Psychonomic Bulletin & Review
2003, 10 (3), 556-569

"Nonparametric" A' and other modern misconceptions about signal detection theory

RICHARD E. PASTORE
Binghamton University, SUNY, Binghamton, New York

EDWARD J. CRAWLEY
Marywood University, Scranton, Pennsylvania

and

MELODY S. BERENS and MICHAEL A. SKELLY
Binghamton University, SUNY, Binghamton, New York

Many modern descriptions of signal detection theory (SDT) are, at best, distorted caricatures of the Gaussian equal-variance model of SDT (G-SDT). The distortions have sometimes led to important, but unwarranted, conclusions about the nature of cognitive processes. Some researchers reject using d' and β because of concerns about the validity of explicit underlying assumptions (that are shared with most inferential statistics), instead using either the supposedly "nonparametric" measures of A' and B'' or measures known to confound ability and bias. The origins, development, and underlying assumptions of SDT are summarized, then contrasted with modern distortions and misconceptions. The nature and interpretation of common descriptive statistics for sensitivity and bias are described along with important pragmatic considerations about use. A deeper understanding of SDT provides researchers with tools that better evaluate both their own findings and the validity of conclusions drawn by others who have utilized SDT measures and analyses.

Signal detection theory (SDT) is a powerful, statistically based research tool that, if used properly, can enhance modern research. However, if used improperly, SDT, like other important research tools (e.g., inferential statistical tests), can lead to incorrect conclusions. Unfortunately, SDT is poorly understood by many modern researchers and thus is sometimes used inappropriately to support or reject important conclusions. Many of the problems with modern uses of SDT, as well as recent arguments against the use of SDT (e.g., Norris, 1995), appear to stem from misunderstandings about the nature of the underspecified decision variable and about the assumptions underlying specific descriptive statistics. For example, having explicit assumptions for d' and β, the respective descriptive statistics for sensitivity and bias that are most typically associated with SDT is apparently of great concern to some researchers, with many opting to use A' and B'' because these alternative measures are assumed to be nonparametric and thus free from assumptions about distribution. The goal of the present work is to examine the nature of the aspects of SDT that seem to be most misunderstood in modern research.

The authors thank John Wixted, John Irwin, Robert Sorkin, and an anonymous reviewer for comments and suggestions that were extremely helpful in our preparing this manuscript in a form that allows difficult material to be presented in a clear and easily followed manner. Preparation of this paper was supported in part by a grant from National Science Foundation. Correspondence should be addressed to R. E. Pastore, Department of Psychology, Binghamton University, Binghamton, NY 13902-6000 (e-mail: pastore@binghamton.edu).

Origin of SDT: Statistical Decision Theory

In all aspects of psychology, the researcher is faced with problems that stem from the need to use simple behavioral measures to study the complex processes involved in *all* human behavior. Furthermore, all decisions that lead to overt behavioral responses are based on internal (thus unobservable) and largely unspecified multidimensional sources of evidence, and they are influenced by strategic and motivational considerations. Although the many problems faced by the researcher seem obvious for complex behaviors, the same problems exist with supposedly simple tasks such as the detection of a stimulus, where the observable physical dimensions are anything but isomorphic with unobservable sensory, perceptual, or cognitive dimensions. A major simplifying strategy throughout psychology is to model behavior and decision-making statistically, with the actual observer viewed as a less than perfect, and sometimes biased, version of a computer or an ideal observer modeled to process the same information. If the multidimensional evidence were precise (without variability or uncertainty), the decision would not be statistical. Because the processes being studied are characterized statistically, it makes sense to use descriptive statistics to characterize the behavior of the processes and to use inferential statistics to test hypotheses about the processes.

Statistical decision theory, developed by Wald (1950), provides models of optimum decision strategies for dealing with complex and uncertain sources of evidence. Statistical decision theory is not a model of evidence, but a model of optimum decisions based on statistically defined (thus, imprecise, noisy, or variable) evidence. When the distribution of evidence is known, the optimum statistical decision is a Bayesian solution that compares the probability that each alternative decision is true or valid. The optimum decision variable, then, is the probability ratio; and the optimum solution is the choice of the alternative with the greatest probability. Wald (1950) developed statistical decision theory to address statistical decision problems that include conditions in which "an a priori distribution cannot be postulated, and, in those cases where the existence of an a priori distribution can be assumed, it is usually unknown to the experimenter and therefore the Bayes solution cannot be determined" (Wald, 1950, p. 16). These are the most common conditions faced by researchers in psychology and in many other disciplines. In his formal derivation of statistical decision theory, Wald (1950) demonstrated that most optimum decision strategies are based on likelihood estimates of the truth or validity of the alternative decisions, with the likelihood estimates derived from the available evidence. The optimal decision solution is functionally equivalent to a Bayesian solution and thus a ratio, but the ratio is now one of likelihood estimators rather than probabilities.[1] The optimal statistical decision variable is the likelihood ratio.

Statistical decision theory was intended to provide models of experimentation (Wald, 1950), but the basic principles of the theory can provide models of human decision processes equally well, and this is the basis for SDT. SDT was originally derived from statistical decision theory (Wald, 1950), with the goal "initially aimed at unifying a considerable area of psychoacoustics within the statistical-decision framework" (Licklider, 1964, p. 119). Although SDT "stems from the direct application of Statistical Decision Theory to the detection of auditory signals embedded in noise (Peterson & Birdsall, 1953) . . . it was immediately recognized that this approach could be easily applied to other sensory domains" (Kadlec & Townsend, 1992, p. 182), as well as to a wide range of discrimination, recognition, and classification tasks in which the stimuli are either simple or complex (see, e.g., Green & Swets, 1966; Swets, 1964) and performance is assumed to reflect either simple low-level sensory processing or complex higher level (cognitive) processing, such as memory (e.g., Egan, 1975) or even medical diagnosis (e.g., Swets, 1996). In keeping with this wide range of decision applications, Macmillan and Creelman (1991, p. 25) broadly define "detection theory" to "mean a theory relating choice behavior to a psychological decision space."

The decision variable. A central concept in statistical decision theory is the nature of the decision variable, and the original, formal conceptualization of the SDT decision variable stems directly from that variable in statistical decision theory. In its original, formal derivation, SDT "specifies as the optimal decision function either likelihood ratio . . . or some monotonic function of likelihood ratio" (Swets, Tanner, & Birdsall, 1961; p. 47 in Swets, 1964).[2] For the present discussion, we will temporarily continue to treat the decision variable as the likelihood ratio. As in statistical decision theory, a single observation can be represented "as a point in an *m*-dimensional space." (Swets et al., 1961; p. 8 in Swets, 1964), where *m* refers to the number of dimensions in the observation. For each dimension, an ideal statistical observer evaluates the likelihood that the observed value has arisen from each of the possible decision alternatives or hypotheses (e.g., in a detection task, the likelihood that the specific dimensional value arose when the signal was present or absent). These evaluations are combined statistically to produce an overall likelihood that the specific *m*-dimensional event has arisen from each of the alternative decision conditions. From the perspective of a statistical decision process, the optimal decision should reflect the choice of the hypothesis with the largest likelihood and thus should be based on a comparison of the likelihood of the multidimensional observation arising from each of the alternative hypotheses.

For a decision between one pair of alternative hypotheses (a binary choice), this likelihood comparison can be expressed most effectively as a single variable, the ratio of the contrasted likelihood values, and thus as the likelihood ratio (e.g., Irwin & Hautus, 1997; Noreen, 1981; Wald, 1950). A likelihood ratio decision variable is unidimensional, but that unidimensional variable reflects an evaluation of evidence distributed across many (i.e., *m*) dimensions. For a given observation, the value of the decision statistic reflects the evaluation by the observer of the relative likelihood that the set of *m* dimensional values might have occurred if each of the contrasted decision outcomes had been valid. For example, in a recognition memory task, the old/new decision would be based on the relative likelihood that the strength of *m* dimensions of evidence elicited by the presented stimulus would have occurred if the stimulus had been, versus had not been, previously presented. The unidimensional likelihood-ratio decision variable reflects the relative strength of the multidimensional evidence (as evaluated by the observer) favoring the two possible decision alternatives. The unidimensional nature of the decision variable for a binary choice thus does *not* make SDT incapable of describing the statistical nature of complex, multidimensional sources of evidence (e.g., word–nonword lexical decisions), despite claims to the contrary (see, e.g., Norris, 1995).

General concepts of sensitivity and decision criterion. The distributions of observed (or expected) likelihood values for alternative decisions reflect the statistical properties of the evidence as evaluated by the decision maker (e.g., in a simple detection task, the distribution of noise and signal-plus-noise).[3] The statistical properties of evidence distributions across the decision variable (e.g., difference in central tendency [e.g., mean] relative to dispersion [e.g., variance]) set a limit on the maximum accuracy that the observer can achieve, and thus they determine maximum ability or sensitivity. The decision process that we have described is directly analogous to inferential statistics, where the decision variable is some measure of evidence (t, F, etc.), and the distributions of evidence for the pair of alternative decision outcomes determine the sensitivity of the statistical test. Furthermore, the ability to distinguish statistically between the alternatives reflects the magnitude of the difference between the central tendency of evidence for the alternatives and the variability in evidence.

The other property of decision models is the decision rule or criterion. A decision rule or decision process that always selects the alternative with the larger likelihood is described as *unbiased*; the decision is based only on the available evidence. When the alternative decisions are equally probable, the unbiased decision rule results in maximum accuracy. Expressed as a likelihood ratio decision variable, unbiased performance reflects a decision criterion of 1, accepting the numerator decision outcome when the likelihood ratio is greater than 1, and accepting the alternative outcome when the ratio is less than 1. Any other decision rule (criterion) represents bias in favor of one of the alternative decisions. Without getting into some subtle mathematical distinctions (see, e.g., Egan, 1975; Wickens, 2002) (and delaying a more detailed discussion of other decision criteria until later), we can say that the SDT criterion statistic, β, is functionally this likelihood ratio decision criterion. Thus, the observer is unbiased when $\beta = 1$ and is biased for other values of β.

The analysis of variance (ANOVA) is a common example of using a biased criterion decision rule for an evidence-ratio decision statistic. The F-ratio decision statistic reflects evidence for "noise" alone (denominator) versus an effect of a manipulation superimposed on the noise. For the ANOVA, the scientific community has accepted a criterion ratio value that, on the basis of the expected distribution of the decision statistic, specifies the maximum false alarm rate (the probability of false positive decision) at .05 for a "significance" decision. The criterion value of the (F statistic) evidence ratio is not 1; it is biased in favor of incorrectly accepting the null hypothesis.

Specific Models of SDT: Assumptions About Distribution of Evidence

Other than indicating that statistical decision theory was developed to model optimal decision strategies for situations in which the underlying distributions are unknown, we have not discussed assumptions about the distribution of evidence. As a general theory derived from statistical decision theory, SDT does not make assumptions about the underlying distribution. In contrast, specific models of SDT are based on specific distributional assumptions. The most common model of SDT is the Gaussian equal variance model (hereafter, G-SDT), which is based on the assumption that the two evidence distributions are Gaussian and equal in variance.[4] This assumption operationalizes a theoretical statistical theory into a model that can be readily used in the study of human decision processes. Specifically, it allows one to calculate two separate descriptive statistics, d' and β (or some other criterion statistic; see immediately below). These respective statistics reflect sensitivity and criterion bias in the manner described earlier for statistical decision theory and for inferential statistics. The descriptive statistic for sensitivity, d', is the z-score distance between the means of the evidence distributions and thus reflects separation of the distributions in units of variance. The criterion statistic, β, the ratio of the ordinates of the two distributions at the criterion, is in principle an instantiation of the likelihood-based decision rule adopted by the observer (McNicol, 1972). If the assumptions of G-SDT are valid, the descriptive statistics for sensitivity and criterion will be orthogonal.

For G-SDT, a number of alternative descriptive statistics for criterion have been developed. The original statistic, β, is excellent for use in mathematical models of decision processes; it represents the Bayesian solution (Wald, 1950), and it is not based on any assumptions about underlying distributions. However, β is probably a poor choice for a descriptive statistic in empirical research. By definition, β is a ratio scale, and the appropriate measure of central tendency for a ratio scale is the geometric mean. Because most behavioral researchers have learned to use descriptive and inferential statistics (e.g., arithmetic mean, ANOVA) that assume an underlying interval metric, it is not unusual for researchers to report the (arithmetic) mean value of β and to use standard inferential statistics. Macmillan and Creelman (1991) recommend the use of c, an alternative descriptive criterion statistic. This statistic, with standard G-SDT distribution assumptions, reflects the position of the criterion in (interval scale) z-score units relative to the midpoint between the two distribution means. The unbiased criterion, corresponding to $\beta = 1$, is at $c = 0$. There are a number of alternative descriptive criterion statistics (see Irwin, Hautus, & Francis, 2001; Macmillan & Creelman, 1991; Snodgrass & Corwin, 1988).

Uniqueness of Gaussian assumptions. A number of modern researchers appear to be extremely concerned over the possibility that the assumptions of G-SDT, and, specifically, Gaussian distributions with equal variance, may be violated. Some researchers thus favor using "nonparametric" SDT descriptive statistics (discussed later) or question the validity of conclusions based on the use of G-SDT measures (e.g., Norris, 1995). It is true that, if the underlying distributions exhibit significant differences in variance, the descriptive statistics for sensitivity and criterion will not be fully independent, and large differences in criteria can result in large changes in sensitivity. How much concern should there be about possible violations of the assumptions of the G-SDT model? We will answer this question by looking at the assumptions of typical inferential statistics.

Statistics are an essential component to all empirical research in many sciences, including psychology. Whenever one uses statistics, whether inferential or descriptive, one is making implicit, if not explicit, assumptions about underlying distributions. Researchers however, seem to be more aware of the nature of these assumptions when they are considering some descriptive statistics (e.g., d', but not either mean or standard error) than when they are using inferential statistics, even when the assumptions are essentially equivalent. Thus, one can easily find researchers reporting significance based on common statistical tests while also expressing significant concern over the assumptions underlying the use of d' and either β or c. However, "the chi-square distribution rests directly on the assumption that the population is normal" (Hays, 1963, p. 353). Since "the F variable is the ratio of two independent chi-square variables, each divided by its degrees of freedom. . . . the F distribution also rests on the assumption of two (or more) normal populations" (Hays, 1963, p. 353). Furthermore, the error variance in each population must be both independent and equal (e.g., Hays, 1963). The Student t statistic, the difference between distribution means scaled in standard error units, is based on similar assumptions. Specifically, "in order to find the exact distribution of t, one must assume that the basic population distribution is normal," and the actual t distribution must be "a unimodal, symmetric, bell-shaped distribution having a graphic form much like a normal distribution" (Hays, 1963, p. 305) that, in its limit, becomes a normal distribution. Thus, the researcher always needs to consider whether the assumptions of equal-variance Gaussian distributions are reasonable; but that consideration is no more or less important for use of inferential statistics (statistical tests) or descriptive statistics (mean, variance, d', etc.).

Common Misconceptions About SDT and the SDT Decision Variable

The formal foundation in statistical decision theory provided a solid theoretical basis for SDT, but also appears to have contributed to many modern misconceptions about SDT. The nature of the decision variable is at the heart of both statistical decision theory and SDT, and many of the modern misconceptions about SDT appear to reflect problems in researchers' understanding of the decision variable. In its derivation within statistical decision theory, the decision variable is probably overspecified as a statistical concept and underspecified as a concept that can be easily understood and operationalized. Our brief summary of statistical decision theory should provide a solid foundation for identifying and correcting modern misconceptions about the decision variable.

In many modern textbooks, the SDT has become G-SDT, and the decision variable is characterized as a single, simple sensory dimension. There are probably two reasons for this characterization. First, because the decision variable of statistical decision theory is not easy to understand, it is a challenge for undergraduate textbook authors to explain SDT at more than a superficial level. Second, not only does the name of the theory (signal detection theory) include the term "detection," but the initial development and validation of SDT was meant for a narrowly defined task, the detection of an auditory signal embedded in "white" (Gaussian) noise, with the dominant variability or uncertainty of that specific application arising from the noise. The application of principles from statistical decision theory to this specific detection task actually contrasts with Wald's theory (and with the many applications of SDT to complex decision situations that began within a few years after SDT's origin; see, e.g., Egan, 1975; Swets, 1964, 1996), where the distribution of evidence is not known. However, one effective test of a theory that claims to make no assumptions about underlying distribution is to evaluate its performance when the distribution is known. Likewise, the initial test of a model that makes specific distributional assumptions is to evaluate its performance when the assumptions are known to apply. (We will return to this question when we evaluate "nonparametric" SDT measures.)

In the original detection task, two distributions reflected noise alone and signal added to noise. In that specific application, the decision variable did reflect sensory information, the two distributions of evidence were each Gaussian, and the difference in distribution means did reflect the difference in sensory information or magnitude associated with the addition of the signal to the noise. This detection task was an excellent choice for a tool to validate the application of statistical decision theory principles to the modeling of human decision performance. Unfortunately, this specific application of G-SDT has become synonymous with SDT. The decision variable has been equated largely with *sensory* activity. The G-SDT sensitivity statistic, d', has been taken to reflect only *sensory* ability. This unfortunate metamorphosis does appear to address the problems associated with the need for a characterization of the SDT decision variable that is understandable to undergraduates, but it actually adds new problems associated with an inaccurate specification of SDT and its decision variable.

There are many examples of this simplified, incorrect characterization of the SDT and its decision variable. Coren, Ward, and Enns (1999), Levine and Shefner (1991), and Schiffman (2000) all describe the decision axis (and thus the decision variable) as being the level of sensory activity, with the label "noise" (originally intended to reflect statistical variability) being interpreted literally and thus described as being based largely on spontaneous neural activity. For example, "the noise referred to by signal detection theory . . . is an ever-varying level of neural activity of a type exactly like the nervous system's response to the signal. There is a background level of activity in the nervous system, and sensory signals are superimposed on this activity" (Levine & Schefner, 1991, p. 27). Likewise, "signal detection theory assumes that any stimulus must be detected against the background of endogenous noise in our sensory systems" (Coren et al., 1999, p. 20). Some authors allow that, in addition to (internal) noise in the nervous system, noise may also come from an external source (e.g., Haberlandt, 1994; Payne & Wenger, 1998). When higher level processes are mentioned, it is in terms of additions to the noise. Specifically, noise reflects, "in addition to spontaneous sensory-neural activity, . . . the unpredictable, random effects of fatigue and the effects of nonsensory response biases such as the observer's fluctuating level of attention and motivation to the detection task" (Schiffman, 2000, p. 28). Likewise, "these fluctuations in noise level are caused by physiological, attentional, and other variables in sensory and perceptual systems of the observer as well as by random fluctuations in the environment" (Coren et al., 1999, p. 22). The decision axis is thus labeled "level of sensory activity" (Schiffman, 2000, p. 28), "sensory activity level" (Coren et al., 1999, p. 23), "level of activation of sensory system" (Levine & Shefner, 1991, p. 28), "sensory evidence" (Payne & Wenger, 1998, p. 82), "magnitude of sensory impression" (Kantowitz, Roediger, & Elmes, 2001, pp. 170–174), "intensity" (Haberlandt, 1994), or "perceptual effect." Rare exceptions specify the decision axis as "what the subject experiences on each trial" (Goldstein, 1999, p. 557; 2002, p. 587), or simply do not label the axis (Jahnke & Nowaczyk, 1998, p. 75). Sekuler and Blake even consistently change the name of SDT to "sensory decision theory," allowing that the theory is "sometimes called signal detection theory in recognition of its origins in electrical engineering" (Sekuler & Blake, 1990, p. 497; 2002, p. 601). Undergraduate textbooks are thus inaccurate sources of information about SDT, and they should not be cited as an SDT source in research articles (e.g., Norris, 1995).

Clearly, the decision variable is being described in very limited, concrete terms, and the concept of noise is taken almost literally, rather than as reflecting the uncertainty or variability associated with statistical processes. Furthermore, some researchers have equated sensitivity or ability (the distribution of the decision variable) with sensory or perceptual processes, and criterion placement with postperceptual processes. In the semantic priming literature, this partitioning of sensory/perceptual decision statistic processes and postperceptual criterion processes has then been used in arguments about perceptual encapsulation and cognitive modularity (e.g., Rhodes, Parkin, & Tremewan, 1993), sometimes with the qualification that "the claim that changes in measured sensitivity are a direct reflection of changes in the sensitivity of some early perceptual process is only true under the specific set of assumptions made by signal detection theory" (Norris, 1995, p. 936; see also Norris, McQueen, & Cutler, 2000). Thus, in parts of the modern psychological literature, SDT is believed to be only a very limited model of sensory processes, with this belief used to support broad, theoretical positions. SDT is *not* a model of sensory processing, and the decision variable is not a simple, concrete physical or sensory dimension. Rather, SDT is a general model of decision processing of evidence, and "the decision variable is essentially unobservable" (Laming, 1986, p. 39). SDT definitely "does not require one to be specific about the axis on which the decision is made" (Wickens, 2002, p. 150).

Specifying the decision variable. Even with an understanding of the actual nature and goals of SDT, the typical researcher still seems to want an answer to the question, "What is the decision variable?" This question has really been answered earlier. The decision axis provides "some measure of evidence for (or against) a particular alternative, . . . formulating it using a decision strategy known as likelihood-ratio testing" (Wickens, 2002, p. 150). SDT does not explicitly assume that the decision variable is the likelihood ratio, only that it is a monotonic transform of the likelihood ratio. This means that as the value of the decision variable is increased, the weight of evidence will shift in a systematic fashion from strongly favoring one alternative, to less strongly favoring that alternative, through favoring neither alternative, and then from weakly to strongly supporting the alternative variable. Thus, in the modeling of false memory decision processes using SDT (Wixted & Stretch, 2000), or using both SDT and statistical decision theory (Wickens & Hirshman, 2000), the decision variable is appropriately labeled "strength of evidence." A greater specification of the nature of the unobservable decision variable depends on the research design, on the processing model being employed by the researcher, and on assumptions about the nature of the relevant evidence. For example, in studies of recognition memory, research manipulations are often designed to alter variables that are assumed to be reflected in the concept of either familiarity or perceptual fluency. With the further assumption that the evidence is restricted to the specific construct being manipulated, the decision axis is labeled "familiarity" or "perceptual fluency." This is appropriate if the assumptions about the nature of evidence and the relevance of the research manipulations are accurate.

The full implications of this discussion will become even clearer when we turn again to common inferential statistics (e.g., ANOVA, *t* test) and descriptive statistics (e.g., mean, median, variance) for examples of generic evidence variables. The simple use of an ANOVA or a *t* test does not define the nature of the evidence being evaluated. The theory being tested, the design of the experiments, and the assumed nature of the independent and dependent variables determine the labels that define the nature of the evidence and allow interpretation of the inferential statistical analyses. SDT is no more a theory of sensory or perceptual process, or of familiarity or perceptual fluency, than is an ANOVA or a *t* test.

Independence of Sensitivity and Bias

Everyone seems to understand that if there are severe violations of the assumptions of G-SDT, then sensitivity and criterion statistics will not be independent, and, as a result, changes only in criterion in the absence of changes in the underlying evidence distributions will result in changes in measures of both criterion and sensitivity (e.g., β and d'). Stated more generally and in the positive, if the assumptions of G-SDT are valid, or if "nonparametric" SDT descriptive statistics are truly nonparametric, then the model's descriptive statistics for sensitivity (respectively, d' and A') will be independent measures of the model's descriptive statistics for criterion (respectively, β or c and B''); the nonparametric measures are addressed in a separate section below. The meaning of independence however, is easily misunderstood. In principle, the independence of sensitivity and criterion measures should mean that neither descriptive statistic is dependent on or a function of the value of the other descriptive statistic. The independence of these two descriptive statistics is sometimes (at least implicitly) interpreted as allowing an experimental manipulation to alter either sensitivity or criterion, but not both descriptive statistics (e.g., Rhodes et al., 1993). In fact, neither of these statements is completely accurate. This is both because the concept of criterion is more complex than is typically realized and because the relationship between the descriptive sensitivity and criterion statistics is not necessarily symmetric or reciprocal.

The relationship that is simple to understand is that the value of the criterion should not affect sensitivity. In statistical decision theory and in SDT, there is effectively an ordered relationship between evaluating the evidence to compute the value of the decision variable and the use of the decision criterion. A change in criterion should not alter the distribution of evidence that defines the decision variable, and it is the distribution of evidence that limits the optimal decision ability.[5] Descriptive statistics for sensitivity are designed to accurately reflect the magnitude of this ability, but such statistics need to be based on specific assumptions about the distribution of the decision variable. Thus, if the assumptions of G-SDT are reasonably valid, the value of d' reflects the underlying ability defined by the distribution of available evidence, and it should therefore be constant across changes in β or c. This aspect of independence, alone, is important, since it allows the researcher to identify conditions that involve only criterion changes.

The consequences of change in the underlying evidence distributions that complicate the nature of independence. A variety of decision strategies are available to the decision maker, and several different descriptive criterion statistics are available to the researcher (e.g., β and c), with each related in a different way to the evidence distributions. Thus, the relationship between a descriptive decision statistic and the *different* types of descriptive criterion statistics is not simple. As a result, changes in the evidence distribution that produce a change in sensitivity may not alter the criterion employed by the observer, but they will often alter the value of the criterion statistic computed by the researcher. A basic understanding of the various descriptive statistics for criteria and the nature of alternative criterion strategies available to the decision maker will allow an understanding of possible relationships between descriptive statistics for sensitivity and criterion.

Different criteria. A statistically optimal decision maker, as described earlier, will adopt a decision criterion based on the likelihood ratio. The descriptive statistic, β, reflects the value of the likelihood ratio or the slope of the receiver operating characteristic (ROC) curve (in simple probability coordinate space) at the criterion. The descriptive statistic, c, is the distance of the criterion from the equal bias point measured in z-score units (Macmillan & Creelman, 1991) and is *not* a direct transformation of β. For fixed evidence distributions (i.e., constant d'), there is an isomorphic mapping between β and c, but that mapping is *not* invariant across changes in the evidence distributions that result in a change in d'. Thus changes in the evidence distributions that result in a change in d' will also alter the mapping between β and c.[6] If the observer establishes a criterion that is constant for one of these two criterion statistics, then a change in the evidence distributions will not alter the observer's criterion (we have stated that it is held constant) or the value of the descriptive statistic for that specific type of criterion, but it will alter the descriptive statistic for the other criterion. For example, if d' changes from 1.0 to 2.0, and then to 3.0 for an observer whose decision strategy maintains c at 0.5, the values of β will be 1.65, 2.72, and 4.48. The researcher, seeing a change in both the sensitivity and criterion statistics, might even conjecture that the apparent change in criterion caused a change in sensitivity (e.g., Norris, 1995). For a more detailed illustration, Stretch and Wixted (1998) provide an excellent analysis of the effects of changes in d' on alternative confidence rating decision strategies in recognition memory with six response categories, and thus five criteria. They analyze three different decision models: the likelihood ratio model with criteria at fixed values of β; the lockstep model with criteria at fixed values of c; and the range model with an equal partitioning of distance between endpoint criteria with constant values of c.

The β and c descriptive criterion statistics, as well as the decision criterion strategies reflected in these statistics, are characterized by relationships between the two evidence distributions being evaluated. However, a perfectly logical decision strategy is to specify the criterion in terms of constant error rate (or accuracy) for one of the two evidence distributions or decision outcomes. For example, in inferential statistics, the decision criterion holds constant the probability of a Type I error (e.g., .05). A different decision strategy is to use a criterion that holds constant the probability of a Type II error. In principle, a researcher's descriptive criterion statistic could be the probability (or z-score value) for either a Type I or a Type II error. Modifying our earlier example that compared β and c, we now will have our hypothetical decision-maker adopt a criterion that holds constant the probability of a Type I error. A change in the underlying evidence distributions will alter d', and a researcher investigating the decision-maker, using any criterion measure other than P(Type I error) [e.g., the researcher might use β, c, c', or P(Type II error)], will report a change in criterion or bias, yet the observer will have held his/her own criterion constant.

From this analysis, it should be clear that, unless the criterion statistic measured by the researcher and the criterion adopted by the observer are based on the same principles, a change in the evidence distributions that alters measured sensitivity (e.g., d') will typically result in a change in measured criterion or bias. With such findings, the researcher can accurately report that, for the type of criterion defined by the researcher, there has been a change in the criterion. However, the researcher cannot make any stronger statement. The empirical finding of changes in both sensitivity and criterion (bias) can mean any of several things. There might be a severe violation of the assumptions of G-SDT that has resulted in a lack of independence in descriptive statistics, so that a change in criterion has altered d' in the absence of any change in underlying distributions of evidence. Alternatively, the assumptions of G-SDT might be reasonably valid, in which case the findings could reflect a change in the underlying distributions (change in d') coupled either with an actual change in the observer's criterion or with the researcher's having used a criterion statistic that is different from the observer's invariant decision strategy. To distinguish among these alternatives for tasks that require the placement of multiple criteria, Stretch and Wixted (1998) appropriately suggest analyzing the results in ROC space (described below). Finally (for reasons that will be developed next), the findings may reflect a flaw in the experimental design. Thus, just as with inferential statistics, one needs to be careful to consider possible procedural causes for patterns of obtained results.

Procedural considerations for independence. Beyond the specific assumptions of the particular SDT model being employed, there is another important statistical aspect of independence for d' and either β or c. In the computing of the SDT descriptive statistics, there are two independent evidence distributions. In G-SDT (and other operational models of SDT), the value of one statistic [e.g., p(H or Hit) or p(F or False alarm)] is computed from each of these independent distributions, and the two statistics are independent. The SDT descriptive statistics for sensitivity and criterion are effectively the sum and the difference of these two independent variables (e.g., Macmillan & Creelman, 1991). Since the sum and the difference of two independent random variables are themselves independent of each other, the two SDT descriptive statistics are independent of each other. The basis for this additional statistical consideration is not the nature of the distributions, but rather the independence of the hit and false alarm probability estimates. If one computes d' and β (or c) for several within-subjects conditions, the descriptive statistics will be independent if both p(H) and p(F) are independently determined across conditions, but they will not be independent if either probability is common across conditions. Several examples will illustrate problems that can occur when there is a failure to appreciate the importance of this computational component to independence.

In the false memory literature, there is typically the finding of a higher false alarm rate to critical lures than to other previously nonpresented items (e.g., Roediger & McDermott, 1995, 1999). Miller and Wolford (1999) argued that the difference in false alarm rate is due to a change in criterion, not a change in sensitivity, and thus does not reflect a false memory. Both Wixted and Stretch (2000) and Wickens and Hirshman (2000) present decision models that address the criterion change explanation in terms of the relationship between the underlying distributions and descriptive statistics for sensitivity and criterion. Figure 1 illustrates some basic principles that reflect both general procedural considerations about independence of the two measures that we have just discussed and the specific problem with Miller and Wolford's conceptualization. The solid curves in Figure 1 represent the distribution of evi-

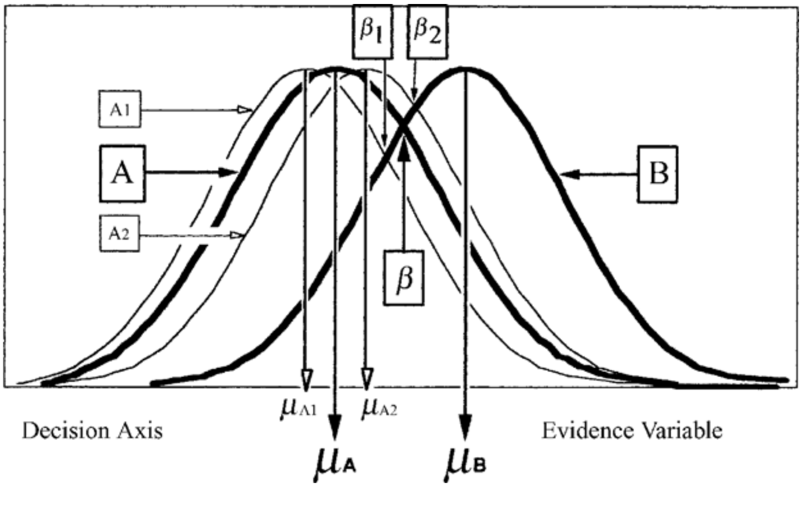

Figure 1. Illustration of independence problems when two conditions (A_1 and A_2) are evaluated together against a common decision alternative (B). For the decision-maker, the A_1 and A_2 distributions, which the researcher treats as separate, are functionally a single, combined distribution, A. When the decision-maker sets a fixed criterion, β, defined by the relationship between the A and B distributions, the researcher will measure four descriptive statistics (two values of d' and two values of β) that are not independent of each other.

dence for two alternative event classes, here labeled A and B. For a recognition memory task, A and B represent new and old memory items, respectively. The decision axis represents the relative amount of evidence favoring the decision alternatives. The observer knows only that there are two possible sets of items and thus is dealing conceptually with only two underlying distributions of evidence. In our illustration, the observer uses an unbiased criterion, and thus β, at the intersection of the two distributions, has a value of 1.0. Sensitivity is reflected by the difference in the means of the two distributions, and thus the distance between μ_B and μ_A (in Figure 1, $d' = 1.3$). The criterion statistic is β, and thus the ratio of the ordinates of the two distributions at the criterion ($\beta = .323/.323 = 1.0$). Probability estimates for hits and false alarms are from independent distributions, and thus independence between the criterion and sensitivity statistics requires only that the assumptions of the SDT model be reasonably valid.

The researcher, however, has treated the critical lures differently from other nonpresented items. Thus, the distribution of A actually consists of two distributions, with a mean of μ_{A1} for noncritical nonpresented items and μ_{A2} for critical nonpresented items.[7] From the perspective of the observer, the fact that there are two different underlying distributions for A is irrelevant, with the observer's criterion (at $\beta = 1$) based on the combined distribution. From the perspective of the researcher, A_1 and A_2 are different distributions. The computed d' sensitivity measure for the two nonpresented items is based on the distance between the mean of each distribution and μ_B. The resulting d' values are 1.0 for A_2 (nonpresented critical lures) and 1.6 for A_1 (noncritical nonpresented lures). Because there is a difference in location of the distribution means for A_1 and A_2 (the mean for distribution B is constant), the effective value of the criterion (at the overall location of β) is decreased ($\beta = .323/.375 = 0.86$) for A_1 and increased ($\beta = .323/.254 = 1.27$) for comparisons based on A_2. It would thus appear that the critical nonpresented items have resulted in both reduced sensitivity and a stricter criterion.

Our current example has really described only the basic point (presented earlier) that changes in sensitivity can, and typically do, alter criterion measures. However, there is a more important lesson about independence in this example. The researcher may be interested in two values of sensitivity, or two values of criterion, or all four of these values. However, with only three independent underlying distributions, the four descriptive statistics are not independent of each other. In fact, if there is any difference between μ_{A1} and μ_{A2}, there will be an equal (z-score) difference in the values of d', and the criterion value of β (or c) will always be more liberal for the condition with the higher value of d'. This is because one of the two parameters in the computation of these statistics is common across the two values. Here, the hit rate z score is common across the two values of d' (and also the two values of the criterion statistic, c), and the ordinate of the hit rate is the numerator in the computation of both values of β.

Hicks and Marsh (2000) provide an example of an equivalent lack of independence in their reported measures for three recognition memory experiments, but with a reversal of the conditions described in Figure 1 (false alarm rate is common, with two different hit rates). Each memory item was presented in one of two conditions that were assumed to result in different levels of encoding. Recognition memory was then evaluated using a set of stimuli that combined one set of nonpresented foils (to determine false alarm rate) and a mixture of the two sets of previously presented items. As in our Figure 1 example, the observer still has only two distributions (old and new items) and adopts a single criterion, but (in contrast to the Figure 1 example) the researcher determines one false alarm rate and two different hit rates. Hicks and Marsh report differences in hit rate and similar differences in d', with β becoming more lax for the larger d' condition, discussing possible encoding-based reasons for the change in criterion. From our earlier analysis, we know that for conditions with one of the component parameters in common (here, false alarm rate), the other component parameter (here, hit rate) and the computed descriptive statistics, d' and β, are not independent. Hicks and Marsh also report RT, finding that RT is faster for conditions with higher d' values. If one posits that RT should be faster for easier conditions, RT should also be at least partially correlated with these other statistics. Thus, all of the reported descriptive statistics across encoding conditions are correlated, and separate discussions of causal factors for differences in each measure are inappropriate.

Roles of Uncertainty

The concept of uncertainty is relevant to both the evidence distributions of the decision variable and the placement of the criterion, and uncertainty is not the exclusive attribute of either component to the decision process. In statistical decision theory, the distribution of evidence is unknown and is specified statistically, thus implying considerable uncertainty. The statistically defined evidence that is the basis of the SDT decision variable does not just appear, but rather is the result of the individual observer's evaluation of the nature and quality of the available evidence; the distribution of evidence thus reflects both the evidence available and the observer's knowledge about possible evidence. Therefore, the statistical distributions of evidence and the decision variable that reflects the statistical distribution can be expected to differ across individuals. Furthermore, the separation of the contrasted evidence distributions is altered when the individual observer gains knowledge about the dimensions of evidence that are, and are not, relevant to the decision, and when there is improved knowledge about the distributions of evidence values for each of the relevant dimensions. A knowledgeable, and thus less uncertain, observer will exhibit greater ability or sensitivity than a naive, highly uncertain observer. Therefore, uncertainty is an important component of sensitivity.

Ideally, the observer implements a decision rule by establishing a fixed criterion. If the criterion is defined by a fixed likelihood ratio, one might assume that the criterion could be invariant, but it is more realistic to assume some degree of variability or uncertainty even in a likelihood ratio criterion (i.e., β). If the criterion is specified as a fixed value of c, or a constant Type I or Type II error probability, then, by its very nature, the criterion requires that the observer possess some knowledge about the underlying distributions that define normalized (z-score) distances for c and cumulative distribution functions that define the Type I or II error rate. Depending on the accuracy and precision of such knowledge, there will be uncertainty in the criterion specification. Uncertainty in the criterion is equivalent to variability in the criterion.

In at least a portion of the literature on perceptual encapsulation or cognitive impenetrability, the notions of independence of the descriptive statistics and the role of uncertainty have sometimes acquired special meaning. Here, there is an implicit, and sometimes explicit, assumption that sensory and perceptual processes are automatic, and there is no place for uncertainty in automatic processes. Thus, uncertainty is a property only of higher level or cognitive processing (e.g., Norris, 1995; Norris et al., 2000). In his checking model, for example, Norris argues that semantic priming is purely postperceptual and is modeled by a temporary altering of the criterion or threshold of words. The conjectured, purely postperceptual, changes in criterion alter uncertainty, and the changes in uncertainty result in changes in SDT measures of sensitivity without altering the decision variable. In a perceptually encapsulated, cognitively impenetrable system (e.g., Norris et al., 2000), recognition criterion "changes operate in a manner which leads to changes in d' as well as beta. This is due to the fact that the priming . . . almost inevitably lead[s] to a reduction in stimulus uncertainty" (Norris, 1986, p.126). The assumption that uncertainty is only a property of the decision criterion has been combined with the incorrect assumption (discussed earlier) that d' reflects only sensory or perceptual processing (e.g., Norris, 1995; Rhodes et al., 1993). The conjoined assumptions then are used in the debate about whether changes in uncertainty that alter d' represent evidence against perceptual encapsulation (Rhodes et al., 1993) or evidence that the multidimensional nature of cognitive processes are too complex to be evaluated using G-SDT (Norris, 1995; Norris et al., 2000).

Criterion variability and sensitivity. G-SDT alone has no inherent ability to distinguish between uncertainty associated with the evidence variable and uncertainty associated with the criterion. Both forms of uncertainty represent variability in the decision process, and in G-SDT variability is functionally lumped together as the denominator in the theoretical formula for d' (e.g., Green & Swets, 1966). This is because d' is a z-score distance, and the z-score transformation is a normalization relative to variance. In principle, this pooling of variance should be a concern, but the effect of criterion variability on the value of d' may not be a great concern under many conditions. In a broad sense, there is equivalence in d' computed for a single subject with criterion variability across trials and d' computed from data pooled across subjects who are identical in underlying sensitivity but who differ in criterion placement. A theoretical analysis of the latter conditions by Macmillan and Kaplan (1985, p. 185) indicates that criterion variability "underestimates true average d' to only a small degree in most cases." Finally, with the appropriate procedures and a G-SDT–based theoretical model, one can evaluate the relative contribution of different sources of variability (e.g., Berliner & Durlach, 1973; Braida & Durlach, 1972; Durlach & Braida, 1969; Macmillan, Braida, & Goldberg, 1987; Macmillan, Goldberg, & Braida, 1988).

Nonparametric Model of SDT

It is not unusual in the modern research literature to find researchers who express the belief that "nonparametric" SDT measures (A' and B'') are functionally equivalent to parametric descriptive SDT statistics [d' or A_Z (defined below) and β, c, etc.], but are superior because they are nonparametric. For example, Rhodes et al. (1993, p.157) are fairly typical in describing these measures as being "nonparametric in that no assumptions about the signal and noise distributions are required (McNicol, 1972)." Masson and Borowsky (1998) describe the "nonparametric" quality of these measures in terms of not requiring strict adherence to the assumptions of equal-variance Gaussian distributions. Similar beliefs have been expressed by Norris (1995), Goldinger (1998), and many others. In this conceptualization, A' and B'' would seem to be ideal descriptive statistics that are far superior to d' and β, possessing attributes that seem to be (and, as we will see, actually are) too good to be true.

Macmillan and Creelman (1996) provide a brief history of "nonparametric" SDT measures, and this history allows us to understand the origin of the "nonparametric" label. The "nonparametric" SDT statistics are based on the ROC curve that plots hit rate (correct acceptance of the primary hypothesis) on the ordinate and false alarm rate (incorrect acceptance of the primary hypothesis) on the abscissa. (Figure 2 is a plot of ideal ROC curves for parametric G-SDT and nonparametric statistics that will be used below to illustrate problems with the nonparametric measure of sensitivity.) The complete operating characteristic for a given observer (with fixed underlying distributions along the decision variable) is the plot of all possible data points as criterion is varied. The (complete) theoretical ROC curve would be generated by systematically moving the criterion along the decision variable, computing the area in the upper tail of each underlying distributions (hits and false alarms) for each decision variable value. With the evidence variable distribution constant, the ROC curve represents constant sensitivity. Macmillan and Creelman (1996) attribute the initial conceptualization of a nonparametric measure based on the ROC curve to a paper presented by Green. Green (1964) reported, "in the 2-alternative forced-choice tasks all models agree that the percent correct is equal to the area under the Yes–No ROC curve."[8] In a two-alternative forced-choice (2AFC) task, the observer is presented with two stimuli that are examples of the two alternative decision states and is asked which of the alternatives matches the specified state (e.g., in a lexical decision task, the observer is presented with a word and a nonword stimulus and must decide which of the two stimuli is a word). The observer is assumed to independently evaluate a separate likelihood ratio (e.g., word/nonword likelihood ratio) for each of the presented stimuli and then to make a decision based on a comparison of the likelihood ratios. No assumptions are needed to compute the two likelihood estimators, and a decision process based on comparison of the likelihood ratios alternatives should tend to minimize bias (Green & Swets, 1966). Expanding on these ideas, Green & Swets (1966, p. 50) state, "the area under the yes–no or rating ROC curve . . . is distribution free, since no assumption about the form or character of the underlying distributions is made by calculating the area." Therefore, the area under a multipoint ROC curve is a nonparametric measure of ability or sensitivity that is easily interpreted because it corresponds to optimal percent correct. In a multipoint ROC curve, the shape of the curve is determined by the underlying distributions, whatever the nature of those distributions.

A' and B'' descriptive statistics. Pollack and Norman (1964; see also Pollack & Hsieh, 1969) developed a procedure for a descriptive statistic that reflects an estimate of the area under an ROC curve that is specified by only a single data point; a single data point in ROC space is not an ROC curve. Their measure, A', is the average area for the two linear ROC curves that maximizes and minimizes hit rate.[9] B'' was developed by Grier (1971) as a corresponding nonparametric, ROC-based measure of bias for a single data point. The computation of A' and B'' is nonparametric in the sense that the computation requires no a priori assumption about the underlying distributions. This does *not* mean that these measures are an accurate reflection of their theoretical origin (i.e., that A' reflects the area under a reasonable ROC curve) or that they are distribution-free measures. Finally, it does not mean that A' and B'' are independent of each other.

The first concerns about these nonparametric SDT descriptive statistics was actually quick in coming, though largely ignored. McNicol (1972) used computer simulations of possible theoretical distributions to evaluate these measures. He reported that the computational formula for A' "gives the same value as P(A) only if the observer is unbiased and does not have a tendency to give a larger proportion of S or N responses. The greater his response bias in either direction, the more $P(\overline{\text{A}})$ will underestimate the true area under the ROC curve" (p. 39). Snodgrass and Corwin (1988) independently reached the same conclusion. Thus, A' and B'' are not independent descriptive statistics, and, when there is significant bias in criterion placement, A' is not even an accurate measure of area under the ROC curve. "This means that $P(\overline{\text{A}})$ can only be used safely as a sensitivity estimate when response biases under the experimental conditions being examined are equivalent." McNicol (1972, p. 40). Despite this early (if somewhat obscure) warning, A' and B'' became popular as supposedly independent, nonparametric, distribution-free SDT measures (Macmillan & Creelman, 1996). In a few instances, the researcher may be aware of the McNicol warning and thus may add a qualification, such as, "McNicol notes that if there is a bias in either direction, A' will underestimate sensitivity" (Rhodes et al., 1993, p.157).

Macmillan and Creelman (1996) provide a more detailed analysis of A' and B''. They report that "area under the one-point ROC, as estimated by Pollack and Norman's (1964) method, is consistent with logistic distributions for low sensitivities and rectangular distributions at high levels; in neither case does it deserve the label 'nonparametric.'" (Macmillan & Creelman, 1996, p. 169). Logistic distributions are the basis for the α [or $\log(\alpha)$] and β_L, the sensitivity and criterion statistics in Luce's choice model (Luce, 1963). Rectangular distributions, evident at higher levels, are indicative of threshold behavior (Egan, 1975; Macmillan & Creelman, 1990, 1991, 1996). We suspect, however, that many researchers do not appreciate the importance of Macmillan and Creelman's conclusion. We will try a different, and hopefully simpler, approach to demonstrate that the "nonparametric" SDT statistics do make implicit assumptions about the underlying distributions.

Comparing A' and d'. An ROC curve is not some abstract entity that is devoid of underlying distributional attributes, as is implied by any descriptive statistics that are described as being nonparametric or distribution free. Instead, the shape of the theoretical isosensitivity contour reflects the explicitly or implicitly assumed nature of the distributions that define decision structure for hits (H) and false alarms (F). Just as the shape of the theoretical isosensitivity contour is determined by the assumed properties of the distributions underlying the measure being plotted, the shape of the empirical ROC curve reflects the actual properties of the decision variable distributions. One can (and should, but seldom does) test the validity of the equal-variance Gaussian assumptions of the d' statistic by evaluating the similarity between the theoretical and empirical ROC curve; if the two curves are equivalent, the assumptions of the measure are reasonable (e.g., Egan, 1975; Green & Swets, 1966; Pastore & Scheirer, 1974). One can also use ROC curves to compare the nature of the distributions underlying different descriptive statistics, as is done throughout the excellent texts by Macmillan and Creelman (1991) and by Egan (1975). Figure 2 provides this comparison between the descriptive statistics of A' and d'.

If A' is truly a nonparametric measure of observer ability, it must make no assumptions about the underlying distributions. The computational formula developed to compute area from a single point in ROC space, however, imposes implicit assumptions about the distribution that must exist to generate a set of all points in ROC space that exhibit the same value of A'. All points along an isosensitivity contour for A' yield the identical value of A', and no other points in ROC space generate that value of A'. Each point along the isosensitivity contour represents a different criterion, with magnitude of bias increasing with increasing distance from the negative diagonal. Ability, and thus the value of A', increases as the isosensitivity contour moves from the chance line (positive diagonal, where $A' = .5$) to perfect performance [the upper left point in ROC space; (F,H) = (1,1)]. Plotting A' isosensitivity contours for different values of A' allows us to examine the behavior of A' as a function of ability and criterion.

In Figure 2, sets of theoretical isosensitivity contours for A' and d' are plotted in ROC probability space, with hit and false alarm probabilities on the ordinate and abscissa, respectively. For each descriptive statistic, a theoretical isosensitivity contour is generated by holding sensitivity (here, A' or d') constant, solving the computational formula for hit rate as false alarm rate is systematically varied. The positive diagonal in the ROC space always represents chance performance ($A' = .5$ or $d' = 0$) and the negative diagonal (not shown) represents zero bias ($\beta = 1$, $c = 0$, $B'' = 0$). Since each d' contour is based on the assumption of equal-variance Gaussian distributions, one will obtain an equivalent isosensitivity contour when these underlying assumptions are reasonably met, and one will see deviation from the theoretical contour when the underlying assumptions are violated (see, e.g., Pastore & Scheirer, 1974). The validity of the underlying assumptions for d', or any other measure, can be tested by evaluating the correspondence between the empirical and the theoretical ROC curves.[10]

In Figure 2, the symbols with broken lines represent theoretical isosensitivity contours for A' values ranging from 0.5 (the positive diagonal chance line) to 1.0 (the upper left corner of the ROC space with coordinates of 1,1) in steps of 0.1. For each of these isosensitivity curves, the broken lines (with symbols) provide a systematic mapping of all possible hit and false alarm rates for the given value of A' across the full range of possible criteria. The A' isosensitivity curves indicate that A' does imply underlying distributions and thus is parametric. The nature of the distributions underlying A' is not known, but aspects of the behavior of this measure can be understood through comparison with d'. The isosensitivity contours for d' are plotted for the values of d' that have the identical hit and false alarm rates at the zero-bias, minimum-error criterion (negative diagonal; $\beta = 1.0$ and $c = 0$). For low levels of sensitivity, the A' and d' contours are equivalent. Thus, when sensitivity is poor (and bias is not extreme), both A' and d' reflect equivalent underlying distributions; both will be independent of bias when the assumption of equal-variance Gaussian distributions has been met. As sensitivity is increased, the isosensitivity contours continue to be equivalent when bias is low (around the negative diagonal in ROC space), but not for extreme bias. The difference at extreme bias is that the isosensitivity contour for d' (and most other measures) begins and ends at the extremes of the positive diagonal (at points 0,0 and 1,1); however, the geometric-based averaging principle used to estimate A' from a single point results in isosensitivity contours that begin with a hit rate that is greater than 0 and end with a false alarm rate that is less than 1.[11] Thus, when estimated from a single point, A' will have a lower value than an unbiased d' measure based on Gaussian distributions, as well as an unbiased α measure based on logistic distributions (Macmillan & Creelman, 1996), or a measure based on any other distribution (McNicol, 1972). Stated another way, because a constant value of A' for increasing bias (in either direction) requires that the hit rate will have to be artificially inflated (or false alarm rate artificially reduced), A' underestimates sensitivity.

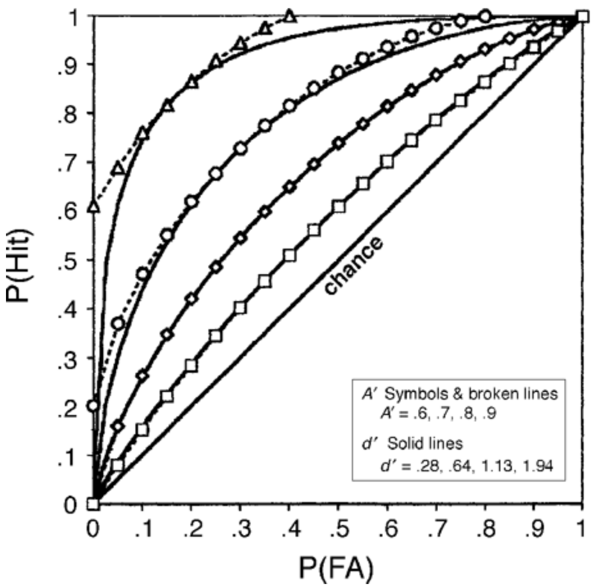

Figure 2. Isosensitivity contours for A' (symbols with broken lines) and d' (solid lines) in ROC space. Contours for A' are in .1 increments from $A' = .5$ (chance line) to $A' = .9$. Contours for d' are based on d' value computed from hits (H) and false alarms (FA) for A' at $B'' = 0$ (the negative diagonal). When sensitivity is low, there is a close correspondence in isosensitivity contours for A' and d'. With increasing sensitivity, A' tends to underestimate performance relative to d' for large biases; specifically, for a given false alarm rate, A' requires a higher hit rate to achieve constancy.

Snodgrass and Corwin (1988) raised an additional concern about independence for A' and B''. They note that these measures "show marked dependence of bias at low levels of discrimination" (p. 39). This problem is not unique to these "nonparametric" statistics, but rather is shared with d' and log(β) or ln(β), where the logarithmic transform of β changes it from a ratio to interval measure. However, Snodgrass and Corwin report that d' and c (the alternative criterion measure discussed above) do meet their criterion for independent measures of sensitivity and bias.

Final considerations about A' and d'. It is useful, at this point, to pull together information from our discussion of the G-SDT descriptive statistics and about the "nonparametric" SDT statistics. The G-SDT statistics for sensitivity (d') and bias [e.g., β, log(β), c] have the disadvantage of requiring specific assumptions about the underlying distributions, but if those assumptions are reasonably valid, then d' and β or c are orthogonal or independent. Since the explicit assumptions are no different than the implicit assumptions of the most common inferential statistics used in research, the typical researcher is already assuming the reasonable validity of the assumptions. The A' "nonparametric" SDT statistic has the advantage of being easily interpreted as percent correct, but has the disadvantages of not being distribution free, not being independent of bias, and underestimating ability by an amount that is a function of magnitude of bias and decision ability. The one advantage for A' can be achieved by either computing $p(c)_{max}$ (the unbiased proportion correct that can be computed from d'; Macmillan & Creelman, 1991) or A_Z, the best-fitting Gaussian ROC curve (Swets, 1986, 1996; Swets, Dawes, & Monahan, 2000).

More About ROC Curves

In the previous section, we used the ROC curve to evaluate, compare, and contrast alternative descriptive statistics for sensitivity. Earlier, we discussed problems in interpreting whether an observer's criterion, as opposed to the researcher's criterion statistic, has changed. These, and other problems, such as a lack of reasonable validity to the assumptions of G-SDT, are minimized when ROC curves are generated. The behavior in ROC space of different descriptive criterion statistics either has been documented (e.g., Egan, 1975; Macmillan & Creelman, 1991) or is obvious (a constant Type I or Type II error rate). The validity of the G-SDT assumptions can be easily evaluated and, if they are not reasonably valid, other descriptive statistics are available (see, e.g., Green & Swets, 1966; Swets, 1996), including the truly nonparametric measure of area under the ROC curve. The major disadvantage to ROC analyses is the cost, in both time and effort, of generating sufficient data. If the research suggests significant concerns about the validity of the G-SDT assumptions, then the researcher needs to be concerned about any descriptive and inferential statistics, and the extra costs of an ROC analysis are probably justified. However, if the researcher is confident enough to utilize standard inferential statistics, then concerns about the assumptions of G-SDT also should be minimal.

Ideal Observer

An important aspect of SDT is the *ideal observer*, a model of the best possible decision performance based on the information available (e.g., Tanner, 1961). Models of SDT (e.g., G-SDT) provide a common basis for evaluating the performance of both ideal and actual observers. There is greater uncertainty in the performance of human observers, resulting in poorer performance than that of an equivalent ideal observer, but the pattern of abilities should be similar when the actual and model observers utilize the same information. Thus, ideal observer models provide an important tool for evaluating the capabilities of the actual observer. For example, in a detection task, the monaural human listener can be modeled as phase insensitive, and as sampling the signal-to-noise ratio output of a flexible set of narrow-band filters. Likewise, the researcher can develop ideal observer models that can be used to evaluate the performance of groups of observers (e.g., a committee or jury) that must reach joint decisions (e.g., Sorkin & Dai, 1994; Sorkin, Hays, & West, 2001). G-SDT also provides a conceptual framework for analyzing and evaluating contrasting theoretical explanations for complex concepts. The analyses of criterion change explanations of memory (Stretch & Wixted, 1998; Wickens & Hirshman, 2000; Wixted & Stretch, 2000), illustrated in Figure 1, are examples of the use of this important SDT tool. Similarly, SDT has been used to model possible strategies used by observers in what might seem to be simple experimental paradigms. For example, a *same–different* task is far more complex than a *yes–no* task; the analysis by Macmillan and Creelman (1991) demonstrates that the task could be performed using two criteria and a bivariate decision space, with Irwin et al. (2001) demonstrating alternative decision strategies.

Some Concluding Remarks

SDT is a theory of decision making about the evidence being evaluated. This theory offers the modern researcher a wide range of excellent, powerful statistical tools, including easily conceptualized descriptive statistics and a formal structure for describing and analyzing alternative information processing strategies. SDT has much in common with modern inferential and descriptive statistics. Like these more common statistical tools, the use of SDT, and the reasonable interpretation of empirical results expressed as G-SDT descriptive statistics, requires a reasonable (but not necessarily expert) understanding of the underlying theoretical basis. A wide range of problems in the modern use of SDT, and incorrect characterizations of SDT, reflect problems in that understanding. The major goal of the present article is to provide the basis for the needed understanding, describing SDT in a manner that is deeper and more accurate than that found in modern textbooks, that is more understandable than the original literature and early texts devoted to SDT, and that has effectively highlighted reasonable pragmatic concerns.

REFERENCES

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception & Performance*, **25**, 1-18.

Berliner, J. E., & Durlach, N. I. (1973). Intensity perception. Resolution in roving-level discrimination. *Journal of the Acoustical Society of America*, **53**, 1270-1287.

Braida, L. D., & Durlach, N. I. (1972). Intensity perception: II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America*, **51**, 483-502.

Coren, S., Ward, L. M., & Enns, J. T. (1999). *Sensation and perception*. New York: Harcourt Brace.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception: I. Pre-

liminary theory of intensity resolution. *Journal of the Acoustical Society of America*, **46**, 372-383.
Egan, J. P. (1975). *Signal detection theory and ROC analysis.* New York: Academic Press.
Goldinger, S. D. (1998). Signal detection comparisons of phonemic and phonetic priming: The flexible-bias problem. *Perception & Psychophysics*, **60**, 952-965.
Goldstein, E. B. (1999). *Sensation and perception* (5th ed). Pacific Grove, CA : Brooks/Cole.
Goldstein, E. B. (2002). *Sensation and perception* (6th ed). Pacific Grove, CA : Wadsworth.
Green, D. M. (1964). General prediction relating yes–no and forced-choice results [Abstract]. *Journal of the Acoustical Society of America*, **36**, 1042.
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 indices for sensitivity and bias: Computing formulas. *Psychological Bulletin*, **98**, 185-199.
Haberlandt, K. (1994). *Cognitive psychology.* Boston: Allyn & Bacon.
Hays, W. L. (1963). *Statistics for psychologists.* New York: Holt, Rinehart & Winston.
Hicks, J. L., & Marsh, R. L. (2000). Toward specifying the attentional demands of recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1483-1498.
Irwin, R. J., & Hautus, M. J. (1997). Likelihood-ratio decision strategy for independent observations in the *same–different* task: An approximation to the detection-theoretic model. *Perception & Psychophysics*, **59**, 313-316.
Irwin, R. J., Hautus, M. J., & Francis, M. A. (2001). Indices of response bias in the *same–different* experiment. *Perception & Psychophysics*, **63**, 1091-1100.
Jahnke, J. C., & Nowaczyk, R. H. (1998). *Cognition.* Upper Saddle River, NJ: Prentice-Hall.
Kadlec, H., & Townsend, J. T. (1992). Signal detection analyses of dimensional interactions. In F. G. Ashby (Ed.). *Multidimensional models of perception and cognition* (pp.181-227). Hillsdale, NJ: Erlbaum.
Kantowitz, B. H., Roediger, H. L., III, & Elmes, D. G. (2001). *Experimental psychology* (7th ed.). Pacific Grove, CA: Wadsworth.
Laming, D. (1986). *Sensory analysis.* London: Academic Press.
Levine, M. W., & Shefner, J. M. (1991). *Fundamentals of sensation and perception* (2nd ed.). Pacific Grove, CA: Brooks/Cole.
Licklider, J. C. R. (1964). Theory of signal detection. In J. Swets (Ed.), *Signal detection and recognition by human observers* (pp. 92-121). New York: Wiley.
Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103-189). New York: Wiley.
Macmillan, N. A., Braida, L. D., & Goldberg, R. F. (1987). Central and peripheral processes in the perception of speech and nonspeech sounds. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 28-45). Dordrecht: Martinus Nijhoff.
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.* New York: Cambridge University Press.
Macmillan, N. A., & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review*, **3**, 164-170.
Macmillan, N. A., Goldberg, R. F., & Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America*, **84**, 1262-1280.
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.
Masson, M. E. J., & Borowsky, R. (1998). More than meets the eye: Context effects in word identification. *Memory & Cognition*, **26**, 1245-1269.
McNicol, D. (1972). *A primer of signal detection theory.* London: George Allen & Unwin.
Miller, M. B., & Wolford, G. I. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review*, **106**, 398-405.
Mood, A. M., & Graybill, F. A. (1963). *Introduction to the theory of statistics.* New York: McGraw-Hill.
Noreen, D. L. (1981). Optimal decision rules for some common psychophysical paradigms. In S. Grossberg (Ed.), *Proceedings of the Symposium in Applied Mathematics of the American Mathematical Society and the Society for Industrial and Applied Mathematics* (Vol. 12, pp. 237-279). Providence, RI: American Mathematical Society.
Norris, D. (1986). Word recognition: Context effects without priming. *Cognition*, **22**, 93-136.
Norris, D. (1995). Signal detection theory and modularity: On being sensitive to the power of bias models of semantic priming. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 935-939.
Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral & Brain Sciences*, **23**, 299-370.
Pastore, R. E., & Scheirer, C. J. (1974). Signal detection theory: Considerations for general applications. *Psychological Bulletin*, **81**, 945-958.
Payne, D. G., & Wenger, M. J. (1998). *Cognitive psychology.* New York: Houghton Mifflin.
Peterson, W. W., & Birdsall, T. G. (1953). *The theory of signal dectectability* (Tech. Rep. No. 13). Ann Arbor: University of Michigan, Electronic Defense Group.
Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of d'_e. *Psychological Bulletin*, **71**, 161-173.
Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, **1**, 125-126.
Rhodes, G., Parkin, A.J., & Tremewan, T. (1993). Semantic priming and sensitivity in lexical decision. *Journal of Experimental Psychology: Human Perception & Performance*, **15**, 154-165.
Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 803-814.
Roediger, H. L., III, & McDermott, K. B. (1999). False alarms about false memories. *Psychological Review*, **106**, 406-410.
Schiffman, H. R. (2000). *Sensation and perception: An integrated approach.* New York: Wiley.
Sekuler, R., & Blake, R. (1990). *Perception.* New York: McGraw-Hill.
Sekuler, R., & Blake, R. (2002). *Perception* (4th ed.). New York: McGraw-Hill.
Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.
Sorkin, R. D., & Dai, H. (1994). Signal detection analysis of the ideal group. *Organizational Behavior & Human Decision Processes*, **60**, 1-13.
Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal detection analysis of group decision making. *Psychological Review*, **108**, 183-203.
Stevens, S. S. (1961). In W. A. Rosenblith (Ed.), *Sensory communication: Contributions to the symposium on principles of sensory communication* (pp. 1-34). Cambridge, MA: MIT Press.
Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1397-1410.
Swets, J. A. (Ed.) (1964). *Signal detection and recognition by human observers: Contemporary readings.* New York: Wiley.
Swets, J. A. (1986). Indicies of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin*, **99**, 100-117.
Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers.* Mahwah, NJ: Erlbaum.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, **1**, 1-26.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. (1961). Decision processes in perception. *Psychological Review*, **68**, 301-340.

Tanner, W. P., Jr. (1961). Physiological implications of psychophysical data. In Annals of the New York Academy of Sciences (Vol. 89, pp. 752-765). New York: New York Academy of Sciences.

Treisman, M. (2002). Is signal detection theory fundamentally flawed? A response to Balakrishnan (1998a, 1998b, 1999). *Psychonomic Bulletin & Review*, **9**, 845-857.

Wald, A. (1950). *Statistical decision functions.* New York: Wiley.

Wald, A. (1955). Basic ideas of a general theory of statistical decision rules. In T. W. Anderson et al. (Eds.), *Selected papers in statistics and probability by Abraham Wald* (pp. 656-668). New York: McGraw-Hill. (Original work published 1952)

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Wickens, T. D., & Hirshman, E. (2000). False memories and statistical decision theory: Comments on Miller and Wolford (1999) and Roediger and McDermott (1999). *Psychological Review*, **107**, 377-383.

Wixted, J.T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, **107**, 368-376.

NOTES

1. In Wald's derivation, "the notions of Bayes solutions and a priori distributions are used here merely as mathematical tools to express some results concerning complete classes of decision rules, and in no way is the actual existence of an a priori distribution . . . postulated" (Wald, 1952/1955, p. 663). Alternatively, "every Bayes Strategy is a likelihood-ratio test" (Mood & Graybill, 1963, p. 284).

2. A likelihood ratio is, by definition, a ratio scale. More important for the analysis of data, β also is a ratio. Most descriptive and inferential statistics assume an interval, not a ratio, scale, and there are very different permissible transformations for the two types of scale (e.g., arithmetic vs. geometric mean; see Stevens, 1961). In contrast to β, log β and c are both interval scales, and thus compatible with standard descriptive and most inferential statistics.

If the likelihood ratio reflects the product of the weighted independent probabilities of the specific observed value along a large number of independent dimensions, and if the decision variable approximates a logarithmic transformation of the likelihood ratio, then the distribution of observations represents the sum of a large number of independent random variables and thus should be Gaussian in nature. This was one theoretical basis for positing Gaussian distributions that became the basis of the Gaussian models of SDT (e.g., Green & Swets, 1966).

3. If the decision variable is related to a likelihood ratio, the important statistical properties are relative, rather than absolute. For example, in the detection tasks used to validate SDT there are two probability density functions, one for N and one for S+N. The likelihood ratio then reflects the relative probability that the specific observation occurred due to S+N relative to N being presented.

4. The general Gaussian model of SDT assumes that the underlying distributions are Gaussian, but does make the equal variance assumption (e.g., Green & Swets, 1966), whereas even less common, alternative models assume different underlying distributions (see Egan, 1975; Swets, 1964). A different decision model, Luce's choice model (Luce, 1963) assumes underlying logistic distributions. Finally, threshold decision models make no assumptions about underlying distributions other than with regard to the number of discrete decision categories or states (e.g., Green & Swets, 1966; Macmillan & Creelman, 1991).

5. Balakrishnan (1999) has argued that changes in criterion do alter the variability of the decision variable. Evidence for his argument is from rating tasks that are designed to create data sets with minimal, but nonzero, entries. Theorists have argued that Balakrishnan's findings are the expected consequence of the analyses' minimal data sets and the statistics that Balakrishnan developed (Treisman, 2002).

6. One problem with c is that (except at $c = 0$) its value often needs to be interpreted relative to the distribution means as a function of d', and this can pose problems in interpreting bias magnitude (this is also true for β). For this reason, Macmillan and Creelman (1991) suggest an alternative descriptive statistic for criterion, c', which equals c/d'; the c' statistic is clearly not independent of d'.

7. All probability density functions in Figure 1 have been adjusted to reflect the same total area; the difference in mean between A1 and A2 is sufficiently small for the variance of the combined distribution, A, to be similar to the component distributions.

8. Snodgrass and Corwin (1988) attribute the basis of A' to an equivalent statement by Green and Moses (1966).

9. An ROC curve can be plotted in simple probability space (with linear probability coordinates) or in a space that reflects probability scaled to assumed distributions. With normal or probit coordinate scales, the probability axes reflect linear z-score distances. The geometric solution for A' is based on the ROC curve plotted in simple probability space. There are two implicit points for all ROC curves, (0,0) and (1,1), the endpoints on the positive diagonal that represents chance performance. For a point (F,H) in ROC space, the linear ROC curve that is a line from (0,0) through this point represents minimum hit rate (and maximum false alarm rate), whereas a line that ends at (1,1) represents maximum hit rate (and minimum false alarm rate).

10. If the underlying distributions of possible event classes are not equal in variance, d' will not be independent of the criterion. SDT offers the researcher alternative descriptive statistics (e.g., d'_a, d'_s, d'_e) derived from ROC curves (e.g., Egan, 1975; Green & Swets, 1966; Macmillan & Creelman, 1991), but use of these alternative statistics requires either knowledge of the underlying variance ratio (e.g., see Green & Swets) or the generation of an empirical multipoint ROC curve.

11. Our point is that, except for chance performance ($A' = .5$), all isosensitivity contours for A' never contain the extreme values of ROC space, (0,0) and (1,1). When $A' > .50$ (greater than chance), hit rate will always be greater than 0 and false alarm rate will be less than 1.0. When of $A' < .50$ (less than chance), the isosensitivity would be equal to the contour for $1-A'$, but rotated at (or flipped along) the positive diagonal (the chance line .50).

(Manuscript received October 23, 2001;
revision accepted for publication July 12, 2002.)