View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by NASA Technical Reports Server

N94-11531

Signal Detection Theory and Methods for Evaluating Human Performance in Decision Tasks.

Kevin O'Brien
(primary contact)
Lockheed Engineering & Sciences Co.
2400 NASA Road 1
C95
Houston, Tx. 77058-3711
(713) 333-7130

Evan M. Feldman
Rice University
Department of Psychology
P. O. Box 1891
Houston, Tx. 77251
(713) 527-8101 ext. 2216
emfeld@ricevm1.rice.edu

Signal Detection Theory (SDT) can be used to assess decision making performance in tasks that are not commonly thought of as perceptual. SDT takes into account both the sensitivity and biases in responding when explaining the detection of external events. In the standard SDT tasks, stimuli are selected in order to reveal the sensory capabilities of the observer. SDT can also be used to describe performance when decisions must be made as to the classification of easily and reliably sensed stimuli. Numbers are stimuli that are minimally affected by sensory processing and can belong to meaningful categories that overlap. Multiple studies have shown that the task of categorizing numbers from overlapping normal distributions produces performance predictable by SDT. These findings are particularly interesting in view of the similarity between the task of categorizing numbers and that of determining the status of a mechanical system based on numerical values that represent sensor readings. Examples of the use of SDT to evaluate performance in decision tasks are reviewed. The methods and assumptions of SDT are shown to be effective in the measurement, evaluation, and prediction of human performance in such tasks.

INTRODUCTION

The purpose of this paper is to discuss the relevance of Signal Detection Theory (SDT) to the evaluation of human decision making. SDT is typically thought of in terms of observers detecting faint, experimental stimuli in the hopes of revealing something about the sensory system of the observer. Such an experiment takes advantage of only a part of the information SDT can provide and assumes that SDT is only applicable to describing sensory functions. This paper will review the scope of SDT and report specific examples of application of SDT to more cognitive tasks.

The contribution of SDT is its attempt to explain detection performance by taking into account both the sensitivity and the response bias of the observer. An observer, say a fighter pilot, is aware of an object in the distance. As the distance between the pilot's craft and the perceived object decreases, the pilot's ability detect an object would be expected to increase. In addition, the pilot would be expected to more accurately identify the object as a hostile aircraft, simply a dark spot in clouds, or as some other uninteresting object. The pilot's decision that an object is a hostile aircraft is also a function of willingness to report a target and in doing so to risk sounding a false alarm. SDT offers explanations for the difficulty encountered in detecting or discriminating objects, and how a criterion for responding is established. SDT has contributed greatly to the revitalization of interest in the study of psychophysics, but the theory and methods are not limited to the study of sensory stimuli.

This research was performed under contract from the National Aeronautics and Space Administration, Johnson Space Center. The research was conducted in the Human-Computer Interaction Laboratory in the Flight Crew Systems Division at Johnson Space Center. Marianne Rudisill was the manager of the laboratory.

Psychophysics is the study of the relationship between physical events and the resulting mental events. Two basic questions are asked: does a physical event result in a mental event, and do incremental changes in physical events result in equal increments in mental events. For example, can the observer detect the onset of a single pixel on a dark computer screen? Can the observer tell the difference between the onset of one pixel and two pixels? These questions address the sensitivity of the visual system, the readiness of the observer to report an event, and the scale of perceptual change.

In psychophysics we often assume that the observer could appropriately assign a response to the event if only the event could be clearly sensed. On the other hand, some events that are clearly sensed are difficult to assign to a response. Classification of a bat as a mammal or a bird would be difficult on the basis of a limited set of information because so many of the obvious characteristics of the bat seem to match those we attribute to bird. While SDT has always been used to understand sensory tasks, SDT methods are becoming more widely used in addressing classification tasks. Swensson (1980) used SDT to describe the performance of radiologists in interpreting chest x-rays. Swets (1988) argued for the use of SDT methods in measuring the accuracy of diagnostic systems providing examples from the medical field, weather forecasting, and materials testing. Parasuraman and Wisdom (1985) suggest the use of SDT to evaluate the rules of expert systems and as a guide for designing systems in which automated expert systems assist human operators. Sorkin and colleagues (Sorkin & Woods, 1985, Sorkin & Robinson, 1984, Sorkin, Kantowitz, & Kantowitz 1986) have dealt with the issue of automated expert based alarms in system control environments. In each of these cases, SDT is applied to problems of categorizing easily detected information as being either meaningful or inconsequential. SDT can be used to describe the process by which one category is distinguished from another and how response biases affect responding. The body of this paper details the applicability of SDT to these problems and describes the use of SDT methods to examine the processing of multiple sources of information.

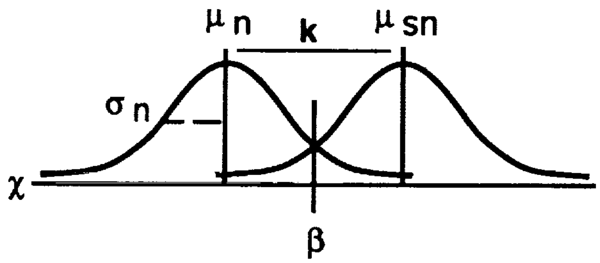


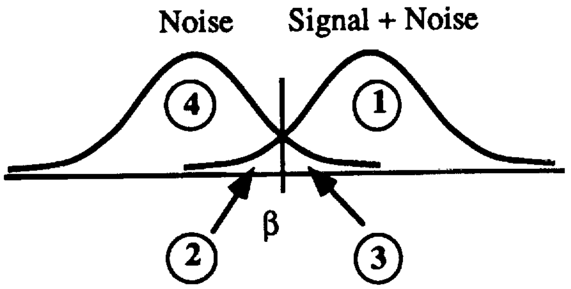
Figure 1. Hypothetical probability distributions for noise (n) and signal-noise (sn).

SDT: THEORY, METHODS, & METRICS

Theory

SDT attempts to account for differences in sensitivity and response bias starting from the assumption that uncertainty surrounds the processing of an event. Uncertainty is present because a variable level of background noise surrounds every interesting event, or in SDT terms, signal. Borrowing an example from Baird and Noma (1978), consider the circumstance that you are listening to the stereo and the phone rings. The sound from the stereo is background noise and the phone ring is the signal of interest. The more distinct the signal is from the noise, the more likely the signal, in this case the ringing phone, will be detected. Uncertainty arises from the fact that on some occasions what you have heard could as easily be attributed to the stereo alone as to the ringing phone with the stereo in the background.

Figure 1 helps us to think of the uncertainty of detecting a signal in a more detailed manner. The continuum χ is the evidence gathered by the observer from some event. The noise present at any given time is expected to be a random observation from a distribution of noise events having a mean μ_n and a variance σ_n. The presence of a signal along with the noise adds a constant, k, to the values in the noise distribution resulting in the signal-noise distribution with a mean μ_{sn} and variance equal to that of the noise distribution. As can be seen in the figure, intermediate levels of evidence are included under the distributions for both the noise and signal-noise distributions; and therefore, the



	Yes	No
Signal	1 Hit	2 Miss
Noise	3 False Alarm	4 Correct Rejection

Figure 2. The response criterion, β, separates the noise and signal-noise distributions into four response categories: hit, miss, false alarm, and correct rejection.

assignment of evidence to one distribution or the other is uncertain. The more variable the noise or the smaller the change introduced by adding the signal, the more uncertain the assignment of intermediate levels of evidence to one distribution or the other.

Two additional factors affect the observer's response selection: the establishment of a response criterion and the likelihood of a signal occurring. When an observer establishes a fixed response criterion, β, responding "yes" to evidence above β and "no" to evidence equal or below, responses are relegated to one of four categories (see Figure 2). Responding "yes" will result in either a correct detection (**hit**) given that a signal was present or a **false alarm** given only noise. Responding "no" will result in either a **correct rejection** given noise or a missed (**miss**) signal given the presence of a signal. An optimal placement of β would minimize the likelihood of false alarms and misses while maximizing the number of hits and correct rejections. Alternative criteria are possible. In the example of the ringing phone above, if one were expecting a very important call (but not important enough to turn off the stereo!) one would be more willing to risk picking up the phone when it hadn't rung (false alarm) than miss a real call. This liberal strategy would move β to the left on the χ axis. A conservative strategy would move β to the right and result in the commission of few false alarms at the expense of missing some calls. Varying the costs and benefits of different responses alters the placement of β by an observer. The likelihood of a signal also alters the response selection of the observer. Up to this point we have been assuming that the chance of a signal was equal to that of noise alone. On the other hand, as the likelihood of a signal declines from 50% to 10%, we would expect to see a similar reduction in the number of yes or signal present responses.

In summary, SDT is based on the assumption that there is uncertainty regarding the classification of an event. That uncertainty is related to the variability of noise and the resulting overlap between the noise and signal-noise distributions. Ability to detect the signal in noise increases as the overlap of the distributions decrease. Responding is also affected by response bias in terms of willingness to respond yes and/or the expectation regarding signal frequency.

Methods

The methods proposed by SDT involve manipulation of the signal, the responses, and the expectation/reward for a particular type of responding. As implied in the above examples, the task generally involves a observer being directed to make an observation and report whether the interval of observation contained a signal or only noise. In this case, the presence of a signal is contrasted with the absence of a signal. SDT can also be used to describe the processing of multiple signals (see Macmillan & Creelman, chap. 10, 1991). The use of multiple signals allows investigation of the observer's ability to identify the signals (signal A versus signal B) and the ability to detect the combination of multiple signals against noise.

The responses required of the observer can also be varied. The two most common variations being the yes/no response used in the preceding examples and the multiple interval rating. The yes/no response produces a single estimate of the response criterion used by the observer. The rating method requires the observer to provide a measure of the certainty of responding. For instance, the observer could be told to respond using the numbers one through six with one representing absolute certainty of a noise event and six representing absolute certainty of a signal event. The ratings can then be summed into five levels of criterion with rating 1 versus the other five being the most liberal criterion. The advantage of collecting rating responses lies in being able to determine sensitivity at varying levels of response bias. If the sensitivity varied across levels of response bias, it would indicate that σ_n is unequal to σ_{sn}. An assumption of equal variance might result in inappropriate comparisons among levels of sensitivity, a condition that can be avoided when rating responses are collected (see Macmillan & Creelman, 1991, pg.82-85). Responding can also be manipulated by altering the likelihood of a signal, or the reward/punishment for exceeding a level of one of the four response categories (hit, miss, false alarm, correct rejection).

Metrics

The primary metrics developed in conjunction with SDT quantify two kinds of information: the sensitivity of the observer and the observer's response bias. Sensitivity is measured as the distance between the means of the noise and signal-noise distributions taking into account the variance of the distributions and is measured as d′.

$$d' = \frac{\mu_{sn} - \mu_n}{\sigma_n}$$

Response bias, β, is measured by the ratio of noise to signal probabilities multiplied by the difference between correct rejections (cr) and false alarms (fa) divided by the difference between hits (h) and misses (m).

$$\beta = \frac{p(n)}{p(sn)} \times \frac{cr-fa}{h-m}$$

Returning to Figure 1, it should be clear that given equal likelihood of noise and signal-noise, the optimal β would divide the distributions into equal proportions of hits and correct rejections, thereby resulting in a β of 1. Changing the rewards for a particular response type, say punishing false alarms, necessarily shifts β one way or the other. The resulting change in the distribution of responses among the four possible outcomes provides a measure of β.

UNCERTAINTY

Uncertainty with regard to signal and noise lies at the heart of SDT. The uncertainty is attributed to variability in the production and processing of the noise and signals. In many psychophysical experiments, the signals are taken to be relatively stable. Variability is introduced by the processing channel through which the signal is encoded. For instance, a visual stimulus is expected to be relatively constant. On the other hand, the perception of the stimulus is made variable by random neural firings, the effects of spatial summation, and the retinal location on which the image is projected to name a few. Each of these effects serves to increase the overlap between the noise and signal-noise distributions.

The critical point for this presentation is that variability can be introduced in other ways as well. Consider the task of sorting a box of school photos into two classes: former classmates versus persons unknown to you. The photo is a fixed image and your perception is not degraded by the only getting a brief look at the photo or the angle at which the photo is displayed. You could describe the photo with a clarity that would allow some other person to select it from the box. The difficulty that you encounter in classifying the photo is not related to your processing the image. Instead, the difficulty is related to your ability to extract from memory the characteristics that would allow you to distinguish former classmates from people you have never seen before. The similarity in facial features and the difficulty with assigning features to names results in a noise distribution.

Noise and signal-noise distributions can be produced using numbers. Numbers are reliably and accurately identified by most adults, yet meaningful categories of numbers can have a great deal of overlap. For instance, the heights of men and women have different means, yet if you were given an intermediate height, say 5'6", there would be uncertainty with regard to knowing whether the height was that of a woman or a man. In an effort to determine whether β is fixed or changes over time, Kubovy, Rapoport, & Tversky (1971) conducted an experiment using the height classification task. The observed d′ was consistent with the d′ expected given the means and variance of the numerical distributions used as stimuli. Measures of the criterion supported a deterministic, fixed β strategy for criterion setting as opposed to a probabilistic, variable β strategy. Numbers have also been used as stimuli in studies examining the convergence of various psychophysical methods on perceptual scaling (Weissmann, Hollingsworth, & Baird, 1975), and the independence of sequential presentations of stimuli with respect to responding (Ward, Livingston, & Li, 1988).

In system control environments, the task of deciding whether the numerical temperatures from a cooling mechanism are more representative of a normal operating condition than a malfunction is very similar to deciding whether a given height is more likely to represent that of a man or a woman. The certainty with which the operating status of a system can be determined from an observation on the system is in part a function of the distance between the means of the normal and malfunction distributions and the extent to which the distributions overlap. Therefore, the performance of an operator deciding that the system is okay or failing can be evaluated in terms of SDT.

APPLICATION

To illustrate the application of SDT to a human decision making problem, we will describe the method and analysis used in a study we conducted. In this research, we are interested in how people use information from a variety of sources, particularly when one source, the expert advisor, is expected to be a more accurate source of information. Previous studies have looked at the effects of varying the criterion information provided by an expert advisor. The basic method used by Sorkin, Kantowitz, & Kantowitz (1988) was to compare decisions made by observers using system data (digital gauges) with observers using the gauges in conjunction with expert advice. The expert advice is provided in either a 1 bit (nominal, malfunction) or a 2 bit (certain nominal, possible nominal, possible malfunction, certain malfunction) message indicating the criterion used for a given event. The study showed that the addition of expert advice improved decision accuracy and gave some indications of extra advantage for the 2 bit message over the 1 bit.

Sensitivity, or d′, of the system was established by the mean and standard deviation of the normal distributions from which the values for a nominal and malfunction event were taken. The sensitivity of the expert advisor was set at a level higher than that of the four gauges. This difference made it difficult to determine whether the improvement in performance was the result of combining the information from the gauges with the expert advice or simply the result of relying on the expert advice. We set out to determine what information the observers used by conducting a study in which the sensitivity of the expert advice varied from worse than that of the gauges to better than that of the gauges.

The generation of stimuli and the analysis of these studies are both dependent on SDT. The stimuli were generated in much the same way as described in the height example above. The two categories, system normal and system malfunction, were defined as having numerical means of 3 and 4, respectively. For the gauges, the standard deviation for each distribution was set at 1.54, yielding an expected d′ for each gauge of .649 and a combined d′ of 1.298 for the four gauges, as will be explained below. The standard deviation of the distribution on which the automated expert based its advice was varied, resulting in d′ levels of .191, .929, 1.667, and 2.774. In order to elicit measures of sensitivity across a variety of response criteria, responses were collected using a rating scale method with six response categories.

Manipulating the d′ of information from the expert advisor allowed us to examine the differential effect of the advice in the face of a constant level of information from the gauges. From a theoretical perspective, the

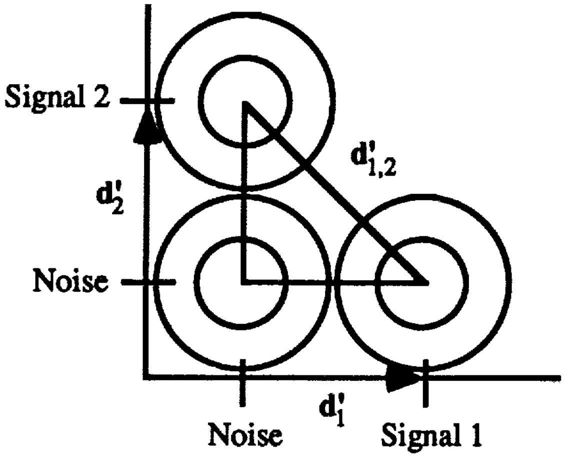


Figure 3. The discriminability between two stimuli, signal 1 and signal 2, is predicted by the distance between their means in d′ units.

use of information could be hypothesized as either the result of combining all the information or the result of selective filtering of the available information. SDT provides predictions regarding the combination of information (Tanner, 1956, Macmillan & Creelman, chap. 10, 1991, or for a comparison between the predictions of SDT and other theories see Massaro & Friedman, 1990). A common prediction discussed in the SDT literature takes advantage of d′ as a measure of distance. Take the case of two signals: sn_1 and sn_2 (see Figure 3). Independence is represented in the 90° angle of the intersection of the vectors. The two sources of information are assumed to share a common noise distribution ($n_{1,2}$). Using geometry, if sensitivity to sn_1 is described by d'_1 and sensitivity to sn_2 is described by d'_2, then the Pythagorean theorem allows us to predict the discriminability of sn_1 from sn_2 as

$$d'^2_{1,2}=d'^2_1+d'^2_2-2d'_1d'_2\cos(\theta).$$

When the signals are independent ($\theta=90^o$, $\cos(90^o)=0$) and d'_1 is equal to d'_2 then

$$d'_{1,2}=\sqrt{2}d'_1.$$

This line of thinking can be extended to predicting the detectability of the combined evidence, sn_1 and sn_2, against the noise distribution by changing our focus from calculating the distance between the means of sn_1 and sn_2 to calculating the distance between $sn_{1,2}$ and $n_{1,2}$. Changing the orientation of the legs of the triangle, it is obvious that the calculation remains the same. This prediction, referred to as the Euclidean metric, can be extended to m independent information sources having equal d′s by the formula $\sqrt{m}d'$. The Euclidean metric predicts that performance will exceed that expected from any of the component parts (see Figure 4, panel 1).

The simplest filtering prediction suggests that one source of information will be processed to the exclusion of other sources. Two sources are possible: the expert advice and the gauges. If only the expert advice is used, then one would expect a linear increase in performance corresponding to the increase in d′ of the expert advice (see Figure 4, panel 2). On the other hand, if the gauges were used, one would expect no change in performance as the d′ of the advice improved (see Figure 4, panel 3).

An alternative model based on filtering would suggest that through repeated exposure, the observer learns the relative sensitivity of available sources, and in some manner weights the contribution of the sources in accordance with the d′. The

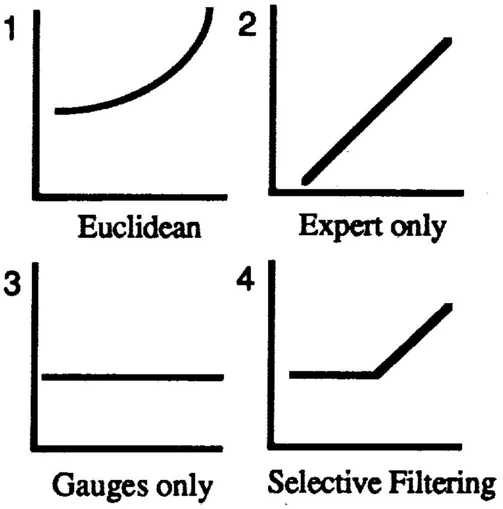


Figure 4. The effect of increasing expert system sensitivity on overall performance for four hypothesized outcomes based on combination of information (Euclidean) and filtering (expert only, gauges only, and simple selection).

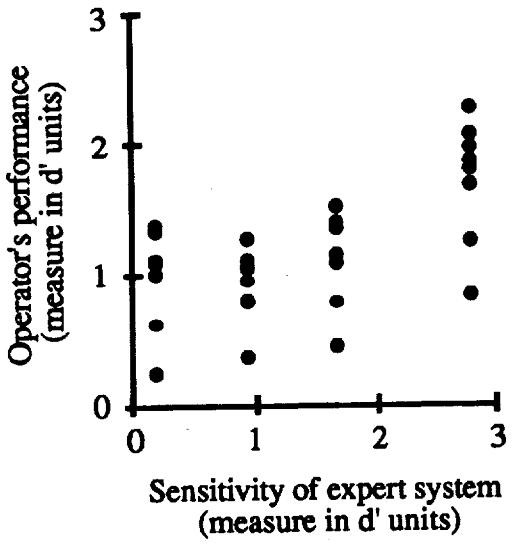


Figure 5. Observed performance showing the effect of expert system sensitivity on operator performance in the system control task.

simplest model based on this approach would predict that the observer would use only the source having the highest d′. The result would be a decision based on the information from the gauges when the advice is less sensitive than the gauges and on the advice when the advice was more sensitive than the gauges (see Figure 4, panel 4).

The experiment was conducted and the best fit for the data was a curve that increased at an increasing rate (see Figure 5). This result rules against accepting the filtering models that are based on only one source of information. Differentiating between the Euclidean metric model and filtering based on d′ is more difficult. Both would be predicted to be curvilinear and increasing at an increasing rate given the d′s provided in the task. The primary distinction between the two models relies on location of the curve. The Euclidean model predicts performance that exceeds that of either source. The filtering model predicts performance equal to the more sensitive of the sources. In practice, observed d′s frequently fall below predicted d′s. As such, selecting between models based on location of the curve has problems in addition to the variability of the data. Additional experiments manipulating the observer's knowledge of the sensitivity of each information source are being conducted to distinguish between the Euclidean and filtering models.

CONCLUSION

In conclusion, SDT provides an assessment of both the decision maker's sensitivity and response bias. Sensitivity can be a function of the variability of noise and signal processing inherent in sensory processes, or, as with numbers, a function of the uncertainty with which individual numbers are categorized. In numerous studies in which numbers have served as stimuli the theory and methods of SDT have been shown to be a valuable tool for explaining the decision making performance of observers. This is particularly valuable in view of the similarity between assigning numbers in a laboratory task and the task of using numbers to categorize the status of a mechanical system. Studies currently being conducted demonstrate the value of SDT in describing and predicting the influence of automated expert system advice on decision making. In one instance, SDT has been used to demonstrate that decision makers process both numerical system data and expert system advice in a task requiring assessment of system status.

REFERENCES

Baird, J. C., & Noma, E. (1978). Fundamentals of Scaling and Psychophysics. New York: J. Wiley & Sons.

Kubovy, M., Rapoport, A., & Tversky, A. (1971). Deterministic vs probabilistic strategies in detection. *Perception & Psychophysics, 9*, 427-429.

Macmillan, N. A., & Creelman, C. D. (1991). Detection Theory: A User's Guide. Cambridge: Cambridge University Press.

Massaro, D. & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review, 97*, 225-252.

Parasuraman, R. & Wisdom, G. (1985). The use of signal detection theory in research on human-computer interaction. *Proceedings of the Human Factors Society - 29th Annual Meeting*, 33-37.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors, 30*, 445-459.

Sorkin, R. D., & Robinson, D. E. (1984). Alerted monitors: Human operators aided by automated detectors. (National Technical Information Service Report No. DOT/OST/p-34/85-021). Washington, DC: U.S. Department of Transportation.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human-Computer Interaction, 1*, 49-75.

Swensson, R. (1980). A two-staged detection model applied to skilled visual search by radiologists. *Perception & Psychophysics, 27*, 11-16.

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

Tanner, W. P. Jr. (1956). Theory of recognition. *Journal of the American Acoustical Society of America, 28*, 882-888.

Ward, L. M., Livingston, J., & Li, J. (1988). On probabilistic categorization: The Markovian observer. *Perception & Psychophysics, 43*, 125-136.

Weissmann, S. M., Hollingsworth, S., & Baird, J. (1975). Psychophysical study of numbers: III. Methodological applications. *Psychological Research, 38*, 97-115.