, the ROC area is called A_z, where the z (which refers to z scores) indicates that the noise and signal distributions are assumed to be normal.

In calculating A_z, no assumptions are made concerning the decision variable's standard deviation. This differs from yes/no tasks, where one of the d' assumptions is that the standard deviation for signals equals that for noise. In fact, rating tasks may be used to determine the validity of the assumption regarding equal standard deviations.

The ROC area can be calculated from yes/no data, as well as from rating data. One method involves using d' to estimate the ROC area; the resulting measure is called $A_{d'}$. This measure is valid only when the d' assumptions are met. A' also estimates the ROC area, and does so without assuming that the decision variable has a particular (e.g., normal) distribution. However, A' is problematic in other respects (Macmillan & Kaplan, 1985; W. D. Smith, 1995). In general, A_z (when it can be calculated) is the preferred measure of the ROC area and, thus, of sensitivity (Swets, 1988a; Swets & Pickett, 1982, pp. 31–32).

Forced-Choice Tasks

In a forced-choice task, each trial presents one signal and one or more noise stimuli. The subjects indicate which stimulus was the signal. Tasks are labeled according to the total number of stimuli presented in each trial; an m-alternative forced-choice (mAFC) task presents one signal and $m - 1$ noise stimuli. For example, a 3AFC task for the study of recognition memory presents one old item and two distractors in each trial. The subjects indicate which of the three stimuli is the old item.

Each stimulus in an mAFC trial affects the decision variable. Thus, each mAFC trial yields m decision variable scores. Subjects presumably compare these m scores (or their corresponding likelihood ratios) with each other and determine which is the largest and, thus, the most likely to have been generated by a signal. Because this comparison does not involve a criterion, mAFC tasks are only suitable for measuring sensitivity.

SDT states that, if subjects do not favor any of the m alternatives a priori, the proportion of correct responses on an mAFC task is a measure of sensitivity unaffected by response bias (Green & Swets, 1966, pp. 45–49). This measure typically ranges from $1/m$ (chance performance) to a maximum of 1 (perfect performance). Values less than $1/m$ may result from sampling error or response confusion; the minimum possible value is 0.

FORMULAE FOR CALCULATING SIGNAL DETECTION THEORY MEASURES

Few textbooks provide any information about SDT beyond that just presented. This section provides the mathematical details that textbooks generally omit. Readers who would rather not concern themselves with formulae may skip this section entirely; familiarity with the underlying mathematical concepts is not required to calculate SDT measures. However, readers who dislike handwaving, or who desire a deeper understanding of SDT, are encouraged to read on.

In the discussion below, H is used to indicate the hit rate. This rate is found by dividing the number of hits by

the total number of signal trials. Similarly, the false-alarm rate, F, is found by dividing the number of false alarms by the total number of noise trials.

Some SDT measures can only be calculated with the aid of two mathematical functions. One of these, the Φ ("phi") function, converts z scores into probabilities. This same conversion is used to perform a z test. However, the Φ function determines the portion of the normal distribution that lies to the *left* of the z score; larger z scores yield higher probabilities. The z test, by contrast, determines the portion to the *right* of the z score; larger z scores yield smaller probabilities. Furthermore, the Φ function is one-tailed, whereas many z tests are two-tailed. For example, $\Phi(-1.64) = .05$, which means that a probability of .05 is associated with a z score of -1.64, not 1.96 (the critical value for a two-tailed z test).

The second mathematical function sometimes needed to calculate SDT measures is the Φ^{-1} ("inverse phi") function. This complements the Φ function and converts probabilities into z scores. For example, $\Phi^{-1}(.05) = -1.64$, which means that a one-tailed probability of .05 requires a z score of -1.64.

Yes/No Tasks

d' can be calculated as follows (Macmillan, 1993):

$$d' = \Phi^{-1}(H) - \Phi^{-1}(F). \tag{1}$$

Thus, d' is found by subtracting the z score that corresponds to the false-alarm rate from the z score that corresponds to the hit rate.

A' can be calculated as follows (Snodgrass & Corwin, 1988):

$$A' = \begin{cases} .5 + \dfrac{(H-F)(1+H-F)}{4H(1-F)} & \text{when } H \geq F \\ .5 - \dfrac{(F-H)(1+F-H)}{4F(1-H)} & \text{when } H < F \end{cases}. \tag{2}$$

Some publications list only the formula to be used when $H \geq F$ (e.g., Grier, 1971; note also the typographical error in Cradit, Tashchian, & Hofacker, 1994). The need for two different formulae is awkward, but Equation 2 can be rewritten into a single formula, as follows:

$$A' = .5 + \left[\operatorname{sign}(H-F) \frac{(H-F)^2 + |H-F|}{4\max(H,F) - 4HF} \right], \tag{3}$$

where $\operatorname{sign}(H-F)$ equals +1 if $H - F > 0$ (i.e., if $H > F$), 0 if $H = F$, and -1 otherwise, and $\max(H, F)$ equals either H or F, whichever is greater.

β may be calculated in a variety of ways. Many authors (e.g., Brophy, 1986) suggest using the formula

$$\beta = \frac{e^{-.5[\Phi^{-1}(H)]^2}}{\sqrt{2p}} \div \frac{e^{-.5[\Phi^{-1}(F)]^2}}{\sqrt{2p}}. \tag{4}$$

However, this can be rewritten as

$$\beta = e^{\left\{ \frac{\left[\Phi^{-1}(F)\right]^2 - \left[\Phi^{-1}(H)\right]}{2} \right\}^2}, \tag{5}$$

which is simpler and less prone to round-off error. The natural logarithm of β is then

$$\ln(\beta) = \frac{\left[\Phi^{-1}(F)\right]^2 - \left[\Phi^{-1}(H)\right]^2}{2}. \tag{6}$$

Thus, the natural logarithm of β is found by squaring the z score that corresponds to the false-alarm rate, subtracting the square of the z score that corresponds to the hit rate, and dividing the result by 2.

The formula for c (Macmillan, 1993) is

$$c = -\frac{\Phi^{-1}(H) + \Phi^{-1}(F)}{2}. \tag{7}$$

Thus, c is found by averaging the z score that corresponds to the hit rate and the z score that corresponds to the false-alarm rate, then multiplying the result by negative one. Some authors (e.g., Snodgrass & Corwin, 1988) omit the minus sign, which simply means that negative values of c indicate a bias toward responding *no*, rather than *yes*.

Grier's B'' can be found as follows (Snodgrass & Corwin, 1988):

$$B'' = \begin{cases} \dfrac{H(1-H) - F(1-F)}{H(1-H) + F(1-F)} & \text{when } H \geq F \\ \dfrac{F(1-F) - H(1-H)}{F(1-F) + H(1-H)} & \text{when } H < F \end{cases}. \tag{8}$$

When Grier (1971) first proposed this measure, he published only the formula to be used when $H \geq F$. If this formula is applied when $H < F$, B'' has the correct magnitude but the wrong sign. B'' may be found with a single formula, as follows:

$$B'' = \operatorname{sign}(H-F) \frac{H(1-H) - F(1-F)}{H(1-H) + F(1-F)}. \tag{9}$$

Rating Tasks

To calculate A_z, the $r - 1$ pairs of hit and false-alarm rates must first be found. An iterative procedure is used to accomplish this. In describing this procedure, it is assumed that numerically higher ratings indicate greater confidence that the stimulus was a signal.

First, ratings greater than 1 are considered to be *yes* responses, whereas ratings of 1 are considered to be *no* responses. The resulting hit and false-alarm rates are then determined. Next, ratings greater than 2 are considered to be *yes* responses, whereas ratings of 1 or 2 are considered to be *no* responses. This yields a second pair of hit and false-alarm rates. This procedure is repeated until all $r - 1$ pairs of hit and false-alarm rates have been determined.

The Φ^{-1} function is then used to find the z scores for each pair of hit and false-alarm rates. This is analogous to

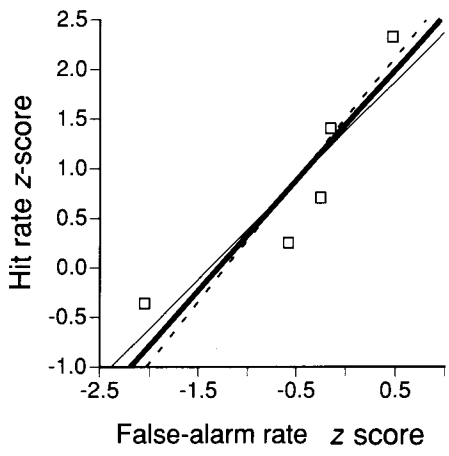

Figure 3. Receiver operating characteristic (ROC) curve plotted in z space, for a rating task with six ratings and $A_z = .82$. The three diagonal lines represent different "best" fits to the five points on the ROC curve (open squares). The thin line predicts hit rate z scores from false-alarm rate z scores; the broken line predicts false-alarm rate z scores from hit rate z scores. The heavy line is a compromise between the two other lines.

plotting the ROC curve in z space, where the axes are z scores rather than probabilities (see Figure 3). Continuing this analogy, the slope and intercept of the line that best fits the z scores in z space are found. A_z can then be found as follows (Swets & Pickett, 1982, p. 33):

$$A_z = \Phi\left[\frac{\text{Intercept}}{\sqrt{1+(\text{Slope})^2}}\right]. \quad (10)$$

The slope of the best-fitting line in z space equals the noise distribution standard deviation divided by the signal distribution standard deviation. Thus, the d' assumption of equal standard deviations for the signal and noise distributions can be tested by determining whether the slope equals 1 (or whether the logarithm of the slope equals 0, which is the better test from a statistical viewpoint).

Forced-Choice Tasks

Sensitivity in a forced-choice task is determined by simply dividing the number of correct responses by the total number of trials.

Comparing Performance Across Different Types of Tasks

d' may be used to estimate the ROC area as follows (Macmillan, 1993):

$$A_{d'} = \Phi\left(\frac{d'}{\sqrt{2}}\right). \quad (11)$$

If the d' assumptions are satisfied, $A_{d'}$ should equal the proportion of correct responses that would have been obtained had subjects performed a 2AFC task instead of a yes/no task. In other words, $A_{d'}$ and the proportion correct on a 2AFC task quantify performance with the same metric, thereby allowing comparisons to be made between yes/no and 2AFC tasks (see, e.g., Stanislaw, 1995, 1996).

The proportion of correct responses on a 2AFC task should equal the area under the ROC curve. Thus, if the decision variable is normally distributed, A_z from a rating task should equal the proportion correct on a 2AFC task, which in turn should equal $A_{d'}$ from a yes/no task, if the decision variable has the same standard deviation for both types of stimuli (Macmillan, 1993).

Unfortunately, no such prediction can be made for A'. In fact, A' may differ from $A_{d'}$, even though both measures estimate the area under the ROC curve. For example, in Figure 1, $A_{d'} = .92$ and $A' = .89$.

Converting Between z Scores and Probabilities

Clearly, calculation of many SDT measures requires methods for converting z scores into probabilities (the Φ function) and probabilities into z scores (the Φ^{-1} function). Neither conversion is straightforward, because closed-form solutions to the underlying equations do not exist. However, several alternatives have been developed.

The most accurate of these—at least in theory—involve iterative calculations that converge on the exact solution. For example, the Φ function can be represented by a power series (Zelen & Severo, 1972, Equation 26.2.10). Adding terms to the series increases accuracy, but, in practice, a limit is imposed by round-off errors.

Another approach involves the use of closed-form approximations with known accuracy bounds. For example, one approximation to the Φ function (Zelen & Severo, 1972, Equation 26.2.17) yields probabilities that are accurate to within $\pm 7.5 \times 10^{-8}$.

Hit and False-Alarm Rates of Zero or One

Regardless of the approach used for the Φ and Φ^{-1} functions, problems may arise when the hit or false-alarm rate equals 0, because the corresponding z score is $-\infty$. Similarly, a hit or false-alarm rate of 1 corresponds to a z score of $+\infty$. These extreme values are particularly likely to arise when signals differ markedly from noise, few trials are presented (so that sampling error is large), or subjects adopt extremely liberal or conservative criteria (as might occur if, for example, the consequences of a false alarm are severe).

If both rates have extreme values, d' and $A_{d'}$ can still be calculated. When $H = 1$ and $F = 0$, $d' = +\infty$ and $A_{d'} = 1$. When $H = 0$ and $F = 1$, $d' = -\infty$ and $A_{d'} = 0$. When both rates are 0 or both rates are 1, most researchers assume that $d' = 0$ and $A_{d'} = .5$. However, if one rate has an extreme value and the other does not, d' and $A_{d'}$ are indeterminate.

Several solutions to this problem have been proposed. One possibility is to quantify sensitivity with nonparametric measures, such as A' (Craig, 1979). These measures eliminate reliance on the Φ and Φ^{-1} functions but are controversial. Macmillan and Creelman (1996) have argued against the use of A'; however, Donaldson (1993) suggests

that A' may estimate sensitivity better than d' when the signal and noise distributions are normal but have different standard deviations.

Another alternative is to combine the data from several subjects before calculating the hit and false-alarm rates (Macmillan & Kaplan, 1985). However, this approach complicates statistical testing and should only be applied to subjects who have comparable response biases and levels of sensitivity.

A third approach, dubbed *loglinear*, involves adding 0.5 to both the number of hits and the number of false alarms and adding 1 to both the number of signal trials and the number of noise trials, before calculating the hit and false-alarm rates. This seems to work reasonably well (Hautus, 1995). Advocates of the loglinear approach recommend using it regardless of whether or not extreme rates are obtained.

A fourth approach involves adjusting only the extreme rates themselves. Rates of 0 are replaced with $0.5 \div n$, and rates of 1 are replaced with $(n - 0.5) \div n$, where n is the number of signal or noise trials (Macmillan & Kaplan, 1985). This approach yields biased measures of sensitivity (Miller, 1996) and may be less satisfactory than the loglinear approach (Hautus, 1995). However, it is the most common remedy for extreme values and is utilized in several computer programs that calculate SDT measures (see, e.g., Dorfman, 1982). Thus, it is the convention we adopt in our computational example below.

METHODS FOR CALCULATING SDT MEASURES

In this section, we describe three general approaches that can be used to calculate the measures prescribed by SDT: tabular methods, methods that use software specifically developed for SDT, and methods that rely on general purpose software.

Tabular Methods

Elliott (1964) published one of the first and most popular listings of d' values for particular pairs of hit and false-alarm rates. Similar tables followed. The most extensive of these is Freeman's (1973), which also lists values of β. (See Gardner, Dalsing, Reyes, & Brake, 1984, for a briefer table of β values.)

Tables for d' are not restricted just to yes/no tasks. Elliott (1964) published an early d' table for forced-choice tasks; more recent versions that correct some errors have since appeared (Hacker & Ratcliff, 1979; Macmillan & Creelman, 1991, pp. 319–322). Tables for other tasks have been published by Craven (1992), Hershman and Small (1968), Kaplan, Macmillan, and Creelman (1978), and Macmillan and Creelman (1991, pp. 323–354).

Tabular methods have relatively poor accuracy. Some tables contain incorrect entries, but even error-free tables can be used only after the hit and false-alarm rates are rounded off (usually to two significant digits). Rounding introduces errors; changing the fourth significant digit of the hit or the false-alarm rate can often affect the second significant digit of d'. Interpolation between tabled values can minimize the impact of rounding errors, but the proper interpolation method is nonlinear. Furthermore, even linear interpolation requires calculations that the tabular method is specifically designed to avoid.

When SDT was first developed, most researchers were forced to rely on the tabular approach. However, this approach is difficult to justify today. Computers can quantify SDT performance far more quickly and accurately than can tables. Computers also provide the only reasonable means of analyzing rating task data; tables can determine neither the slope nor the intercept of the best-fitting line in z space.

Some computer programs (e.g., Ahroon & Pastore, 1977) calculate SDT measures by incorporating look-up tables, thus gaining a slight speed advantage over closed-form approximations and iterative techniques. However, speed is likely to be of concern only in Monte Carlo simulations involving thousands of replications. Even here, tables are of questionable utility, because of their limited accuracy. Thus, the tabular approach should be used only as a last resort.

Signal Detection Theory Software

Ahroon and Pastore (1977) were the first authors to publish programs (written in FORTRAN and BASIC) for determining values of d' and β for yes/no tasks. However, these programs should be used with caution, as they rely on look-up tables for the Φ^{-1} function. Furthermore, β is calculated with Equation 4 rather than Equation 5. This imposes speed and accuracy penalties, thus offsetting whatever advantage might be gained from the tabular approach.

Better choices for analyzing yes/no data are Brophy's (1986) BASIC program and the Pascal program published by Macmillan and Creelman (1991, pp. 358–359). Other published programs are written in APL (McGowan & Appel, 1977) and Applesoft (which is similar to BASIC; Gardner & Boice, 1986). Programmers who wish to write their own code can refer to algorithms for the Φ^{-1} function (see Brophy, 1985, for a review) and then apply the appropriate equation. Algorithms for calculating d' for forced-choice tasks can be found in J. E. K. Smith (1982).

More extensive programs are required to analyze data from rating tasks. Centor (1985) has published a spreadsheet macro for this purpose, but his program uses linear extrapolation and thus tends to underestimate A_z. (See Centor & Schwartz, 1985, for a discussion of this problem.) A better choice is RSCORE, written by Donald Dorfman. Source code is available in both FORTRAN (Dorfman, 1982) and BASIC (Alf & Grossberg, 1987). The latest version, RSCORE4, may be downloaded (from ftp://perception.radiology.uiowa.edu/public/rscore). A comparable program, Charles Metz's ROCFIT, may be downloaded (from ftp://random.bsd.uchicago.edu/roc). Both programs are available in source code (FORTRAN) and compiled form (PC-compatible for RSCORE4; PC-compatible,

Table 1
Commands Needed to Calculate *d′* in Various Computer Packages

Package	Command(s)
Excel	DPRIME = NORMSINV(H) − NORMSINV(F)
Mathematica	<<Statistics`NormalDistribution` DPRIME = Quantile[NormalDistribution[0,1],H] −Quantile[NormalDistribution[0,1],F]
Minitab	InvCDF 'H' c3; normal 0 1. InvCDF 'F' c4; normal 0 1. Name c5 = 'DPRIME' Let 'DPRIME' = c3 − c4
Quattro Pro	DPRIME = @NORMINV(H,0,1) − @NORMINV(F,0,1)
SAS	DPRIME = PROBIT(H) − PROBIT(F)
SPSS	COMPUTE DPRIME = PROBIT(H) − PROBIT(F)
SYSTAT	LET DPRIME = ZIF(H) − ZIF(F)

Note—The hit and false-alarm rates are called H and F, respectively, and *d′* is returned in DPRIME.

Macintosh, and Unix for ROCFIT). Both ftp sites also contain programs that can be used for other types of SDT analyses.

ROCFIT and RSCORE4 are both based on Dorfman's (1982) program. Thus, they usually provide similar results. However, RSCORE4 may perform better when hit or false-alarm rates equal 0 or 1, as it uses a sophisticated iterative method for dealing with these cases (Dorfman & Berbaum, 1995).

General Purpose Software

A' and B'' can be computed with virtually any spreadsheet or statistics package. For example, A' can be calculated in SPSS with the following COMPUTE statement:

$$\begin{aligned}\text{APRIME} = {} & 0.5 + (\text{ABS}(H - F) / (H - F)) \\ & * ((H - F)**2 + \text{ABS}(H - F)) \\ & /(4 * \text{MAX}(H,F) - 4 * H * F), \quad (12)\end{aligned}$$

where H is the variable containing the hit rate and F is the variable containing the false-alarm rate. SPSS lacks the sign function, so the statement $\text{ABS}(H - F)/(H - F)$ is used instead.

Extensions of Equation 12 to other software packages are straightforward but may require slight changes. For example, some packages square a value with the ^2 operator, rather than the **2 operator. Packages that lack a function for determining the maximum must use the two formulae found in Equation 2, rather than the single formula used in Equation 3.

Spreadsheets (and some statistical packages, such as Minitab) require cell identifiers, rather than variable names. This is illustrated by the following Excel statement, which can be used to calculate B'' from the hit rate (stored in cell a1) and the false-alarm rate (stored in cell a2):

$$\begin{aligned}= {} & \text{SIGN}(a1 - a2) * (a1 - a1 * a1 - a2 + a2 * a2) \\ & /(a1 - a1 * a1 + a2 - a2 * a2). \quad (13)\end{aligned}$$

Table 2
Commands Needed to Calculate β in Various Computer Packages

Package	Command(s)
Excel	BETA = EXP ((NORMSINV(F)^2 −NORMSINV(H)^2) / 2)
Mathematica	<<Statistics`NormalDistribution` BETA = Exp [(Quantile[NormalDistribution[0,1],F]^2 −Quantile[NormalDistribution[0,1],H]^2) / 2]
Minitab	InvCDF 'H' c3; normal 0 1. InvCDF 'F' c4; normal 0 1. Name c5 = 'BETA' Let 'BETA' = EXP ((c4**2 − c3**2) / 2)
Quattro Pro	BETA = @EXP ((@NORMINV(F,0,1)^2 −@NORMINV(H,0,1)^2) / 2)
SAS	BETA = EXP ((PROBIT(F)**2 − PROBIT(H)**2) / 2)
SPSS	COMPUTE BETA = EXP ((PROBIT(F)**2 − PROBIT(H)**2) / 2)
SYSTAT	LET BETA = EXP ((ZIF(F)^2 − ZIF(H)^2) / 2)

Note—The hit and false-alarm rates are called H and F, respectively, and β is returned in BETA.

Table 3
Commands Needed to Calculate *c* in Various Computer Packages

Package	Command(s)
Excel	C = −(NORMSINV(H) + NORMSINV(F)) / 2
Mathematica	<<Statistics`NormalDistribution C = −(Quantile[NormalDistribution[0,1],H] +Quantile[NormalDistribution[0,1],F]) / 2
Minitab	InvCDF 'H' c3; normal 0 1. InvCDF 'F' c4; normal 0 1. Name c5 = 'C' Let 'C' = −(c3 + c4) / 2
Quattro Pro	C = −(@NORMINV(H,0,1) + @NORMINV(F,0,1)) / 2
SAS	C = −(PROBIT(H) + PROBIT(F)) / 2
SPSS	COMPUTE C = −(PROBIT(H) + PROBIT(F)) / 2
SYSTAT	LET C = −(ZIF(H) + ZIF(F)) / 2

Note—The hit and false-alarm rates are called H and F, respectively, and *c* is returned in C.

Other sensitivity and response bias measures can only be calculated with software that provides direct access to the Φ and Φ^{-1} functions. All the statistical packages utilize these functions to perform significance tests, but only some allow the functions to be used while creating new variables. Even fewer spreadsheets provide such access; notable exceptions are Excel and Quattro Pro (but not Lotus 1-2-3). However, any spreadsheet that permits macro programming can use the Φ^{-1} algorithms reviewed by Brophy (1985).

The Φ^{-1} function is needed to calculate d', β, and c. This function is called NORMSINV or NORMINV in Excel and Quattro Pro, ANORIN in IMSL, normQuant in JMP, Quantile[NormalDistribution[0,1]] in Mathematica, InvCDF in Minitab, PROBIT in SAS and SPSS, and ZIF in SYSTAT. Users must sometimes specify $M = 0$ and $SD = 1$ (e.g., when using NORMINV, but not when using NORMSINV). Calculation of β requires exponentiation (called EXP in most packages) and squaring (using the **2 or ^2 operator).

The Φ function is needed to calculate $A_{d'}$. This function is called NORMSDIST or NORMDIST in Excel and Quattro Pro, ANORDF in IMSL, normDist in JMP, CDF[NormalDistribution[0,1]] in Mathematica, CDF in Minitab, PROBNORM in SAS, CDFNORM in SPSS, and ZCF in SYSTAT. Users must sometimes specify $M = 0$ and $SD = 1$ (e.g., when using NORMDIST, but not when using NORMSDIST). Calculation of $A_{d'}$ also requires taking a square root. Most packages accomplish this with the SQR or SQRT function.

Sample commands for determining d', β, c, and $A_{d'}$ in a variety of packages are listed in Tables 1–4. Commands for other packages should be readily derived from these examples (referring to the appropriate equation, as needed). Note that some packages (e.g., spreadsheets and Minitab) use column or cell identifiers in place of the variable names listed in Tables 1–4.

Researchers with rating task data may be tempted to use standard regression techniques (ordinary least-squares, or OLS) to fit a line to the *z* scores for the hit and false-alarm rates (e.g., Richards & Thornton, 1970). Once the slope and intercept of this line are known, Equation 10 can be used to determine A_z. However, this approach is problematic. OLS assumes that only one variable is measured with error; the best-fitting line minimizes the errors in predicting this variable. In rating tasks, the empirically determined hit and false-alarm rates both contain sampling error, so OLS provides biased estimates of the slope and intercept.

This problem is best solved by relying on a program designed specifically for the analysis of rating task data, such as ROCFIT or RSCORE4. However, another possibility is to perform two OLS regressions. In describing this

Table 4
Commands Needed to Calculate $A_{d'}$ in Various Computer Packages

Package	Command(s)
Excel	AD = NORMSDIST (DPRIME / SQRT(2))
Mathematica	<<Statistics`NormalDistribution` AD = CDF [NormalDistribution[0,1],DPRIME / Sqrt[2]]
Minitab	LET c2 = 'DPRIME' / SQRT(2) Name c3 = 'AD' CDF c2 'AD'
Quattro Pro	AD = @NORMDIST (DPRIME / SQRT(2), 0, 1, 1)
SAS	AD = PROBNORM (DPRIME / SQRT(2))
SPSS	COMPUTE AD = CDFNORM (DPRIME / SQRT(2))
SYSTAT	LET AD = ZCF (DPRIME / SQR(2))

Note—d' is stored in a variable called DPRIME, and $A_{d'}$ is returned in a variable called AD.

Table 5
Hypothetical Number of Responses in Signal and Noise Trials for a Study Using a 6-Point Rating Scale, and Hit and False-Alarm Rates if Numerically Higher Responses Are Considered to Be *Yes* Responses

Response	Signal Trials	Hit Rate	Noise Trials	False-Alarm Rate
1	0	.99*	8	.68
2	4	.92	6	.44
3	8	.76	1	.40
4	8	.60	3	.28
5	12	.36	7	.02†
6	18		0	

*The actual hit rate is 1.00; the entry shown assumes a rate of 49.5 ÷ 50. †The actual false-alarm rate is 0.00; the entry shown assumes a rate of 0.5 ÷ 25.

procedure, we call the z score corresponding to a given hit rate z_H, so that $\Phi^{-1}(\text{hit rate}) = z_H$. Similarly, we call the z score corresponding to a given false-alarm rate z_F. The double regression procedure assumes that the best-fitting "unbiased" line (the solid black line in Figure 3) is midway between the regression line that minimizes the errors in predicting z_H from z_F (the thin black line in Figure 3) and the regression line that minimizes the errors in predicting z_F from z_H (the broken line in Figure 3).

The procedure begins by using OLS to regress z_H on z_F (i.e., predict the hit rate z scores from the false-alarm rate z scores). Call the slope of the resulting regression line Slope_1. Next, regress z_F on z_H (i.e., predict the false-alarm rate z scores from the hit rate z scores). Call the slope of this regression line Slope_2. Average the two slopes to find the slope of the "unbiased" line, as follows:

$$\text{Slope}^* = 0.5\left(\text{Slope}_1 + \frac{1}{\text{Slope}_2}\right), \qquad (14)$$

where Slope_2 is inverted, so that both slopes indicate how a given change in z_F affects z_H. Next, find the mean of all $r - 1$ values of z_F (call this $\overline{z_F}$) and the mean of all $r - 1$ values of z_H ($\overline{z_H}$). Then, find the intercept of the best fitting "unbiased" line with the formula

$$\text{Intercept}^* = \overline{z_H} - (\text{Slope}^* \times \overline{z_F}). \qquad (15)$$

Equation 15 ensures that the best-fitting "unbiased" line predicts a value of $\overline{z_H}$ for z_H when $z_F = \overline{z_F}$. Finally, determine A_z by using the "unbiased" slope and intercept in Equation 10.

COMPUTATIONAL EXAMPLES

Readers may find it useful to work through a computational example that illustrates the application of the procedures described above. Our example presents data from a study in which a subject observes 50 signal trials and 25 noise trials. After each trial, the subject uses a 6-point scale to indicate whether or not a signal was presented. A response of 1 indicates that the subject is *very certain* a signal was *not* presented, whereas a response of 6 indicates the subject is *very certain* that a signal *was* presented. Intermediate values represent intermediate levels of certainty. Hypothetical data from a single subject are presented in Table 5.

The data must first be converted into hit and false-alarm rates. To do this, first consider 1 responses to be equivalent to *no* responses in a yes/no task, and consider the remaining responses to be equivalent to *yes* responses. This yields 17 false alarms, so (dividing by the total number of noise trials) the false-alarm rate is 17 ÷ 25 = .68. Similarly, there are 50 hits, so the hit rate is 50 ÷ 50 = 1.00. This hit rate is replaced by 49.5 ÷ 50 = .99 (see the section entitled "Hit and False-Alarm Rates of Zero or One"). Now consider responses of 1 or 2 to be *no* responses, and consider responses of 3, 4, 5, or 6 to be *yes* responses. This yields 11 false alarms (for a false-alarm rate of .44) and 46 hits (for a hit rate of .92). This procedure is continued until only responses of 6 are considered *yes* responses, which yields a false-alarm rate of 0 (which is replaced by 0.5 ÷ 25 = .02) and a hit rate of .36. The five pairs of hit and false-alarm rates that result from this procedure are listed in Table 5.

Table 6 lists z scores and the values of d', $A_{d'}$, A', β, c, and B'' for each of the five pairs of hit and false-alarm rates. To illustrate how these values are obtained, consider the second row in Table 5, which has a hit rate of .92. Application of the Φ^{-1} function reveals that the corresponding z score is 1.4051. Similarly, the false-alarm rate of .44 corresponds to a z score of −0.1510. Subtracting the false-alarm rate z score from the hit rate z score yields $d' = 1.56$ (Equation 1). Dividing this by $\sqrt{2}$ and applying the Φ function to the result reveals that $A_{d'} = .86$ (Equa-

Table 6
Signal Detection Parameters for the Data Shown in Table 5

Yes Response	$\Phi^{-1}(H)$	$\Phi^{-1}(F)$	d'	$A_{d'}$	A'	β	c	B''
2–6	2.3263	0.4677	1.86	.91	.82	0.07	−1.40	−0.91
3–6	1.4051	−0.1510	1.56	.86	.84	0.38	−0.63	−0.54
4–6	0.7063	−0.2533	0.96	.75	.75	0.80	−0.23	−0.14
5–6	0.2533	−0.5828	0.84	.72	.72	1.15	0.16	0.09
6	−0.3585	−2.0537	1.70	.88	.88	7.73	1.21	0.84

tion 11). Application of Equation 3 yields $A' = .84$. Subtracting the squared hit rate z score from the squared false-alarm rate z score and dividing the result by 2 yields -0.98, which is the natural logarithm of β (Equation 6). Raising e to this power reveals that $\beta = 0.38$ (Equation 5). The mean of the hit rate and the false-alarm rate z scores is 0.6271, so $c = -0.63$ (Equation 7). Finally, application of Equation 9 yields $B'' = -0.54$.

ROCFIT estimates that the ROC curve in z space has a slope of 1.28 and an intercept of 1.52. Application of Equation 10 yields a value of .82 for A_z. When OLS is used to estimate the slope and intercept, $\text{Slope}_1 = 0.99$, $\text{Slope}_2 = 0.81$, and Slope* = 1.11. The mean false-alarm rate z score is -0.51, whereas the mean hit rate z score is 0.87. Thus, Intercept* = 1.44, yielding a value of .83 for A_z.

This example illustrates the danger in using standard OLS to analyze rating data. The correlation between the hit rate and the false-alarm rate z scores is very high ($r = .90$). Even so, the slope obtained by minimizing errors in predicting z_H (0.99) differs markedly from the slope obtained by minimizing errors in predicting z_F (1.23, after inverting Slope_2). Furthermore, OLS gives the misleading impression that the signal and noise distributions have equal standard deviations ($\text{Slope}_1 \approx 1$); in fact, the noise distribution varies far more than the signal distribution. Thus, a program specifically designed to analyze rating data (such as ROCFIT or RSCORE4) should be used whenever possible.

The data in this example violate one of the d' assumptions, because the noise distribution has a far larger standard deviation than does the signal distribution.

This violation explains why the d' values in Table 6—each of which results from a different criterion—vary so widely. Clearly, researchers should not use d' without first obtaining evidence (preferably from a rating task) that its underlying assumptions are valid. Nonparametric measures are no panacea; the A' values are somewhat less variable than the $A_{d'}$ values but also lack stability. Thus, researchers who are primarily interested in sensitivity may wish to avoid yes/no tasks altogether and rely, instead, on forced-choice or rating tasks.

CONCLUSION

The primary contribution of SDT to psychology is the recognition that performance on discrimination tasks involves two separate factors: response bias and sensitivity. Several different measures have been developed to quantify both of these factors. Many of these measures can be determined in a computation-free manner through the use of tables, but a better approach involves using specialized or general purpose computer software. It is to be hoped that the ready availability of this software will encourage more researchers to apply SDT than currently seems to be the case.

REFERENCES

ACHENBACH, T. M., & EDELBROCK, C. S. (1983). *Manual for the child behavior checklist and revised child behavior profile*. Burlington: University of Vermont, Department of Psychiatry.

AHROON, W. A., JR., & PASTORE, R. E. (1977). Procedures for computing d' and β. *Behavior Research Methods & Instrumentation*, **9**, 533-537.

ALF, E. F., & GROSSBERG, J. M. (1987). DORF2R.BAS: Analyzing signal-detection theory rating data in the BASIC programming language. *Behavior Research Methods, Instruments, & Computers*, **19**, 475-482.

BANKS, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, **74**, 81-99.

BROPHY, A. L. (1985). Approximation of the inverse normal distribution function. *Behavior Research Methods, Instruments, & Computers*, **17**, 415-417.

BROPHY, A. L. (1986). Alternatives to a table of criterion values in signal detection theory. *Behavior Research Methods, Instruments, & Computers*, **18**, 285-286.

CENTOR, R. M. (1985). A Visicalc program for estimating the area under a receiver operating characteristic (ROC) curve. *Medical Decision Making*, **5**, 139-148.

CENTOR, R. M., & SCHWARTZ, J. S. (1985). An evaluation of methods for estimating the area under a receiver operating characteristic (ROC) curve. *Medical Decision Making*, **5**, 149-156.

CRADIT, J. D., TASHCHIAN, A., & HOFACKER, C. F. (1994). Signal detection theory and single observation designs: Methods and indices for advertising recognition testing. *Journal of Marketing Research*, **31**, 117-127.

CRAIG, A. (1979). Nonparametric measures of sensory efficiency for sustained monitoring tasks. *Human Factors*, **21**, 69-78.

CRAIG, A. (1985). Field studies of human inspection: The application of vigilance research. In S. Folkard & T. H. Monk (Eds.), *Hours of work* (pp. 133-145). Chichester, U.K.: Wiley.

CRAVEN, B. J. (1992). A table of d' for M-alternative odd-man-out forced-choice procedures. *Perception & Psychophysics*, **51**, 379-385.

DAVIES, D. R., & PARASURAMAN, R. (1982). *The psychology of vigilance*. London: Academic Press.

DONALDSON, W. (1993). Accuracy of d' and A' as estimates of sensitivity. *Bulletin of the Psychonomic Society*, **31**, 271-274.

DORFMAN, D. D. (1982). RSCORE II. In J. A. Swets & R. M. Pickett, *Evaluation of diagnostic systems: Methods from signal detection theory* (pp. 212-232). New York: Academic Press.

DORFMAN, D. D., & BERBAUM, K. S. (1995). Degeneracy and discrete receiver operating characteristic rating data. *Academic Radiology*, **2**, 907-915.

ELLIOTT, P. B. (1964). Tables of d'. In J. A. Swets (Ed.), *Signal detection and recognition by human observers* (pp. 651-684). New York: Wiley.

ERDELYI, M. H. (1974). A new look at the New Look: Perceptual defense and vigilance. *Psychological Review*, **81**, 1-25.

FREEMAN, P. R. (1973). *Tables of d' and β*. Cambridge: Cambridge University Press.

GARDNER, R. M., & BOICE, R. (1986). A computer program to generate signal-detection theory values for sensitivity and response bias. *Behavior Research Methods, Instruments, & Computers*, **18**, 54-56.

GARDNER, R. M., DALSING, S., REYES, B., & BRAKE, S. (1984). Table of criterion values (β) used in signal detection theory. *Behavior Research Methods, Instruments, & Computers*, **16**, 425-436.

GARLAND, H. L. (1949). On the scientific evaluation of diagnostic procedures. *Radiology*, **52**, 309-328.

GOLDSTEIN, E. B. (1996). *Sensation and perception* (4th ed.). New York: Brooks Cole.

GREEN, D. M., & MOSES, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin*, **66**, 228-234.

GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

GRIER, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, **75**, 424-429.

HACKER, M. J., & RATCLIFF, R. (1979). A revised table of d' for m-alternative forced choice. *Perception & Psychophysics*, **26**, 168-170.

HAUTUS, M. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, **27**, 46-51.

HERSHMAN, R. L., & SMALL, D. (1968). Tables of d' for detection and localization. *Perception & Psychophysics*, **3**, 321-323.

HODOS, W. (1970). Nonparametric index of response bias for use in de-

tection and recognition experiments. *Psychological Bulletin*, **74**, 351-354.

HUTCHINSON, T. P. (1981). A review of some unusual applications of signal detection theory. *Quality & Quantity*, **15**, 71-98.

INGHAM, J. G. (1970). Individual differences in signal detection. *Acta Psychologica*, **34**, 39-50.

KAPLAN, H. L., MACMILLAN, N. A., & CREELMAN, C. D. (1978). Tables of *d'* for variable-standard discrimination paradigms. *Behavior Research Methods & Instrumentation*, **10**, 796-813.

KLATZKY, R. L., & ERDELYI, M. H. (1985). The response criterion problem in tests of hypnosis and memory. *International Journal of Clinical & Experimental Hypnosis*, **33**, 246-257.

MACMILLAN, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 21-57). Hillsdale, NJ: Erlbaum.

MACMILLAN, N. A., & CREELMAN, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, **107**, 401-413.

MACMILLAN, N. A., & CREELMAN, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

MACMILLAN, N. A., & CREELMAN, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and bias. *Psychonomic Bulletin & Review*, **3**, 164-170.

MACMILLAN, N. A., & KAPLAN, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, **98**, 185-199.

MCGOWAN, W. T., III, & APPEL, J. B. (1977). An APL program for calculating signal detection indices. *Behavior Research Methods & Instrumentation*, **9**, 517-521.

MCNICOL, D. (1972). *A primer of signal detection theory*. London: Allen & Unwin.

MILLER, J. (1996). The sampling distribution of *d'*. *Perception & Psychophysics*, **58**, 65-72.

NAYLOR, J. C., & LAWSHE, C. H. (1958). An analytical review of the experimental basis of subception. *Journal of Psychology*, **46**, 75-96.

NELSON, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, **95**, 109-133.

NELSON, T. O. (1986). ROC curves and measures of discrimination accuracy: A reply to Swets. *Psychological Bulletin*, **100**, 128-132.

POLLACK, I., & NORMAN, D. A. (1964). A nonparametric analysis of recognition experiments. *Psychonomic Science*, **1**, 125-126.

REY, J. M., MORRIS-YATES, A., & STANISLAW, H. (1992). Measuring the accuracy of diagnostic tests using receiver operating characteristic (ROC) analysis. *International Journal of Methods in Psychiatric Research*, **2**, 39-50.

RICHARDS, B. L., & THORNTON, C. L. (1970). Quantitative methods of calculating the *d'* of signal detection theory. *Educational & Psychological Measurement*, **30**, 855-859.

RICHARDSON, J. T. E. (1994). Continuous recognition memory tests: Are the assumptions of the theory of signal detection really met? *Journal of Clinical & Experimental Neuropsychology*, **16**, 482-486.

SMITH, J. E. K. (1982). Simple algorithms for M-alternative forced-choice calculations. *Perception & Psychophysics*, **31**, 95-96.

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

SNODGRASS, J. G., & CORWIN, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.

SNODGRASS, J. G., LEVY-BERGER, G., & HAYDON, M. (1985). *Human experimental psychology*. New York: Oxford University Press.

STANISLAW, H. (1995). Effect of type of task and number of inspectors on performance of an industrial inspection-type task. *Human Factors*, **37**, 182-192.

STANISLAW, H. (1996). Relative versus absolute judgments, and the effect of batch composition on simulated industrial inspection. In R. J. Koubek & W. Karwowski (Eds.), *Manufacturing agility and hybrid automation: I* (pp. 105-108). Louisville, KY: IEA Press.

STANISLAW, H., & TODOROV, N. (1992). Documenting the rise and fall of a psychological theory: Whatever happened to signal detection theory? (Abstract). *Australian Journal of Psychology*, **44**, 128.

SWETS, J. A. (1973). The relative operating characteristic in psychology. *Science*, **182**, 990-1000.

SWETS, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, **99**, 181-198.

SWETS, J. A. (1988a). Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-1293.

SWETS, J. A. (ED.) (1988b). *Signal detection and recognition by human observers*. New York: Wiley.

SWETS, J. A., & PICKETT, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.

THOMAS, E. A. C., & HOGUE, A. (1976). Apparent weight of evidence, decision criteria, and confidence ratings in juror decision making. *Psychological Review*, **83**, 442-465.

ZELEN, M., & SEVERO, N. C. (1972). Probability functions. In M. Abramowitz & I. A. Stegun (Eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (pp. 925-973). Washington, DC: National Bureau of Standards.

(Manuscript received August 15, 1997;
revision accepted for publication February 9, 1998.)