ANALYSIS OF DIAGNOSTIC TASKS IN ACCOUNTING RESEARCH

Robert J. Ramsay

Arthur Andersen Professor of Accounting
Von Allmen School of Accountancy
Gatton College of Business and Economics
University of Kentucky
Lexington, KY 40506

Richard M. Tubbs*

Department of Accounting
Henry B. Tippie College of Business
University of Iowa
Iowa City, IA 52242

September 5, 2003

This paper is dedicated to Donald Dorfman (1933-2001). We appreciate the helpful comments of Mike Bamber, Bruce Johnson, Kathryn Kadous, Mark Peecher, Sean Peffer, Mort Pincus, Jay Rich, Dan Stone, Scott Vandervelde, and the participants at the University of Iowa and University of Kentucky Accounting Research Workshops and the University of Iowa JDM Seminar. We also thank Joe Carcello and Terry Neal for sharing data with us.

* Corresponding author. Telephone: 319-335-0848; fax: 319-335-1956; e-mail: richard-tubbs@uiowa.edu.
ANALYSIS OF DIAGNOSTIC TASKS IN ACCOUNTING RESEARCH

Abstract

Many accounting judgments are diagnostic tasks in which accountants, auditors, managers, or investors discriminate among possible states and decide which one exists. To measure the accuracy of such decisions, most accounting research employs percentage correct, a measure proven to be invalid and unreliable. This paper describes Signal Detection Theory (SDT), a theoretical model of diagnostic tasks that has been empirically supported in many fields, and discusses advantages of employing SDT in accounting research. The paper also describes an SDT-based reanalysis of data related to two published accounting studies which results in revised conclusions and important insights.

Keywords Diagnostic tasks; Signal Detection Theory; Accuracy; Response bias; Confidence.
ANALYSIS OF DIAGNOSTIC TASKS IN ACCOUNTING RESEARCH

INTRODUCTION

Many accounting-related judgments (e.g., bankruptcy prediction, stock buy-sell recommendations, going concern, error or fraud detection) can be characterized as diagnostic tasks, tasks whose purpose is to discriminate among possible states of the object under study, and to decide which one actually exists (Swets and Pickett 1982, ix). Since diagnostic tasks are prevalent in accounting, it is not surprising that a great deal of accounting research evaluates the performance of a decision maker or a diagnostic system (e.g., analytical tool) in such tasks.¹ This paper has three primary purposes. The first purpose is to demonstrate that Signal Detection Theory (SDT) concepts, methods, and measures (which have been underutilized by accounting researchers) offer many advantages when examining diagnostic tasks. The second purpose is to establish that commonly used non-SDT measures of accuracy can result in imprecise and even misleading findings. The third purpose is to provide adequate technical detail about SDT so that accounting researchers who examine diagnostic tasks can more easily employ these concepts, methods, and measures in their research.

The outcomes in the simplest form of a diagnostic task can be captured in a 2 x 2 table that compares a subject’s (or a system’s) responses to the actual state as follows:

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Response</th>
<th>Signal</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>hits</td>
<td>misses</td>
<td></td>
</tr>
<tr>
<td>Noise</td>
<td></td>
<td>false alarms</td>
<td>correct rejections</td>
<td></td>
</tr>
</tbody>
</table>

in which the presentation of the stimulus (i.e., signal or noise) is captured in the rows and the subject’s responses (i.e., yes or no) are captured in the columns. In the table “hits” represents the number of times the signal is present and the subject correctly indicates that the signal is present, “misses” indicates the number of times the signal is present and the subject incorrectly reports that the signal is absent, “false alarms” represents the number of times the signal is absent but the subject incorrectly responds that the signal is present, and “correct rejections” represents the number of times the signal is absent and the subject correctly responds that the signal is absent.

SDT provides a theoretical model of the diagnostic task that has been empirically supported in both theoretical and practical fields (Swets 1986a). SDT recognizes that the diagnostic task involves two independent cognitive processes: discrimination and decision. SDT decomposes the outcomes (i.e., hits, misses, false alarms,
correct rejections) of the diagnostic task “so that the discrimination and decision processes can be evaluated independently” (Swets 1996, xiv). Valid methods of examining diagnostic tasks have been developed as part of SDT. In addition, valid and reliable measures of both the discrimination process (often referred to as accuracy measures) and the decision process (often referred to as response bias measures) have been developed. Accuracy is a measure of the capability of the subject in discriminating signal from noise, while response bias is a measure of the effect of decision factors (e.g., prior probability of signal, costs and benefits associated with different outcomes) on the subject responding “yes” or “no.” These methods and measures are widely adopted for the study of diagnostic tasks in fields like information retrieval, weather forecasting, medical diagnosis, recognition memory, aptitude testing, polygraph lie detection, and vigilance (in which operators are required to detect unpredictable and infrequent signals over a long period of time).  

Unfortunately, the vast majority of accounting research in which diagnostic tasks are examined fails to employ the concepts, methods and measures of SDT. For example, most of these studies employ a non-SDT measure of accuracy, percentage correct ((hits + correct rejections) / (hits + misses + false alarms + correct rejections)) (e.g., McDaniel 1990; Liang et al. 1992; Nelson 1993; Ramsay 1994; Simnett 1996; Braun 2000; Knapp and Knapp 2001; Jamal and Tan 2001). While non-SDT measures of accuracy may be high in “face validity” (i.e., they appear to measure what they are intended to measure, see Nunnally 1978, 111), percentage correct \( P(C) \) is an “inadequate index of accuracy” (Swets and Pickett 1982, 24) because it does not separate accuracy from response bias. For instance, percentage correct can vary from \( 1 - P(S) \) to \( P(S) \) (where \( P(S) \) is the probability of signal) because of response bias alone with no accuracy (Swets 1986b, 102). Furthermore, non-SDT measures of accuracy like \( P(C) \) imply theoretical models of the diagnostic task that have not been empirically supported (Macmillan and Kaplan 1985; Swets 1986a). Moreover, these studies largely ignore response biases, although understanding the underlying causes of response biases could be critical for understanding the accountants’ judgments.

Increased use of SDT concepts, methods and measures in accounting research offers many advantages. First, SDT provides theory-based measures of accuracy that are not contaminated by response bias. Second, SDT provides theory-based measures of response bias. An intuitively appealing measure of confidence can also be calculated from one particular SDT measure of response bias. Third, since measures of accuracy and response bias can be calculated contemporaneously, one can determine whether a manipulation affects accuracy,
response bias, or both. For example, several presumed accuracy effects in the psychology literature were overturned and found to actually be response bias effects once SDT methods and measures were applied (see Swets 1973). Fourth, in certain tasks, optimal accuracy and optimal response bias can be calculated. These calculations can then be compared with actual accuracy and response bias. Finally, the validity of one’s research is likely to be enhanced.

The remainder of this paper is organized as follows. The second section presents detailed guidance in employing SDT theoretical concepts, empirical methods, and empirical measures. The third section discusses the advantages associated with employing SDT concepts, methods, and measures in research. The fourth section describes the experimental studies that employ SDT concepts, methods, and measures that have appeared in accounting journals. The fifth section reanalyzes data related to two published accounting studies, Ramsay (1994) and Carcello and Neal (2000). The reanalysis of experimental data related to Ramsay (1994) illustrates that empirical results (and associated conclusions) can differ depending on whether SDT or non-SDT measures are employed. We also illustrate a potential difficulty with SDT analysis and the preferred solution to that difficulty. The reanalysis of archival data related to Carcello and Neal (2000) illustrates that SDT procedures and measures can be applied to archival data and that important additional insights can be gained. Concluding comments are provided in the final section.

**SIGNAL DETECTION THEORY**

**Theoretical Concepts**

SDT was developed in electrical engineering and based upon statistical decision theory. It is applicable to tasks in which an observer is presented with one of two discrete but not easily distinguishable stimulus categories (e.g., signal and noise), and the observer must choose between responses that indicate which stimulus category has been presented (e.g., yes and no). SDT recognizes that tasks of this nature involve two independent cognitive processes, discrimination and decision. The discrimination process involves the observer’s assessment of the degree to which the evidence in the observation favors signal. The decision process involves the observer’s determination of how strong the evidence favoring signal must be before responding yes (Swets 1996).

The concepts of SDT are illustrated in Figure 1. The ordinate represents the magnitude of the sensory observation (e.g., the familiarity of the stimulus in a recognition task). As illustrated in Figure 1, this sensory observation (X) is assumed to vary continuously on a single dimension (Swets, Tanner, and Birdsall 1961).
abscissa is the probability density function representing the observer’s assessment of the probability of the state given the magnitude of the sensory observation. The left-hand distribution represents the observer’s probability density function for state \( S_1 \) (i.e., noise). The right hand distribution is the probability density function for state \( S_2 \) (i.e., signal). It is assumed that any value of \( X \) can arise from either the \( S_1 \) or \( S_2 \) distributions (Swets, Tanner, and Birdsall 1961). Point \( X_\circ \) is the decision criterion at which the observer begins to respond “yes” (i.e., state that the stimulus is \( S_2 \)). Stimuli to the left of \( X_\circ \) are reported “no” (i.e., state that the stimulus is \( S_1 \)). The area under the \( S_2 \) distribution to the right of \( X_\circ \) (vertical lines) equals the probability of the subject stating “yes” when \( S_2 \) exists (\( P(\text{yes}|S_2) \), or the hit rate (H)); whereas, the area of \( S_1 \) to the right of \( X_\circ \) (horizontal lines) represents the probability of stating “yes” when \( S_2 \) does not exist (\( P(\text{yes}|S_1) \), or the false alarm rate (F)).

The degree of overlap of the \( S_1 \) and \( S_2 \) distributions on the sensory observation axis represents the observer’s ability to discriminate between stimuli that represent noise and stimuli that represent signal (i.e., accuracy). For example, if the noise and signal distributions overlap greatly, then the observer has low accuracy. In contrast, if the noise and signal distributions overlap little, then the observer has high accuracy.

SDT is considered a “variable-criterion” model of the diagnostic task because the observer is assumed to be able to set the decision criterion at any point along the X-axis (Swets, Tanner, and Birdsall 1961). The location at which the observer places the decision criterion, \( X_\circ \), represents the observer’s bias toward responding “yes” or “no.” An observer who is neutral toward responding “yes” or “no” would place the criterion at the point where \( F \) and \( 1-H \) are equal (in Figure 1 where the \( S_1 \) and \( S_2 \) distributions intersect). An observer who is conservative about responding “yes” would place the criterion to the right of this point. Such a placement is referred to as “conservative response bias” or a “strict criterion.” An observer who is liberal about responding “yes” would place the criterion to the left of this point. Such a placement is referred to as “liberal response bias” or a “lenient criterion.”

The observer’s response bias is assumed to be a function of (1) the prior probabilities of the stimulus being noise and signal and (2) the benefits and costs associated with the various decision outcomes (Swets and Pickett 1982). These benefits and costs include the benefit of a correct rejection (\( \text{no}|S_1 \)), the cost of a false alarm, the benefit of a hit, and the cost of a miss (\( \text{no}|S_2 \)). The optimal value of the decision criterion expressed as a likelihood ratio (i.e., \( f(x|S_2)/f(x|S_1) \)) can be calculated as:

\[
\beta_{\text{opt}} = \frac{P(S_1)}{P(S_2)} \times \frac{\text{benefit (no } | S_1 \) }{\text{benefit (yes } | S_1 \) } - \frac{\text{cost (yes } | S_1 \) }{\text{cost (no } | S_1 \) } .
\] (1)
The lower the $\beta$, the farther to the left the decision criterion is placed, and the greater the probability of reporting “yes.”

Movement of the criterion ($X_k$) along the X-axis in figure 1 (i.e., changing the observer’s response bias) yields a series of different (H, F) pairs. However, each (H, F) pair in the series represents a constant degree of accuracy (i.e., the degree of overlap of $S_1$ and $S_2$ is unaltered). The plot of this series of (H, F) pairs with constant accuracy is represented in SDT with a relative-operating-characteristic (ROC) curve. As illustrated in figure 2, F is plotted on the horizontal axis, and H is plotted on the vertical axis of the ROC graph. Since H and F range from 0 to 1, the ROC graph is a unit square. An ROC curve along the major diagonal (i.e., where H=F) represents chance performance. ROC curves shift toward the upper left-hand corner (i.e., where H=1 and F=0) as performance increases. The minor diagonal of the ROC graph (dashed line in Figure 2) represents zero bias (i.e., $F=1-H$). On any ROC curve, points above the minor diagonal represent bias toward responding “yes” (lenient criterion or liberal response bias), while points below the minor diagonal represent a bias toward responding “no” (strict criterion or conservative response bias).

Figures 1 and 2 demonstrate that the relative frequency of hits, misses, false alarms, and correct rejections is a function of both the observer’s accuracy and where the observer places the criterion. For example, a more lenient criterion in figures 1 and 2 would, ceteris paribus, increase the relative frequency of hits but would also increase the relative frequency of false alarms. Accordingly, SDT separates performance into two components – accuracy (the observer’s capability in discriminating between stimuli) and decision criterion (the observer’s bias to report “yes” or “no”). A principal benefit of SDT is that accuracy and decision criterion are independent constructs, and independent measures of these constructs can be calculated (Swets 1996).

**Empirical Procedures and Measures**

**Yes-no Experiments**

In yes-no experiments, a single stimulus drawn from one of two stimulus categories (i.e., signal or noise) is presented on each trial within an experimental session, and the observer must respond from which category the stimulus is drawn (i.e., yes or no). Some proportion of the trials is signal, and the remaining proportion is noise. The observer is typically informed at the beginning of the session what the proportion of signal and noise trials will be and what the values and costs associated with the various decision outcomes will be.
The results of a yes-no experiment can be captured by the 2 x 2 matrix of possible outcomes. For example, the frequencies by response and stimulus categories for a subject follow (example taken from Sprinkle and Tubbs 1998).

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td>8</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Noise</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

The frequencies in each response category are converted to proportions of the total number of signal and noise trials as follows:

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td>.80</td>
<td>.20</td>
<td>1.00</td>
</tr>
<tr>
<td>Noise</td>
<td>.50</td>
<td>.50</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Therefore, in this example the H equals .80, and the F equals .50.

**Empirical Measures Associated with Yes-no Experiments**

The measure of accuracy that is traditionally employed with yes-no experiments is \( d' \). It is considered a distance measure of accuracy because \( d' \) theoretically represents the distance between the means of the signal and noise distributions \( (\mu_2 - \mu_1) \) divided by the standard deviation of the noise distribution \( (\sigma_1) \) (see Figure 1 and Table 1).

\[
\text{(Please insert Figure 1 and Table 1 here)}
\]

The empirical surrogate for \( d' \) is calculated as:

\[
d' = Z(H) - Z(F).^9
\]

The accuracy measure \( d' \) can range from negative to positive infinity. When \( H \) and \( F \) are equal, \( d' \) has a value of 0 which represents chance performance. For the example subject, substituting \( H \) and \( F \) into equation 2 yields a \( d' \) equal to 0.8416.

Two different measures of response bias, \( C \) and \( \beta \), are sometimes employed with yes-no experiments. The first measure, \( C \), is considered a criterion-location measure of response bias. Bias measure \( C \) theoretically
represents the distance between the criterion \( (X_k) \) and the zero-bias point where the signal and noise distributions cross \((0.5[\mu_2 + \mu_1])\) divided by the standard deviation of the noise distribution \((\sigma_1)\) (see Table 1). The empirical surrogate for \( C \) is calculated as:

\[
C = -0.5[Z(H) + Z(F)].
\]  

(3)

Bias measure \( C \) can range from negative to positive infinity. When \( F \) and \( 1-H \) are equal, \( C \) has a value of 0, which represents neutral bias. Positive values of \( C \) represent conservative bias, while negative values of \( C \) represent liberal bias. This measure has the attractive features that (1) \( C \) is statistically independent of \( d' \) and (2) the range of \( C \) does not depend upon \( d' \) (see Macmillan and Creelman 1991). For the example subject, substituting \( H \) and \( F \) into equation 3 yields a \( C \) equal to \(-0.4208\), a weak liberal bias.

The second measure, \( \beta \), is considered a likelihood-ratio measure of response bias. Bias measure \( \beta \) theoretically represents the relative likelihood of the signal distribution versus the relative likelihood of the noise distribution at the criterion. That is, \( \beta \) represents the ordinate of the signal distribution at criterion \( (\phi(x_k|S_2)) \) divided by the ordinate of the noise distribution at criterion \( (\phi(x_k|S_1)) \) (see Table 1). The empirical surrogate for \( \beta \) is calculated as:

\[
\beta = \exp (-0.5[Z(H)^2 - Z(F)^2]).
\]  

(4)

\( \beta \) can range from 0 to positive infinity. When \( F \) and \( 1-H \) are equal, \( \beta \) has a value of 1.0, which represents neutral bias. High values of \( \beta \) (i.e., \( \beta>1 \)) represent conservative bias, while low values of \( \beta \) (i.e., \( 0<\beta<1 \)) represent liberal bias. Unlike \( C \), bias measure \( \beta \) has the unattractive features that (1) \( \beta \) is not statistically independent of \( d' \) and (2) the range of \( \beta \) does depend upon \( d' \) (see Macmillan and Creelman 1991). However, it has the advantage that optimal decision criterion can be specified by \( \beta \) (Swets 1996). For the example subject, substituting \( H \) and \( F \) into equation 4 yields a \( \beta \) equal to 0.7018, a weak liberal bias.

**Confidence Rating Experiments**

The only difference in the execution of confidence rating experiments in comparison to yes-no experiments is the set of responses available to the observer. The set of responses vary from great confidence that one stimulus category has been presented (e.g., definitely, or almost definitely, yes), through relative indifference between the alternatives (e.g., possibly yes), to great confidence that the other stimulus category has been presented (e.g., definitely, or almost definitely, no).
The results of a confidence-rating experiment can be captured by the matrix of possible outcomes. For example, the frequencies by response and stimulus categories for a subject in an experiment with five rating categories follow (example taken from Sprinkle and Tubbs 1998). The rating categories are: (1) definitely, or almost definitely, yes, (2) probably yes, (3) possibly yes, (4) probably no, and (5) definitely, or almost definitely, no.

<table>
<thead>
<tr>
<th>Rating Category</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Noise</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

First, the frequencies in each rating category are converted to proportions of the total number of signal and noise trials.

<table>
<thead>
<tr>
<th>Rating Category</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td>.10</td>
<td>.10</td>
<td>.20</td>
<td>.30</td>
<td>.30</td>
<td>1.00</td>
</tr>
<tr>
<td>Noise</td>
<td>.20</td>
<td>.30</td>
<td>.10</td>
<td>.30</td>
<td>.10</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Second, the proportions by category and signal and noise distribution are cumulated from right to left. This is done to obtain the H and F for each decision criterion the subject can adopt (1 between each category). For example, if the subject were to adopt the criterion between categories 4 and 3, all of the assumed yes responses in categories 3, 2, and 1 for signal trials would be hits (i.e., H=.80). Similarly, all of the assumed yes responses in categories 3, 2, and 1 for noise trials would be false alarms (i.e., F=.50).

<table>
<thead>
<tr>
<th>Rating Category</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal (Hs)</td>
<td>1.00</td>
<td>.90</td>
<td>.80</td>
<td>.60</td>
<td>.30</td>
</tr>
<tr>
<td>Noise (Fs)</td>
<td>1.00</td>
<td>.80</td>
<td>.50</td>
<td>.40</td>
<td>.10</td>
</tr>
</tbody>
</table>

Columns 1-4 of the previous table represent the H (top row), F (bottom row) pairs at four different decision criteria (with column 1 representing the subject’s most strict criterion and column 4 representing the subject’s most lenient criterion). These empirical points (i.e., (H, F) pairs) serve as the basis for generating an ROC curve. A maximum-likelihood estimation, curve-fitting procedure (e.g., Dorfman and Alf 1969) can be employed to generate estimates of SDT parameters assuming an underlying bivariate normal model. These estimates include A.
(estimate of \( \frac{\mu_2 - \mu_1}{\sigma_2} \)), B (estimate of \( \frac{\sigma_1}{\sigma_2} \)), and the estimated (theoretical) points at each of the criteria on the best-fitting ROC curve.\(^{13}\) For this particular subject, A equals .6086, and B equals .9012. The empirical points and the ROC curve generated are graphed in Figure 2. The estimated (i.e., theoretical) points for the subject at each decision criterion are:

<table>
<thead>
<tr>
<th>Criterion</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit Rate (H)</td>
<td>.9099</td>
<td>.7576</td>
<td>.6237</td>
<td>.2982</td>
</tr>
<tr>
<td>False Alarm Rate (F)</td>
<td>.7916</td>
<td>.5398</td>
<td>.3724</td>
<td>.1033</td>
</tr>
</tbody>
</table>

**Empirical Measures Associated with Confidence Rating Experiments\(^{14}\)**

Two closely related measures of accuracy are frequently employed with confidence rating experiments. The first of these measures is \( d_a \), which is analogous to \( d' \). Like \( d' \), \( d_a \) is considered a distance measure of accuracy because it theoretically represents the distance between the means of the signal and noise distributions \((\mu_2 - \mu_1)\) in units of the root-mean-square standard deviation \((0.5(\sigma_2^2 + \sigma_1^2))^{0.5}\) (see Figure 1 and Table 1).

The empirical surrogate for \( d_a \) is calculated as:

\[
d_a = \frac{\sqrt{A}}{(1 + B^2)^{0.5}}\]  \(^{15}\)

The accuracy measure \( d_a \) can range from negative to positive infinity. A value of 0 represents chance performance. For the example subject, substituting the estimates of A and B into equation 5 yields a \( d_a \) equal to .6394.\(^{16}\)

The second measure \( A_z \) is the most widely accepted (e.g., Swets 1986a, 196) measure of accuracy that is employed with confidence rating experiments. \( A_z \) theoretically represents the probability transform of the distance between the means of the signal and noise distribution \((\mu_2 - \mu_1)\) divided by the square root of the sum of the variances of the signal and noise distribution \((\sigma_2^2 + \sigma_1^2)^{0.5}\) (see Figure 1 and Table 1). \( A_z \) is considered an area measure of accuracy because this probability transform is equivalent to the proportion of the area of the ROC graph that lies under the ROC curve (see Figure 2).

The empirical surrogate for \( A_z \) is calculated as:

\[
A_z = \Phi\left(\frac{A}{(1 + B^2)^{0.5}}\right)
\]  \(^{6}\)
where $\Phi(\bullet)$ is the probability transform of the argument and $A$ and $B$ are as defined above.\textsuperscript{17, 18} Accuracy measure $A_z$ can range from 0 to 1. A value of .5 represents chance performance, and a value of 1 represents perfect performance. For the example subject, substituting the estimates of $A$ and $B$ into equation 6 yields an $A_z$ equal to .6744.

The criterion-location measure of response bias that is often employed with the confidence rating procedure is $C_a$. $C_a$ theoretically represents the distance of the criterion ($X_i$) from the zero-bias point

$$\left(\frac{\sigma_1\mu_2 + \sigma_2\mu_1}{\sigma_1 + \sigma_2}\right)$$

in units of the root-mean-square standard deviation \(\left(0.5\left(\sigma_2^2 + \sigma_1^2\right)\right)^{0.5}\) (see Table 1). The empirical surrogate for $C_a$ is calculated as:

$$C_a = -\frac{\sqrt{2}B}{(1+B^2)^{0.5}(1+B)}[Z(H_k)+Z(F_k)].\textsuperscript{19}$$

(7)

Bias measure $C_a$ can range from negative to positive infinity. When $F$ and $1-H$ are equal, $C_a$ has a value of 0, which represents neutral bias. Positive values of $C_a$ represent conservative bias, while negative values of $C_a$ represent liberal bias. Notice that formula 7 reduces to $C_a=-.5[Z(H_a)+Z(F_a)]$ if the standard deviations of the noise and signal distributions are equal (i.e., $B=1$). Substituting the estimate of $B$ and the theoretical ($H$, $F$) pairs into equation 6 yields the following values of $C_a$ at each estimated point for the subject:

<table>
<thead>
<tr>
<th>Criterion</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_a$</td>
<td>-1.0717</td>
<td>-0.3977</td>
<td>0.0051</td>
<td>0.8927</td>
</tr>
</tbody>
</table>

**Empirical Support for SDT**

Swets (1986a) reviews diagnostic task research in the experimental psychology fields of perception, learning, memory, and cognition as well as in the practical fields of medical imaging, information retrieval, weather forecasting, aptitude testing, and polygraph lie detection. Empirical ROC curves from this research are extremely consistent in form. $H$ is a monotonically increasing function of $F$, and the ROC curves are fitted well by a straight line on $z$-coordinates. However, they are not always (1) convex and symmetric about the minor diagonal on linear coordinates and (2) of unit slope on $z$-coordinates. These results support the validity of a continuous, variable-criterion model of the diagnostic task that incorporates normally distributed signal and noise distributions of possibly unequal variance (Swets 1986a, 196). Therefore, methods in which one obtains a sufficient number of data
points (i.e., \((H, F)\) pairs) to generate an ROC curve (e.g., confidence rating experiments) are to be preferred over methods in which one obtains only one data point (e.g., yes-no experiments) (Macmillan and Creelman 1990, 83). Furthermore, measures that imply a continuous, variable-criterion model with normal distributions that can have unequal variance (e.g., \(d_a\), \(A_z\) and \(C_a\)) are to be preferred (see footnotes 15, 18 and 19). Measures which imply a continuous, variable-criterion model with normal distributions with equal variance (e.g., \(d', \ C\) and \(\beta\)) are less preferred but are the best measures available for yes-no experiments (see footnotes 9, 10 and 11) (Swets 1986b). Measures that do not imply a continuous, variable-criterion model (e.g., \(P(C)\)) are least preferred.\(^{20}\)

**ADVANTAGES ASSOCIATED WITH EMPLOYING SDT CONCEPTS, METHODS AND MEASURES**

**Superior Measure of Accuracy**

The first advantage associated with employing SDT is that superior measures of accuracy are available. SDT accuracy measures have many characteristics associated with a good accuracy measure. First, Macmillan and Creelman (1990, 1991) state that a good accuracy index should monotonically increase with \(H\), monotonically decrease with \(F\), and treat \(H\) and \(F\) symmetrically. Symbolically,

\[
\text{accuracy} = v[u(H) - u(F)] \quad (8)
\]

where \(v\) and \(u\) are both monotonic functions. SDT accuracy measure \(d'\) possesses these intuitively appealing qualities.\(^{21}\) Non-SDT measures like the percentage correct, \(P(C)\), (i.e., in terms of \(H\) and \(F\), \(P(C)=P(S_2)H + (1-P(S_2))(1-F)\)) do not (Swets and Pickett 1982).\(^{22}\)

Second, and even more important, a good accuracy index should be invariant to changes in factors other than accuracy (e.g., Swets and Pickett 1982; Macmillan and Creelman 1991). SDT measures of accuracy possess this attribute. The following three examples illustrate that this is not the case with a non-SDT measure like \(P(C)\).

**Example 1:** \(P(C)\) is dependent on \(P(S_2)\), and \(P(C)\) can be as high as \(P(S_2)\) or \((1 - P(S_2))\) by chance (i.e., without discrimination) (Swets 1986b). For example, suppose observers A and B have the following outcomes on a diagnostic task:

<table>
<thead>
<tr>
<th></th>
<th>Observer A</th>
<th></th>
<th>Observer B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td><strong>Signal</strong></td>
<td>8</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td><strong>Noise</strong></td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
P(C) equals 0.80 for observer A and 0.20 for observer B, even though neither observer demonstrates any ability to discriminate (i.e., they respond "yes" to every stimulus). Note that P(C)=P(S₂) for both observers. For both observers, d’ equals 0 (i.e., chance performance), and C equals --∞ (i.e., a strong bias toward saying “yes”). This is one reason why one should not compare P(C) across conditions in which P(S₂) is different.

**Example 2:** P(C) is dependent on the cost and values associated with the decision outcomes. For example, suppose that observer A associates an extremely high cost with a miss and, therefore, always responds “yes.” On the other hand, observer B associates an extremely high cost with a false alarm and, therefore, always responds “no.”

<table>
<thead>
<tr>
<th>Observer A</th>
<th>Observer B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Signal</td>
<td>8</td>
</tr>
<tr>
<td>Noise</td>
<td>2</td>
</tr>
</tbody>
</table>

P(C) equals 0.80 for observer A and 0.20 for observer B, even though neither observer demonstrates any ability to discriminate. For both observers, d’ equals 0. C equals --∞ (i.e., a strong bias toward saying “yes”) for observer A and +∞ (i.e., a strong bias toward saying “no”) for observer B.

**Example 3:** Even if P(S₁)= P(S₂)=0.5, P(C) can take on a wide range of values for a fixed level of accuracy (as measured by d’). For example, if d’=1.37, then either of the following outcomes could occur on a diagnostic task:

<table>
<thead>
<tr>
<th>Observer A</th>
<th>Observer B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Signal</td>
<td>17</td>
</tr>
<tr>
<td>Noise</td>
<td>1</td>
</tr>
</tbody>
</table>

P(C) equals 0.58 for observer A and 0.75 for observer B, even though each observer demonstrates the same ability to discriminate (i.e., d’=1.37). C equals 1.64 (i.e., a strong bias toward saying “no”) for observer A and -0.19 (i.e., a slight bias toward saying “yes”) for observer B.

The second characteristic of a good accuracy index (i.e., invariance to changes in response bias) indicates an ROC curve on which every point is indicative of the same degree of accuracy. As mentioned in section
2, measures which imply a continuous, variable-criterion model with normal distributions that can have unequal variance (e.g., ROC measure d) come closest to possessing this characteristic.

Non-SDT accuracy measures, like P(C), indicate a specific model of the diagnostic task. Since H = [P(C) – (1 – P(S2))(1 – F)]/P(S2) (from the definition of P(C)), (H, F) pairs with equal values of P(C) generate a theoretical ROC curve that is a straight line, with slope of (1 – P(S2))/P(S2), on linear coordinates and curvilinear on z-coordinates (Swets 1986a, 1986b). An ROC curve with such characteristics implies a “double-threshold” model in which the X-axis is not continuous, the criterion is fixed (not variable), and signal and noise distributions are uniform (rectangular) (Macmillan and Kaplan 1985; Swets 1986b). No empirical ROC curves have been found that possess this shape (Macmillan and Kaplan 1985; Swets 1986a). Therefore, as implied by examples 1-3, P(C) is both an invalid and unreliable measure of accuracy (e.g., Swets 1986a).

The use of an SDT measure of accuracy seems appropriate for accounting diagnostic decisions because factors other than discrimination (or accuracy) are often present that can affect these decisions. First, accounting diagnostic decisions often involve situations in which P(S1) ≠ P(S2). For example, in error detection the probability of client firms having a material error is probably less than the probability of client firms not having a material error. Second, accounting diagnostic decisions often involve situations in which the costs and values associated with the decision actions are different. For example, the cost to the firm of incorrectly deciding that a material error is likely to exist (i.e., the cost of additional audit work) is probably less than the cost to the firm of incorrectly deciding that no material error is likely to exist (i.e., the cost of a lawsuit associated with an audit failure). Furthermore, estimates of these costs are likely to differ across auditors.

**Ability to Measure Response Bias and Confidence**

While a great deal of the research employing SDT concepts, methods, and measures has done so in order to use a measure of accuracy that is free of response bias, some research is interested in response bias in its own right. A second advantage associated with SDT is that superior measures of response bias are available. SDT response bias measures have many desirable characteristics. First, a good response bias index should be monotonic with H and F in the same direction and treat H and F symmetrically (Macmillan and Creelman 1990, 1991). Symbolically,

\[
\text{response bias} = v[u(H) + u(F)]
\]
where \( v \) and \( u \) are both monotonic functions. Response bias measures \( C \) and \( C_a \) possess these intuitively appealing qualities; \( \beta \) does not.

Second, a good response bias index should be independent of its associated sensitivity index (Macmillan and Creelman 1990, 1991). Two types of independence have been discussed: statistical independence (i.e., the response bias index and the sensitivity index are independent random variables) and range independence (i.e., the range of the response bias index does not depend on the range of the sensitivity index). For yes-no experiments, bias index \( C \) is statistically independent of \( d' \); \( \beta \) is not (Macmillan and Creelman 1991). The range of \( C \) does not depend on \( d' \), while the range of \( \beta \) does (Banks 1970; Snodgrass and Corwin 1988). For confidence rating experiments, bias index \( C_a \) is statistically independent of \( A_z \) only in cases in which \( B \) (estimate of \( \sigma_1 / \sigma_2 \)) is equal to one. The range of \( C_a \) does not depend on \( A_z \) (Macmillan and Creelman 1991).

Third, in confidence rating experiments it is possible to calculate an intuitively appealing surrogate for confidence in judgment of a subject from the response bias measure \( C_a \). As previously mentioned, the scale used in confidence rating experiments ranges from 1 (item is definitely or almost definitely signal) to \( n \) (item is definitely or almost definitely noise). Intuitively, if one subject classifies a greater proportion of the items in the extreme categories of the rating scale than another subject does, then the former subject is assumed to be more confident than the latter subject is. The \( C_a \) values for the more confident subject will be smaller in absolute value (on average across decision criteria) than for the less confident subject.

Therefore, a good confidence index should be monotonic with the relative frequency of usage of the extreme categories of the rating scale. Sprinkle and Tubbs (1998) employ such an index, the linear trend of \( C_a \), as a surrogate for willingness to rely on memory (i.e., confidence). Symbolically,

\[
\text{confidence} = \sum_{k=1}^{\frac{n}{2}} L_k C_{ak}
\]

(10)

where \( L_k \) is the \( k^{th} \) linear coefficient of orthogonal polynomials with \( n \) categories. Since Sprinkle and Tubbs employ five rating categories, \( n \) equals 4 and the coefficients are \(-3, -1, +1, \) and \(+3\). With this index, more frequent usage of the extreme categories of the rating scale (i.e., greater confidence) results in lower values of the index (for an example, see Sprinkle and Tubbs 1998, 487-489).

Previous accounting research (Brown 1981; Blocher, Moffie, and Zmud 1986; Sprinkle and Tubbs 1998) has examined the effect of various manipulations on response bias and/or confidence. Further examination of the
effect of different situations (e.g., closely-held or public audit client) or different subjects (e.g., staff or senior auditor; Big 5 or non-Big 5 auditor) on response bias and/or confidence are likely to yield important insights regarding the perceived costs and values associated with different decision actions and the perceived prior probabilities.

**Ability to Calculate Accuracy and Response Bias Contemporaneously**

A third advantage associated with SDT is that an accuracy measure and a response bias measure can be calculated contemporaneously for the same manipulation. This ability enables the researcher to test whether a particular manipulation influences accuracy, response bias, or both. For example, when SDT methods and measures were first applied to tasks in several areas of psychology, it was discovered that several presumed accuracy effects from previous literature were actually response bias effects. Other psychology studies demonstrated that certain manipulations changed both accuracy and response bias (for a review, see Swets 1973).

**Ability to Calculate Optimal Accuracy and Response Bias**

Numerous diagnostic situations, both in practice and research, exist in which it is useful to be able to calculate optimal accuracy and/or response bias. A fourth advantage associated with employing SDT is that optimal accuracy and response bias can be calculated. In order to calculate optimal accuracy, it is necessary to know the values of $\mu_1$, $\mu_2$, $\sigma_1$, and $\sigma_2$. In order to calculate optimal response bias, sometimes referred to as the “optimal positivity criterion” (e.g., Swets 1992), it is necessary to know the values of $P(S_2)$, benefit (no|S_1), cost (yes|S_1), benefit (yes|S_2), and cost (no|S_2). These values are often difficult to assess. However, even in cases in which the priors and the benefits and costs associated with decision outcomes are not exactly known, some decision theorists argue that explicit estimation of these unknowns and calculation of the resulting optimal response bias is the preferable approach to decision making (Swets 1992, 525).

Swets (1992) states that while the concept of setting the “optimal positivity criterion” is well accepted in many fields of academic research, it “has yet to influence most practical arenas, including several of substantial import to individuals and society” (e.g., HIV testing and materials testing of aircraft) (522). It is unclear whether optimal positivity criteria are explicitly implemented for accounting decisions (e.g., going concern opinions, bankruptcy prediction, audit client acceptance).
In some research applications it is useful to compare actual accuracy and/or response bias with optimal accuracy and/or response bias. For example, in Brown (1981) the effect of environmental factors on relative accuracy (the ratio of actual accuracy to optimal accuracy) and relative anchoring bias (the difference between actual and optimal response bias scaled by optimal response bias) is investigated.

**Increased Validity**

A final advantage associated with employing SDT concepts and measures is that the validity of research is likely to be enhanced for the following reasons. First, since SDT provides a theoretical model of the diagnostic task, the use of SDT concepts (like accuracy and response bias) contributes to construct validity by improving the “pre-experimental explication of constructs so that definitions are clear” (Cook and Campbell 1979, 60). Second, use of SDT accuracy measures (like $d'$, $d_0$ or $A_z$) contributes to construct validity by ensuring that “variables . . . [are] not . . . dominated by irrelevant factors that make them measures of more or less than was intended” (Cook and Campbell 1979, 61). Clearly, this is not the case with non-SDT accuracy measures like $P(C)$, which may have high “face validity” (Nunnally 1978, 111), that are dependent on irrelevant factors like the probability of signal and the costs and values associated with the decision outcomes (see the third section). Finally, use of SDT measures contributes to statistical conclusion validity (Cook and Campbell 1979, 43) because SDT measures are more reliable than non-SDT measures (see, for e.g., Swets and Pickett 1982; Swets 1986b). For example, use of a non-SDT accuracy measure like $P(C)$ is clearly a threat to statistical conclusion validity because its value can vary widely when accuracy is constant (see the third section). In fact, Sprinkle and Tubbs (1998) demonstrate greater power (i.e., larger effect sizes) in their hypothesis tests with SDT measures than with non-SDT measures.

**ACCOUNTING STUDIES THAT EMPLOY SDT**

Only three experimental studies that employ SDT concepts, methods, and measures have appeared in accounting journals. The following descriptions of these studies include the issue examined, the procedure employed, the measure of accuracy used, the measure of response bias used, the findings, and the reason(s) why it was important to employ SDT.

Brown (1981) examines, in a standard cost variance investigation task, how (1) amount of prior distributional information about states (less information, more information), (2) degree of overlap between prior distributions of states (less overlap, more overlap), and (3) differential cost of decision error types (cost of miss is 3 times the cost of false alarm, cost of false alarm is 3 times the cost of miss) affect (1) relative accuracy (the ratio
of individual decision accuracy to optimal decision accuracy) and (2) relative anchoring bias (the difference between individual response bias and optimal response bias scaled by optimal response bias). The yes-no procedure is employed with responses of “in control” and “out of control.” A distance measure of yes-no accuracy, $d'$, is employed. A likelihood ratio measure of yes-no response bias, $\beta$, is employed. Relative accuracy was found to be greater when the cost of a miss is 3 times the cost of a false alarm (i.e., when the cost structure favors more investigation). Relative anchoring bias was also found to be greater when the cost of a miss is 3 times the cost of a false alarm. It was important to employ SDT in this study because (1) the effect of the independent variables on both relative accuracy and relative anchoring bias is examined and (2) optimal $d'$ and optimal $\beta$ are needed in order to measure relative accuracy and relative anchoring bias.

Blocher et al. (1986) examine how report format (graphic, tabular) and task complexity (low, high) affect the decision accuracy and response bias of internal auditors’ risk judgments. The confidence rating procedure is employed with a 6-point response scale (1 “surely low risk” to 6 “surely high risk”). An area measure of confidence rating accuracy, the proportion of the unit square lying under the line connecting the points provided by the confidence rating procedure (see Macmillan and Creelman 1991, 109 eq. 4.10), is employed. A likelihood ratio measure of yes-no response bias (the 6-point response scale is converted to “yes” [if 1, 2 or 3] and “no” [if 4, 5, or 6]), the Hodos bias measure (see Macmillan and Creelman 1991, 109 eq. 4.11), is employed. Accuracy was found to be higher for the tabular format in the high complexity condition but higher for the graphical format in the low complexity condition. Response bias was found to be lower (i.e., a greater tendency to report “high risk”) for the graphical format in the low complexity condition but lower for the tabular format in the high complexity condition. Furthermore, response bias was found to be lower on average in the high complexity condition. It was important to employ SDT in this study because the effect of the independent variables on both accuracy and response bias is examined.

Sprinkle and Tubbs (1998) examine how level of audit risk of the area (lower risk, higher risk) and degree of importance of an information item within the area (less important, more important) affect the memory accuracy and willingness to rely on memory (pattern of response bias levels) in a working paper review and recognition task. The confidence rating procedure is employed with a 5-point response scale (1 “definitely, or almost definitely, true” to 5 “definitely, or almost definitely, false”). An area measure of confidence rating accuracy, $A_z$, is employed. A criterion location measure of confidence rating response bias, $C_m$, is employed. Accuracy was found to be positively related
to level of audit risk of the area and degree of importance of an information item within the area. Willingness to rely on memory was found to be negatively related to the degree of information importance. It was important to employ SDT in this study because (1) the effect of the independent variables on both memory accuracy and willingness to rely on memory is examined and (2) C is needed in order to measure willingness to rely on memory.

**TWO ACCOUNTING STUDIES THAT COULD HAVE EMPLOYED SDT**

In this section, we reanalyze data from two papers for which we had data access. We find different results using SDT techniques and gain important additional insights using separate measures of accuracy, bias, and confidence. We also demonstrate techniques to overcome data related problems when employing SDT.

**Ramsay (1994)**

Ramsay (1994) is an experimental accounting study that did not employ SDT measures. We reanalyze the data from Ramsay (1994) in order to (1) compare the results of SDT and non-SDT accuracy measures, (2) calculate SDT-based measures of response bias and confidence, and (3) illustrate a potential difficulty and the preferred solution to that difficulty. Ramsay (1994) examines how experience level (senior, manager) and error type (mechanical, conceptual) affect accuracy in a working paper review and recognition task. It is posited that managers will be more accurate in recognizing conceptual errors while seniors will be more accurate in recognizing mechanical errors (an experience level by error type interaction). In the mechanical condition four noise (i.e., an error not present in the working papers) and four signal (i.e., an error present in the working papers) items are presented. In the conceptual condition one noise and seven signal items are presented. The confidence rating procedure is employed with a 6-point response scale (1 “definitely, or almost definitely, false” to 6 “definitely, or almost definitely, true”).

The low number of signal and noise items and the relatively high number of rating categories resulted in many empty cells in subjects’ outcome matrices. Empty cells in an outcome matrix may lead to an unreliable ROC solution (see, for e.g., Metz 1989; Dorfman and Berbaum 1995). One possible solution to the empty cell problem is to analyze the data as if it were from a yes-no experiment. That is, responses in categories 1, 2, and 3 can be considered “no” responses, and responses in categories 4, 5, and 6 can be considered “yes” responses. SDT measures d’ and C can then be calculated. This approach is not effective with the Ramsay data because Z(F) is always infinite for the noise trials in the conceptual condition (i.e., since there is only one noise trial, F=0 or F=1). Infinite Z-scores lead to infinite d’ and C and are, thus, to be avoided because of their effect on the mean and
the variance of the conditions in which they occur (Macmillan and Kaplan 1985; Macmillan and Creelman 1991, 268-273).

Ramsay (1994) treats the data as if it were from a yes-no experiment but employs P(C) as the measure of accuracy. Unlike d', P(C) is always finite, even in situations in which H or F equal zero or one. However, as previously discussed, P(C) is an invalid and unreliable measure of accuracy. Least-squares means and the ANOVA table for P(C) as the dependent variable are presented in Table 2. As hypothesized, a significant experience level by error type interaction with the predicted pattern is found (t=2.1671, one-tailed p=.0185). Furthermore, seniors were significantly more accurate than managers in the mechanical condition (t=−1.7010, one-tailed p=.0466), and managers were of moderately significant greater accuracy than seniors in the conceptual condition (t=1.4156, one-tailed p=.0806).

(Please insert Table 2 here)

The preferred solution to the empty cell problem is to merge the rating-scale judgments made by all subjects in a condition into one outcome matrix, and, then, calculate a group ROC from that matrix. This approach, referred to as pooling, is “the method of choice when data are relatively scarce and when the desire is to gain greater stability for the estimates of performance indices” (Swets and Pickett 1982, 64; also see Macmillan and Kaplan 1985, 196). Macmillan and Creelman (1991, 281) point out that ROC curves obtained by pooling require special estimation procedures and recommend the RSCORE-J program of Dorfman and Berbaum (1986). The RSCORE-J program employs the jackknife method and the RSCORE estimation procedure to yield pseudovalues for each subject within the pool. These pseudovalues can then be analyzed as if they were individual-subject estimates of ROC parameters (Dorfman and Berbaum 1986, 452; for an example, see Franken et al. 1994).

RSCORE-J was employed to yield estimates of $A_2$ (accuracy measure), Mean $C_0$ (bias measure), and Mean $C_v^2$ (confidence measure) for each subject in each condition. Least-squares means and ANOVA tables for these dependent variables are presented in Tables 2 and 3. When $A_2$ is the dependent variable, a moderately significant experience level by error type interaction with the predicted pattern is found (t=1.4549, one-tailed p=.0772). Furthermore, seniors were not significantly more accurate than managers in the mechanical condition ($t=−0.6929$, one-tailed $p=.2453$), and managers were of moderately significant greater accuracy than seniors in the conceptual condition ($t=1.3898$, one-tailed $p=.0844$). In addition, a moderately significant effect of error type is found ($F=3.6144$, $p=.0653$).
When Mean $C_a$ is the dependent variable, a significant effect of error type is found ($F=6.3707, p=.0162$). Mean $C_a$ is significantly lower for conceptual errors ($-0.3213$) than for mechanical errors ($-0.0335$). Subjects, therefore, exhibited greater bias toward saying “yes” (i.e., “error is present”) for conceptual errors. This result is reasonable because the cost of a miss of a conceptual error is likely to be higher than the cost of a miss of a mechanical error.

When Mean $C_a^2$ is the dependent variable, a significant effect of error type is found ($F=4.7180, p=.0365$). Mean $C_a^2$ is significantly lower for conceptual errors ($0.2261$) than for mechanical errors ($0.5102$). Subjects, therefore, exhibited greater confidence for conceptual errors. Perhaps greater confidence for conceptual errors is the result of spending more time reviewing conceptual errors because conceptual errors are more important. Similarly, Sprinkle and Tubbs (1998) found that confidence was greater for more important items.

Conclusions regarding Ramsay (1994) are altered as a result of employing SDT as opposed to non-SDT measures. First, the primary thesis of an experience level by error type interaction upon accuracy is no longer significant (at $a=.05$). Second, seniors are not more accurate than managers at detecting mechanical errors. Third, auditors are of moderately significant greater accuracy in identifying mechanical (as opposed to conceptual) errors. Fourth, auditors’ decision factors (e.g., perceived prior probability of errors, perceived benefits and costs of different decision outcomes) appear to cause them to exhibit a greater bias toward saying “yes” for conceptual (as opposed to mechanical) items. Finally, auditors are more confident in identifying conceptual errors than mechanical errors. We believe these are important additional findings regarding reviewers’ judgments.

**Carcello and Neal (2000)**

As previously mentioned, SDT procedures and measures associated with yes-no experiments can be employed with archival data. For example, Carcello and Neal (2000) examine the effect of audit committee independence on auditor going-concern reports. A sub-sample of their data (not reported in their paper) can be cross classified with respect to bankrupt or not bankrupt within one year (stimulus) and going-concern or clean report (response).

<table>
<thead>
<tr>
<th></th>
<th>Going Concern</th>
<th>Clean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bankrupt</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Not Bankrupt</td>
<td>49</td>
<td>82</td>
</tr>
</tbody>
</table>
Applying SDT measures to this diagnostic task results in $d'$ equal to 0.93 (moderate discrimination) and $C$ equal to −0.14 (no bias).

The results are much different if the data are partitioned based upon whether the audit committee is independent or affiliated (i.e., at least one “inside” or “gray” director).

<table>
<thead>
<tr>
<th></th>
<th>Independent</th>
<th></th>
<th>Affiliated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Going Concern</td>
<td>Clean</td>
<td>Going Concern</td>
</tr>
<tr>
<td>Bankrupt</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Not Bankrupt</td>
<td>31</td>
<td>26</td>
<td>18</td>
</tr>
</tbody>
</table>

For the independent condition, $d'$ is equal to 1.36 (moderate to high discrimination), and $C$ is equal to −0.79 (a bias toward saying “going concern”). For the affiliated condition, $d'$ is equal to 0.02 (no discrimination), and $C$ is equal to 0.69 (a bias against saying “going concern”).

It appears as if auditors of firms with independent audit committees are more biased toward issuing going-concern reports than auditors of firms with affiliated audit committees. This result is consistent with the conclusion of Carcello and Neal (2000). It also appears as if auditors of firms with independent audit committees are better able to discriminate between bankrupt and non-bankrupt firms than auditors of firms with affiliated audit committees, an issue not examined by Carcello and Neal (2000). The underlying reason for this difference in discrimination ability is beyond the scope of the current paper, but it certainly warrants further investigation. Our purpose is to demonstrate the value of analyzing archival information about decisions on both accuracy and bias dimensions.

**CONCLUSION**

The current paper has presented the following thesis. Diagnostic tasks are prevalent in accounting practice, and have been studied extensively in accounting research. Previous non-accounting research has demonstrated that SDT concepts, methods, and measures are theoretically and empirically valid in the examination of diagnostic tasks. For that reason, SDT has been widely adopted in many fields of research. However, only three experimental studies that employ SDT have appeared in accounting journals. Therefore, it appears as if the increased application of SDT in accounting research is warranted.

Hence, accounting researchers examining diagnostic tasks may want to consider the following recommendations. First, if surrogates for the theoretical constructs of accuracy, response bias, and/or confidence...
are needed in experimental or quasi-experimental research, then one should use the methods and measures of SDT. Second, for yes-no experiments or archival research, d’ should be employed as the measure of accuracy rather than non-SDT measures like P(C). Third, for yes-no experiments or archival research, C should be employed as the measure of response bias rather than β unless the goal is to compare actual with optimal response bias. Fourth, when possible confidence rating experiments are to be preferred to yes-no experiments. Fifth, confidence rating experiments should be designed with a balance of the number of signal items, noise items, and rating categories in order to avoid a high likelihood of empty cells in subjects’ outcome matrices. Sixth, when many empty cells occur in subjects’ outcome matrices, pooling of data and special estimation procedures are necessary. Finally, for confidence rating experiments, one should use A₂ as the measure of accuracy and Cₐ as the measure of response bias. Following these recommendations is likely to enhance the validity of accounting research examining diagnostic tasks.
ENDNOTES

1 Examples of diagnostic tasks examined in accounting research include: error detection in auditor workpaper review (Ramsay 1994; Bamber and Ramsay 1997), bankruptcy prediction (Simnett 1996), error or fraud detection in an audit (McDaniel 1990; Nelson 1993; Braun 2000; Knapp and Knapp 2001), memory for audit evidence previously presented (Moeckel and Plumlee 1989; Johnson 1994), and classification of firms as using LIFO or FIFO (Liang et al. 1992).


3 “Signal A” and “signal B” can be substituted for “signal” and “noise” with no loss of applicability (Swets 1996, xiii).

4 While the discussion in this manuscript is limited to diagnostic tasks with two stimuli, SDT concepts, methods, and measures can also be applied to diagnostic tasks with more than two stimuli (see, for e.g., Macmillan and Creelman 1991, Chapters 9-10).

5 In a 2 x 2 outcome matrix, H=hits/(hits+misses) and F=false alarms/(false alarms+correct rejections).

6 This curve is also referred to as a receiver-operating-characteristic curve or an isosensitivity curve.

7 SDT assumptions imply an ROC curve with particular characteristics. The assumptions of continuous X and variable Xk imply an ROC curve in which H is a monotonically increasing function of F (Green and Swets 1966). These assumptions also imply an ROC curve that passes through (0, 0) and (1, 1). That is, F=0 if and only if H=0, and H=1 if and only if F=1 (Swets 1986b). If it is also assumed that S1 and S2 are normally distributed, the ROC curve on z-coordinates (a graph in which the linear coordinates are rescaled so that normal-deviate values are spaced linearly) will be a straight line (Green and Swets 1966). That is, Z(H)=A + BZ(F) where A is the y-intercept, B is the slope, and Z is the inverse of the normal distribution function, which converts H and F to z-scores. If it is further assumed that S1 and S2 have equal variance, the ROC curve on linear coordinates (as in Figure 2) will be concave and symmetric about the minor diagonal, and the ROC curve on z-coordinates will be of unit slope (Egan 1975). That is, Z(H)=A + Z(F).

8 Procedures and measures associated with yes-no experiments have been successfully employed with archival data (e.g., weather forecasts, medical diagnoses).

9 Use of d’ as a measure of accuracy in yes-no experiments is based upon the assumption that the signal and noise distributions are normally distributed with equal variance (Swets 1986b).

10 Use of C as a measure of response bias in yes-no experiments is based upon the assumption that the signal and noise distributions are normally distributed with equal variance (Macmillan and Creelman 1990).
Use of $\beta$ as a measure of response bias in yes-no experiments is based upon the assumption that the signal and noise distributions are normally distributed with equal variance (Macmillan and Creelman 1990).

The SDT methods and measures associated with confidence rating experiments have also been recommended (e.g., Mason 1982; Levi 1985; Rockette, Gur, and Metz 1992) for responses that are probabilistic (continuous) rather than categorical.

Maximum-likelihood estimation is preferred to linearly regressing $Z(H)$ on $Z(F)$ to estimate the y-intercept ($A$) and slope ($B$) because $H$ and $F$ are both dependent variables (see, for e.g., Macmillan and Creelman 1991, 78).

The following measures could also be employed for yes-no experiments in which either $P(S_2)$ or the costs and benefits of the various decision outcomes are manipulated. In such cases, an ROC curve can be generated.

Use of $d_a$ as a measure of accuracy in rating experiments is based upon the assumption that the signal and noise distributions are normally distributed and can have unequal variance (Macmillan and Creelman 1991, 68-70).

Accuracy measure $d_a$ can also be calculated for individual points on the ROC curve as $\frac{\sqrt{2}}{(1+B^2)^{0.5}} [Z(H_k) - B*Z(F_k)]$.

$A_z$ is a monotonic transformation of $d_a$ (i.e., $A_z = \Phi(d_a/\sqrt{2})$).

Use of $A_z$ as a measure of accuracy in rating experiments is based upon the assumption that the signal and noise distributions (1) are of a form “that can be transformed monotonically to the normal distribution” (Swets 1986a, 183) and (2) can have unequal variances.

Use of $C_a$ as a measure of response bias in rating experiments is based upon the assumption that the signal and noise distributions are normally distributed but may have unequal variances (Macmillan and Creelman 1991).

These indices, which imply threshold models, are “clearly at odds with existing data . . . [and] subject to unnecessary unreliability” (Swets 1986b, 100).

SDT accuracy measure $A_z$ treats $H$ and $F$ symmetrically only when $B=1$ (i.e., when $\sigma_1=\sigma_2$).

$P(C)$ is presented as an example of a non-SDT accuracy measure because $P(C)$ has been frequently employed in accounting research. Alternative non-SDT accuracy measures (see Swets 1986b) are subject to most of the same disadvantages.

According to Nunnally (1978 111), “face validity” is irrelevant if a measure does not validly measure a construct.

However, other experimental (e.g., Brown 1983) and archival (e.g., Hopwood, McKeown, and Mutchler 1994) studies have incorporated SDT-related concepts like the effect of the costs and benefits of different decision outcomes on judgment. In addition, a recent archival accounting paper, Phillips, Pincus, and Rego (2003), employs the area under the
ROC curve as a measure of classification accuracy. Hosmer and Lemeshow (2000) advocate such an approach as a "more complete description of classification accuracy" for logistic regression models (160-164).

Since the probability of a signal is .875 in the conceptual condition (i.e., 7 signal items and 1 noise items), subjects' bias toward saying "yes" in that condition (i.e., Mean $C_a$=−0.3213) results in $P(C)$ being an upwardly biased (toward 1) measure of accuracy (as for observer A - example 1 in section 3). This upward bias is likely to have masked the true effect of error type when $P(C)$ was the measure of accuracy.

In order to avoid an infinite $Z(H)$, we employ the often recommended solution of replacing $H=1$ with $H=1−\frac{1}{2N}$ in which $N$ is the "number of trials on which the relevant stimulus was present" (Macmillan and Kaplan 1985, 191-192).
REFERENCES


<table>
<thead>
<tr>
<th>Measure</th>
<th>Surrogate</th>
<th>Construct</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d')</td>
<td>(Z(H) - Z(F))</td>
<td>(\frac{\mu_2 - \mu_1}{\sigma_1})</td>
</tr>
<tr>
<td>(\beta)</td>
<td>(\exp{-0.5[Z(H)^2 - Z(F)^2]})</td>
<td>(\frac{\phi(X_n</td>
</tr>
<tr>
<td>(C)</td>
<td>(-0.5[Z(H) + Z(F)])</td>
<td>(\frac{X_n - 0.5(\mu_2 + \mu_1)}{\sigma_1})</td>
</tr>
<tr>
<td>(d_a)</td>
<td>(\frac{\sqrt{2} A}{(1 + B^2)^{0.5}})</td>
<td>(\frac{\mu_2 - \mu_1}{(0.5(\sigma_2^2 + \sigma_1^2)^{0.5}})</td>
</tr>
<tr>
<td>(A_a)</td>
<td>(\Phi\left(\frac{A}{(1 + B^2)^{0.5}}\right))</td>
<td>(\Phi\left[\frac{(\mu_2 - \mu_1)}{(\sigma_2^2 + \sigma_1^2)^{0.5}}\right])</td>
</tr>
<tr>
<td>(C_a)</td>
<td>(-\frac{\sqrt{2} B}{(1 + B^2)^{0.5} (1 + B)} [Z(H_n) + Z(F_k)])</td>
<td>(\frac{X_n - (\sigma_2 \mu_2 + \sigma_1 \mu_1)}{(0.5(\sigma_2^2 + \sigma_1^2)^{0.5}})</td>
</tr>
</tbody>
</table>

**Theoretical Concepts**

\(\mu_2\): mean of signal distribution.

\(\mu_1\): mean of noise distribution.

\(\sigma_2\): standard deviation of the signal distribution.

\(\sigma_1\): standard deviation of the noise distribution.

\(x_n\): decision criterion.

\(\phi(x_n | S_n)\): ordinate of \(S_n\) distribution at criterion \(X_n\).

**Empirical Measures**

\(H\): hit rate.

\(F\): false alarm rate.

\(Z(\bullet)\): \(Z\)-score transform of the probability argument.

\(\exp(\bullet)\): constant \(e\) raised to the power of the argument.

\(\Phi(\bullet)\): probability transform of the \(Z\)-score argument.

\(A\): maximum likelihood estimate of y-intercept (i.e., estimate of \(\frac{\mu_2 - \mu_1}{\sigma_2}\) ) of the binormal (i.e., z-coordinates) ROC.

\(B\): maximum likelihood estimate of slope (i.e., estimate of \(\frac{\sigma_1}{\sigma_2}\) ) of the binormal ROC.
### TABLE 2

Least-squares Means and ANOVAs for P(C) and $A_z$

**Dependent Variable: P(C)**

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Experience Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Senior</td>
<td>Manager</td>
</tr>
<tr>
<td>Mechanical</td>
<td>.6938</td>
</tr>
<tr>
<td>Conceptual</td>
<td>.6500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>0.0014</td>
<td>1</td>
<td>0.0014</td>
<td>0.0422</td>
<td>.8385</td>
</tr>
<tr>
<td>Subject(Level)</td>
<td>1.2190</td>
<td>36</td>
<td>0.0339</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>0.0494</td>
<td>1</td>
<td>0.0494</td>
<td>1.3616</td>
<td>.2509</td>
</tr>
<tr>
<td>Type*Level</td>
<td>0.1703</td>
<td>1</td>
<td>0.1703</td>
<td>4.6963</td>
<td>.0369</td>
</tr>
<tr>
<td>Type*Subject(Level)</td>
<td>1.3051</td>
<td>36</td>
<td>0.0363</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Dependent Variable: $A_z$**

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Experience Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Senior</td>
<td>Manager</td>
</tr>
<tr>
<td>Mechanical</td>
<td>.7593</td>
</tr>
<tr>
<td>Conceptual</td>
<td>.5226</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>0.0223</td>
<td>1</td>
<td>0.0223</td>
<td>0.2489</td>
<td>.6209</td>
</tr>
<tr>
<td>Subject(Level)</td>
<td>3.2297</td>
<td>36</td>
<td>0.0897</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>0.3406</td>
<td>1</td>
<td>0.3406</td>
<td>3.6144</td>
<td>.0653</td>
</tr>
<tr>
<td>Type*Level</td>
<td>0.1995</td>
<td>1</td>
<td>0.1995</td>
<td>2.1168</td>
<td>.1544</td>
</tr>
<tr>
<td>Type*Subject(Level)</td>
<td>3.3923</td>
<td>36</td>
<td>0.0942</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 3
Least-squares Means and ANOVAs for mean $C_a$ and mean $C_a^2$

**Dependent Variable: Mean $C_a$**

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Experience Level</th>
<th>Senior</th>
<th>Manager</th>
<th>$C_a$</th>
<th>$C_a^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanical</td>
<td>Senior Manager</td>
<td>−0.1455</td>
<td>0.0785</td>
<td>−0.0335</td>
<td></td>
</tr>
<tr>
<td>Conceptual</td>
<td>Senior Manager</td>
<td>−0.2908</td>
<td>−0.3517</td>
<td>−0.3213</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>−0.2182</td>
<td>−0.1366</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>0.1261</td>
<td>1</td>
<td>0.1261</td>
<td>0.2890</td>
<td>.5942</td>
</tr>
<tr>
<td>Subject(Level)</td>
<td>15.7009</td>
<td>36</td>
<td>0.4361</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>1.5696</td>
<td>1</td>
<td>1.5696</td>
<td>6.3707</td>
<td>.0162</td>
</tr>
<tr>
<td>Type*Level</td>
<td>0.3846</td>
<td>1</td>
<td>0.3846</td>
<td>1.5611</td>
<td>.2196</td>
</tr>
<tr>
<td>Type*Subject(Level)</td>
<td>8.8694</td>
<td>36</td>
<td>0.2464</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Dependent Variable: Mean $C_a^2$**

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Experience Level</th>
<th>Senior</th>
<th>Manager</th>
<th>$C_a^2$</th>
<th>$C_a^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanical</td>
<td>Senior Manager</td>
<td>0.5334</td>
<td>0.4870</td>
<td>0.5102</td>
<td></td>
</tr>
<tr>
<td>Conceptual</td>
<td>Senior Manager</td>
<td>0.2634</td>
<td>0.1888</td>
<td>0.2261</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3984</td>
<td>0.3379</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>0.0695</td>
<td>1</td>
<td>0.0695</td>
<td>0.1441</td>
<td>.7064</td>
</tr>
<tr>
<td>Subject(Level)</td>
<td>17.3487</td>
<td>36</td>
<td>0.4819</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>1.5295</td>
<td>1</td>
<td>1.5295</td>
<td>4.7180</td>
<td>.0365</td>
</tr>
<tr>
<td>Type*Level</td>
<td>0.0038</td>
<td>1</td>
<td>0.0038</td>
<td>0.0116</td>
<td>.9147</td>
</tr>
<tr>
<td>Type*Subject(Level)</td>
<td>11.6709</td>
<td>36</td>
<td>0.3242</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FIGURE 1
Illustrative Probability Density Functions of Noise and Signal Distributions (1 Criterion Included)

\[ S_1 \text{ (Noise)} \]
\[ S_2 \text{ (Signal)} \]

\[ \text{Criterion}(X_\theta) \]

\[ X = \text{Magnitude of Sensory Observation} \]

False alarms

Hits
FIGURE 2
Illustrative Receiver-operating-characteristic (ROC) Curve

Area Under Curve = $A_z = 0.6744$

0 = empirical points from rating task.