The Use of Signal Detection Theory to Evaluate Aphasia Diagnostic Accuracy and Clinician Bias

Donald A. Robin and Malcolm R. McNeil

Diagnosing aphasia involves differentiating it from other speech-language pathologies that resemble it. The validity and reliability of the differential diagnoses among neurogenic and psychiatric pathologies have been approached from a variety of perspectives. For example, Halpern, Darley, and Brown (1973) used the pattern of performance across a variety of speech, language, and other cognitive tasks to construct a differential profile for confused, demented, and aphasic patient performance. Wertz and Rosenbek (1971) also profiled differentiation of apraxia of speech from aphasia.

Recently, discriminant function analyses have been used to construct a statistical profile for the differentiation of aphasia from conditions resembling it. The Porch Index of Communicative Ability (PICA) (Porch, 1971) has been used to differentiate aphasia from malingering (Porch, Frieden, & Porec, 1977), normal (Brauer, McNeil, Duffy, Keith, & Collins, 1990), and from right hemisphere damage (Brauer et al., 1988). The Revised Token Test (RTT) (McNeil & Prescott, 1978) has been used to differentiate left hemisphere damaged-aphasic, right hemisphere damaged-nonaphasic and normal individuals using discriminant analyses (McNeil, Brauer, & Prescott, 1988).

While statistical profile analyses stimulate clinical insight into general areas of function, they provide limited information for generating differential diagnosis standards. Discriminant function studies offer information on the accuracy and bias of a particular test; however, they provide no opportunity to evaluate individual clinicians. It would be useful to have a method of assessing the accuracy and bias for individual clinicians. The theory of signal detection (TSD) quantifies accuracy and response bias for test or individual clinician. This paper (1) outlines the major tenets of TSD, (2) discusses TSD application to aphasia diagnosis, and (3) presents

the results of assessing accuracy and bias in aphasia diagnosis using PICA test data.

OVERVIEW OF TSD AND ITS RELATION TO DIAGNOSTICS

The assumption was made that no diagnostic system, test, or clinician is perfectly accurate, no matter how objective. Moreover, the valid evaluation of the accuracy and bias of diagnostic systems is critical in understanding their potential *clinical value* and diagnostic accuracy must precede the evaluation of treatment efficacy and cost-benefit (Swets, 1988; Swets & Pickett, 1982).

Statistical decision theory, translated into TSD, can be traced to Blackwell in the early fifties (Blackwell, 1953; Blackwell, Prichard, & Ohmart, 1954). For three decades Swets and colleagues (Green & Swets, 1974; Swets, 1959, 1988; Swets & Pickett, 1982) have pioneered the theory's application to diagnostic systems. Originally developed in World War II to assist submarine detection, TSD has been used widely in psychophysical experiments (Green & Swets, 1974; Gescheider, 1985).

In any discrimination task there are two possible states: Noise (N) and Signal-plus-Noise (SN). The observer's job is to determine if SN is present or just N. Noise arises from external sources (the wind, traffic, or human noise such as speech or babble) as well as internal sources (the observer's physiology). Given these two states (N and SN), there are four possible judgment outcomes at any given time. As shown in Table 1, an observer may say SN when SN is present, a *hit* (correct detection or true positive). Or the observer may say N when SN is present, a *miss* (false negative). Likewise, the observer may say SN when N is present, (a *false positive* [*FP*] or false alarm) or the observer may say N when N is present, (a *true negative* [*TN*] or correct rejection).

Knowledge of the raw numbers in each of the cells compared to the total number of trials allows one to calculate the probability of each event. As shown in a hypothetical case in Table 2, there are a total of 200 trials: 100 are SN and 100 are N. The observer said SN 60 times and N 40 times in 100 SN trials. Thus, the probability of a hit, P(hit), is .60 and the probability of a miss, P(miss), is .40. Of the 100 trials in which N was present, the observer said SN 25 times and N 75 times; thus P(FP) is .25 and P(TN) is .75. P(miss) = 1–P(hit) and P(TN) = 1–P(FP). Therefore, the percent of hits and FPs are needed to determine true sensitivity (accuracy) and bias. A point that cannot be overly stressed is that accuracy cannot be assessed based on percent correct (hits) alone; FPs must be accounted for as well. This point will be elaborated below.

TABLE 1. FOUR POSSIBLE JUDGMENT OUTCOMES WHEN PRESENTED WITH SIGNAL PLUS NOISE (SN) OR NOISE (N)

		Subject Response: SN	Subject Response: N
Actual State: SN	SN	hit	miss
Actual State: N	N	FP	TN

TABLE 2. HYPOTHETICAL DATA AND RESPONSE FREQUENCIES WHEN PRESENTED WITH SIGNAL PLUS NOISE (SN) OR NOISE (N)

		Subject Response: SN	Subject Response: N	TOTAL #
Actual State: SN	SN	60	40	100
Actual State: N	N	25	75	100

P (hit) = .60 P (miss) = .40 P (FP) = .25 P (TN) = .75

P (miss) = 1-P (hit) P (TN) = 1-P (FP)

Knowledge of hits, misses, FPs, and TNs (hits and FPs) allows for TDS application to evaluate a diagnostic system. Consider SN to be the abnormal case (e.g., an individual with aphasia) and N to be any other state (e.g., normal). How N is defined can be varied systematically to provide different degrees of precision when evaluating a diagnostic system. Given any diagnostic decision, the clinician is faced with the problem of choosing SN or N, and any of the four outcomes (hits, misses, FP, TN) is possible.

A hypothetical example of diagnostic accuracy will serve to demonstrate the need for a better accuracy measure than percent of hits alone. Clinician Y is 100% correct in the identification of aphasia when aphasia is present (i.e., P[hit] = 1.0 and P[miss] = 0). Clinician Z is only 70% correct

in saying aphasia is present when the person actually has aphasia (P[hit] = .70 and P[miss] = .30). Y, under most traditional criteria, would be considered superior. However, if you were also told that Y identified 90% of the normal persons seen for evaluation as aphasic (i.e., P[FP] = .9 and P[TN] = .1) and Z had an FP rate of only 5% (P[FP] = .05 and P[TN] = .95), you might reconsider your evaluation of the adequacy of these individuals and be forced to conclude that Z was a better diagnostician than Y. Thus, the percent correct alone is an inadequate measure of accuracy. One must account for FPs as well as hits. TSD allows one to derive a better measure of accuracy based on hits and FPs than on percent of hits alone.

Additionally, TSD allows consideration of the bias with which one approaches a task. In TSD, *bias* conventionally refers to the tendency to favor a positive outcome (hits). Both tests and individual clinicians have biases. Numerous factors affect these biases, including: (1) a clinician's background and training, (2) the assumptions a clinician or test (or both) brings to a given diagnostic situation, (3) a clinician's work setting, (4) a clinician's mood, and/or (5) a clinician's cost-benefit assessment of a given situation.

Consider the following hypothetical diagnostic situation. General X is a military radar operator with the job of detecting incoming missiles. General X has family living in the area of his surveillance and wants to make sure that no missiles come close to populated areas, especially those where his family live. Because of his personal and professional situation he adopts a lenient response bias. That is, he wants to make sure that when a missile is detected, he will not miss it. Thus, even though a defensive missile will be launched every time he detects an incoming missile, he launches his defensive missile any time he detects a signal on his radar screen. Because aircraft other than missiles are also detected on the screen, the signal detected may or may not be a missile. As a result, General X has a high FP rate, but never misses an incoming missile. For the General, the cost of missing an incoming missile is greater than the benefit of not taking down an allied or neutral plane. When General X goes off duty, Private A assumes the radar watch. Although the defense of the country from attack is paramount in both General X's and Private A's minds, Private A's family is scheduled to arrive by air sometime during his work shift. He knows that every time he launches defensive missiles he is likely to hit his target because they are very accurate. He also reasons that the attack missiles are fairly inaccurate and the chance of hitting a populated area is low. Private A adopts a bias that dictates that he will launch a defensive missile only when he is absolutely positive that the signal detected on the radar screen is an offensive missile. His hit rate is lower than General X's, but so is his rate of FPs. Private A's bias dictates that the cost of an FP far outweighs the benefit derived from identifying every incoming missile. The distributions of decisions based on N and SN form the data on which TSD is computed.

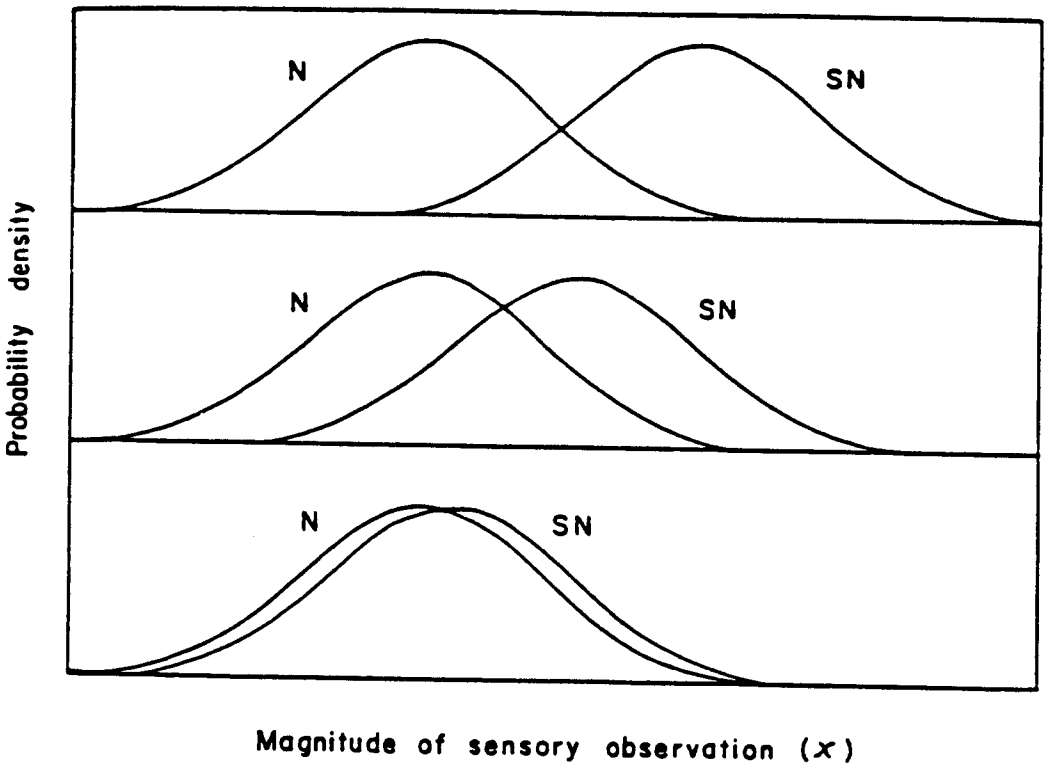

Figure 1. Three different theoretical distributions of noise (N) and signal-plus-noise (SN). Accuracy (d') is defined as the distance between the two peaks of the N and SN distributions. The further the distance between the two peaks, the better the accuracy of and the higher the d'. Thus, the top N and SN distributions represent the highest accuracy and d' is relatively large. The lowest N and SN distributions are almost overlapping, and accuracy is poor; d' is close to 0. Reprinted from G. A. Gescheider, *Psychophysics: Method, theory and application* (2nd ed.), p. 87, with permission of Lawrence Erlbaum Associates, Hillsdale, NJ.

Figure 1 shows two distributions that represent N on the left and SN on the right. The distributions are similar in shape and their variance is equal. (Note that TSD can handle nonparametric data as well.) Notice that the two curves overlap to some degree. That is, in all situations there are instances in which one may report SN when only N is present or only N when SN is present. One main index of accuracy, d', is represented by the difference between the peaks of the N and SN distributions. The greater the overlap of N and SN distributions, the less the distance between the peaks, and the lower the d'. The distributions at the top of Figure 1 have the highest d' and represent the highest accuracy. The bottom distributions with the most overlap have the lowest d' and represent the poorest accuracy. The middle distributions represent an intermediate accuracy

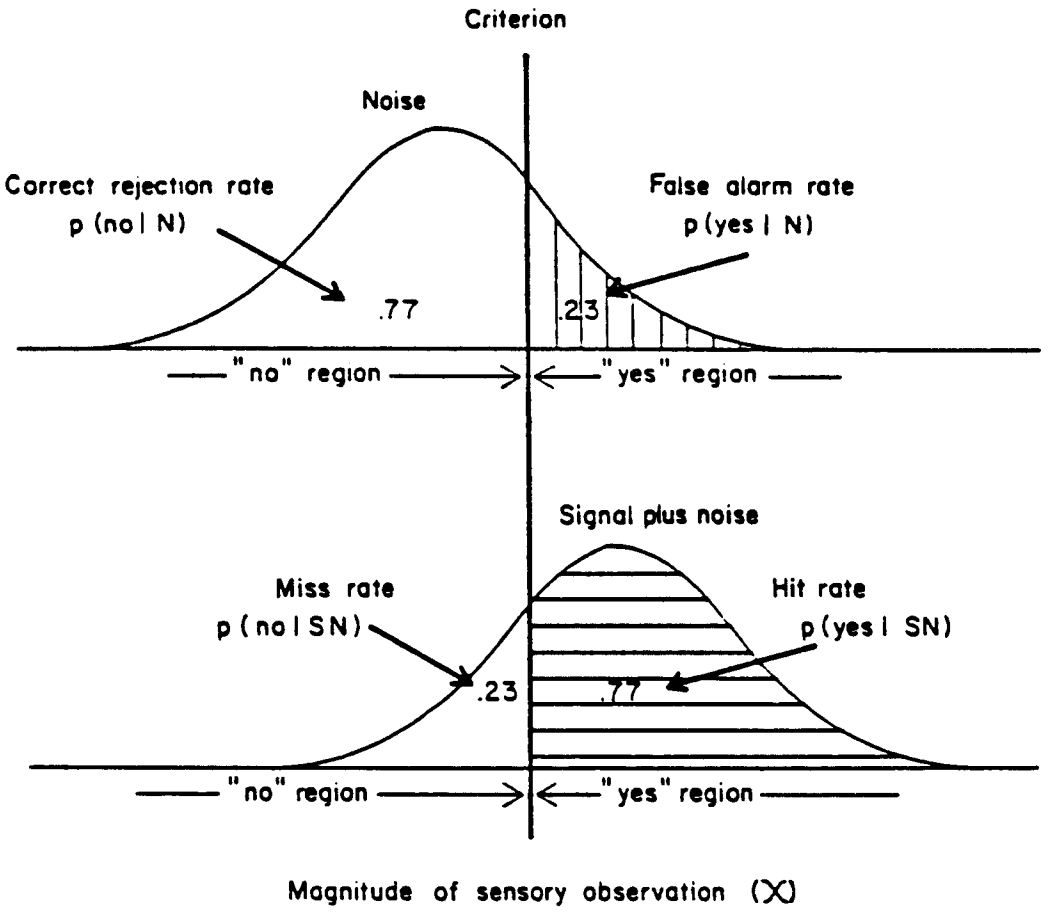

Figure 2. Theoretical distributions of noise (N) on top and signal-plus-noise (SN) on bottom. Note that the distance between the peaks represents accuracy or d'. The perpendicular line represents bias (labeled as criterion in the figure) or β. For responses to the right of the line the clinician says yes, SN present, and to the left of the line the clinician says no, SN not present (i.e., N present). Note the areas of overlap between the two distributions. Thus, all four conditions (Hit, FP, miss, and TN) are represented. Hits (hit rate in figure [P = .77]) are shown by the lined portion of the SN distribution; FPs, false alarm in figure (P = .23), by the lined portion of the N distribution; misses, miss rate in figure (P = .23), by the blank portion of the SN distribution; and TN, correct rejection in figure (P = .77), by the blank portion of the N distribution. As the perpendicular line moves to the left a more lenient bias is adopted, β goes up. In this case both the hits and FPs increase. As the perpendicular line moves to the right, β goes down and a more stringent bias is adopted. In this case both hits and FPs decrease. Reprinted from G. A. Gescheider, *Psychophysics: Method, theory and application* (2nd ed.), p. 90, with permission of Lawrence Erlbaum Associates, Hillsdale, NJ.

level. Accuracy as measured by d′ can range from complete overlap of the distributions (d′ = 0.00) to as high as about 4.67.

Figure 2 represents theoretical N (top) and SN (bottom) distributions. The perpendicular line represents the response bias that in TSD terms is called β . Responses to the right of β are positive for SN (clinician says "yes," SN present) whereas those to the left favor N (clinician says "no"). At the intersection of the two distributions β is 1, which refers to equal bias for SN and N. As β decreases, a more lenient criterion is adopted (hits and FPs increase) and as β increases a more stringent criterion is adopted (hits and FPs decrease). Also note that the areas of the distribution that represent different cells in contingency Table 1 are shown in this figure. These probabilities depend on one's bias. Moreover, accuracy does not change as a function of shifts in bias, although the relative percent of hits and FPs do.

One other aspect of TSD needs to be discussed. In order to apply TSD to diagnostic situations, receiver operating characteristic (ROC) curves—a plot of the P(hits) by the P(FP)—need to be developed. The measure d′ reflects the distance between the N and SN distributions, which is reflected in the position of the ROC curve on the graph. Figure 3 is a graph of an ROC curve with a d′ of 1.0. The diagonal straight line is a d′ of zero, where the two distributions have complete overlap. The N and SN distributions that give rise to the position of the ROC curve are shown along with three different biases (βs) that lie along the curve. Different points along the curve represent differences in bias or β , with the same d′. Thus, d′ is an accuracy measure independent of bias. The construction of ROC curves allows an assessment of accuracy *and* bias. The uppermost N and SN distributions have the bias line far to the left, indicating a low β and a lenient bias. Moving down the figure to the lower distributions, β increases, indicating more stringent biases. Figures 4 and 5 show ROC curves for d′s of 2.0 and 0 respectively, each with three bias points.

If diagnosticians' data are shown to fit the theoretical curves, then d′ and β may be used without developing the entire curve. Convenient tables for d′ and β exist for any known hit and FP rate. To illustrate TSD application to aphasia diagnostics and to explore its value, ROC curves, d′, and β data for individual clinicians using the PICA to diagnose aphasia are presented.

METHOD

Judges

Four of the judges were formally trained and experienced in the administration, scoring, and interpretation of the PICA-evaluated item, subtest,

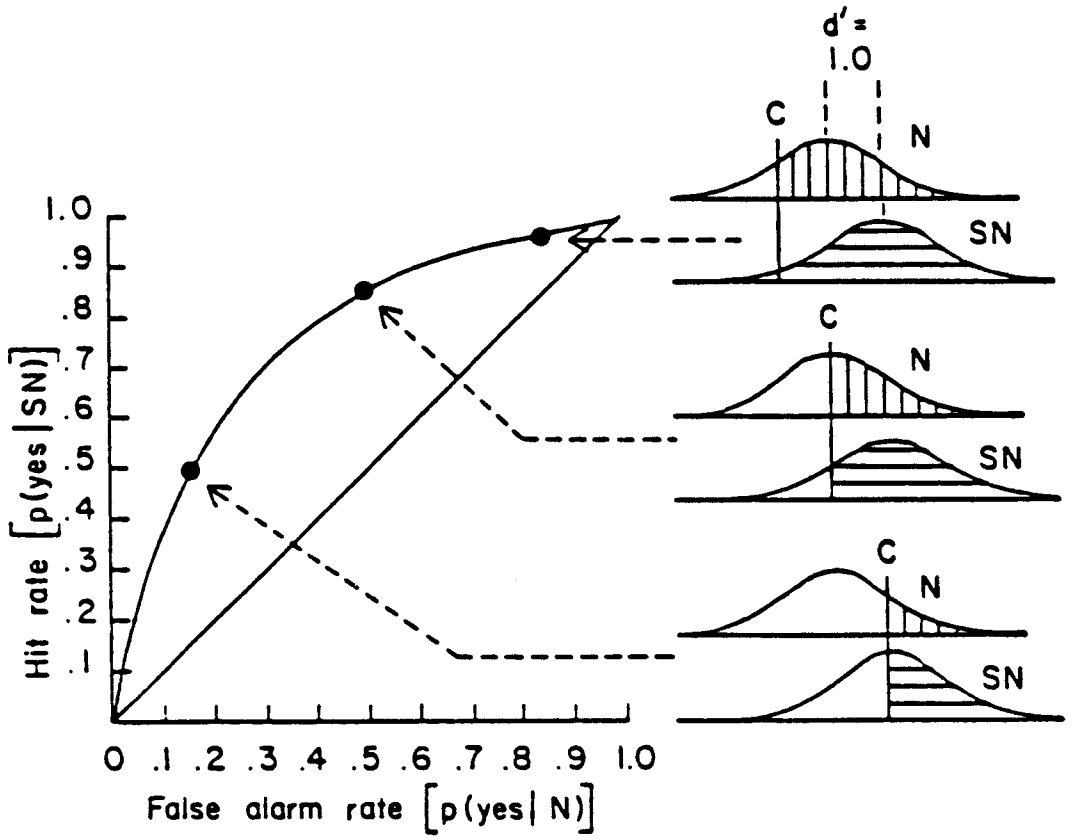

Figure 3. Relation between the ROC curve and the theoretical N and SN distributions that represent a d' of 1.0. The straight diagonal represents a d' of 0 as hits and FPs are equal. Each of the three data points represent a different bias at the same accuracy level (the distance between the peaks of the N and SN distributions is constant), only the perpendicular line moves. The most lenient bias, highest β is represented by the upper distributions for which both hit and FP rates are high. The most stringent bias, lowest β is represented by the bottom distributions in which both hit and FP rates are low. Reprinted from G. A. Gescheider, *Psychophysics: Method, theory and application* (2nd ed.), p. 95, with permission of Lawrence Erlbaum Associates, Hillsdale, NJ.

and overall scores from 246 PICA tests. A fifth judge (subject 3) was a PICA-trained but generally inexperienced graduate student seeking a master's degree in speech-language pathology.

PICA Data

To determine the accuracy of responses, one needs to differentiate, a priori, PICA tests derived with normal subjects from those that used aphasic individuals. In other words, a *gold standard* for diagnosis is required.

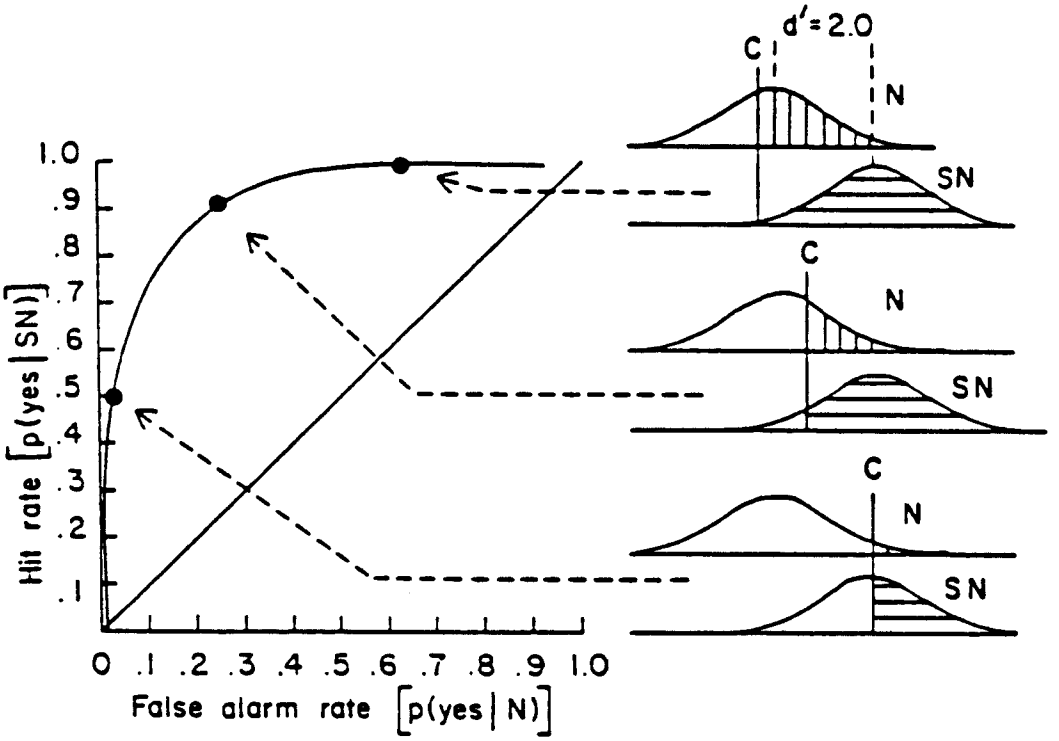

Figure 4. Same as Figure 3, but d' = 2.0. Reprinted from G. A. Gescheider, *Psychophysics: Method, theory and application* (2nd ed.), p. 94, with permission of Lawrence Erlbaum Associates, Hillsdale, NJ.

Although much theory and philosophy underlie the final categorization, only the general criteria and methods used for making the decisions are discussed. First, the normal subject's PICA performance was provided by Duffy and Keith from their 1980 standardization for normal individuals. Subjects met the criteria for "normal" as defined in that investigation. The aphasic PICA data were derived from the speech-language pathology patient records at the VA hospital in Madison WI. The aphasia diagnosis was made or certified by either Wertz, Rosenbek, or Collins. Although the PICA was used in the patient's diagnosis, in no case was the categorization based solely on PICA data. A variety of other standardized, unstandardized, and informal assessment tools, interviews, case histories, and medical records were also used to reach the final diagnosis. Aphasia was defined in a manner similar to Darley (1982). That is, all aphasic subjects had multimodality language deficits that were disproportionate to other intellective deficits. Onset was sudden and secondary to a focal dominant hemisphere lesion. Of the 246 PICA tests evaluated, 103 were from aphasic (SN distribution) and 143 from normal non-brain-damaged individuals (N distribution).

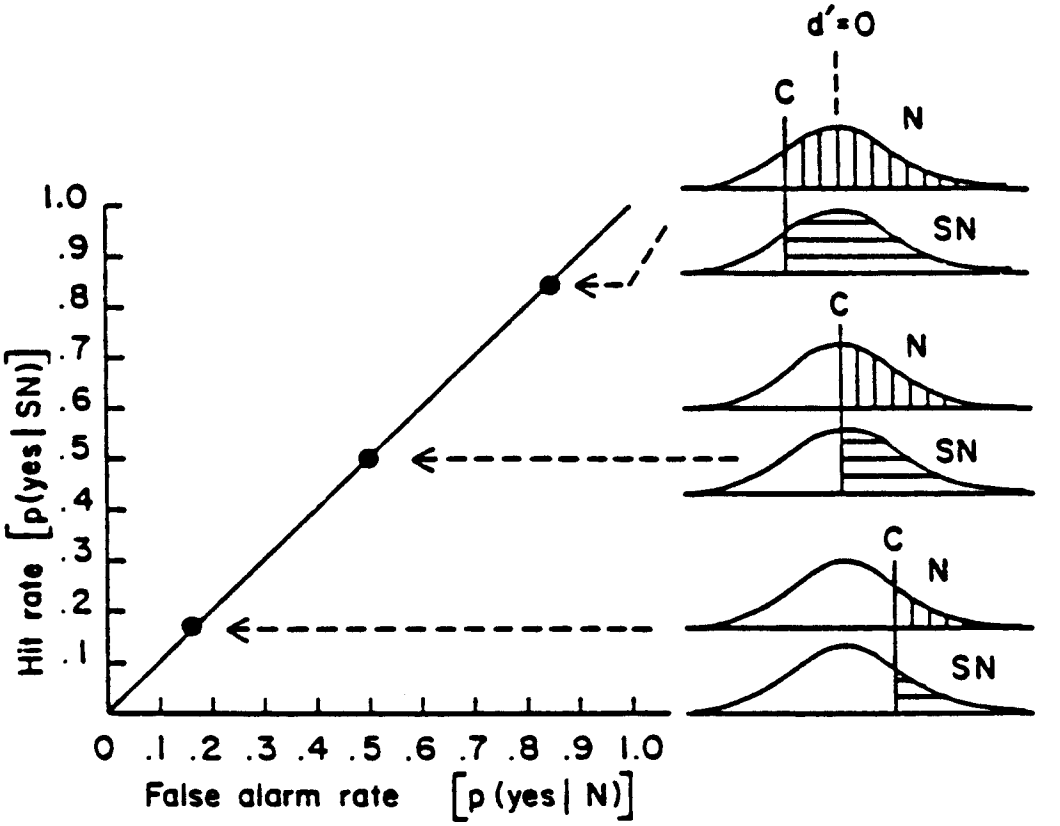

Figure 5. Same as Figure 3, but d' = 0. Reprinted from G. A. Gescheider, *Psychophysics: Method, theory and application* (2nd ed.), p. 96, with permission of Lawrence Erlbaum Associates, Hillsdale, NJ.

Judges were first required to indicate whether the test presented was from an aphasic or normal individual (subjects said yes for aphasia and no for normal). The clinicians then rated the confidence of their decisions on a five-point equal-appearing interval scale from 1 (not very sure) to 5 (very sure). The rating scale method is a cost effective means of applying TSD to clinical situations to generate the ROC curve and test the validity of the approach (Swets, 1988; Swets & Pickett, 1982). Each point on the rating scale represents a different bias. An individual who is very sure of a decision will have a stringent bias (β will be high) while the bias associated with a rating of not very sure will be very lenient (β will be low). In this manner up to four points along an ROC curve, each with differing bias, were obtained. The maximum number of points is four as the most lenient criterion associated with the rating of 1 will always result in a P(hit) and P(FP) equaling 1.0.

From the rating scale data, the P(hit) and P(FP) for each person at the different biases were determined. ROC curves for each person were plotted and the data were compared to theoretical curves. Accuracy and bias mea-

sures were derived from these. The d′ and β for each person, representing their typical levels, were determined.*

RESULTS AND DISCUSSION

Table 3 shows each subject's d′ and β for the diagnostic decision (yes/no portion of the study). These data generally are thought to reflect the typical accuracy and bias of each subject. High accuracy is indicated by a d′ better than 3.5. None of the subjects were close to achieving a high accuracy. The highest d′ was 2.79. This judge, subject 5, had the least experience with aphasia and the PICA other than the student. Subject 4 with approximately 12 years of PICA experience had a very low d′ of 1.0. At this low level of accuracy, at typical bias, this judge had a P(hit) of .86 and a P(FP) of .62. Thus, this judge identified 86% of the aphasic tests as aphasic but also identified 63% of the normal tests as aphasic. By contrast, subject 5, who had the highest accuracy, had a P(hit) of .89 (89% correct aphasic identification) but a P(FP) of only .06 (6% incorrect identification of normal subjects) at the typical bias. The student clinician (subject 3) had an accuracy level similar to most of the other subjects.

Table 3 also shows the bias with which the subjects typically approached the task. While β s were generally indicative of a relatively stringent bias (>1.0), there was a range of biases among the subjects. Subject 4 had a very lenient bias ($\beta = .46$) while Subject 1 was fairly stringent ($\beta = 4.10$). Thus, some individuals were more willing than others to identify an individual as having aphasia given the PICA information they reviewed.

In summary, TSD provided a method of investigating accuracy and bias in aphasia diagnosis. PICA tests alone did not result in acceptable accuracy levels given the extant literature suggesting that a d′ greater than 3.5 is achieved in most good diagnostic systems. However, individual clinicians can use TSD to check their own bias and accuracy. Future studies could systematically evaluate accuracy and bias when differing amounts of information are available to the diagnostician. For instance, accuracy and bias could be determined when other standardized aphasia test data are added to the PICA's. Accuracy and bias could then be determined when a case history or a brain scan is part of the information provided. One could vary *noise* by asking judges to determine if aphasia is present or if the patient is normal, has dementia, or has right hemisphere involve-

*The computer program to calculate the best fit for the data and determine accuracy and bias were kindly provided by Donald D. Dorfman and Kevin Burnbaum at the University of Iowa in the Departments of Psychology and Radiology, respectively. The original program can be found in Dorfman and Alf (1969).

TABLE 3. ACCURACY (d') AND RESPONSE BIAS (β) FOR EACH SUBJECT AT *TYPICAL* RESPONSE LEVEL

Subj #	*d'*	β
1	1.96	4.10
2	2.66	2.56
3	2.42	3.60
4	1.00	0.46
5	2.79	1.72

Note: Subject 3 was a student.

ment; accuracy and bias vary with the amount of information provided and the amount of noise in the system.

TSD is also useful in cost-benefit assessment. One can objectively determine which measures give the highest accuracy. The minimum amount of information that provides the highest accuracy measures can be determined for any given diagnostic decision.

This paper examined the basic principles of TSD and its application to aphasia diagnosis. Future studies will examine if the assumptions of d' (equal variance of N and SN) hold true. If not, other measures of accuracy may be more appropriate. Moreover, instead of using β as the bias measure, $\ln\beta$ (natural log of β) may prove to be a better metric, as lenient biases are indicated by βs between 0 and 1.0 and stringent biases by βs greater than 1.0. The investigation of these and several related methodological issues in the application of TSD to aphasia assessment should help clinical aphasiologists select the most sensitive and specific tools and procedures for their clinical practices.

REFERENCES

Blackwell, H. R. (1953). Psychophysical thresholds: Experimental studies of methods of measurement. University of Michigan's *Bulletin of the Engineering Research Institute, 53.*

Blackwell, H. R., Prichard, B. S., & Ohmart, T. G. (1954). Automatic apparatus for stimulus presentation and recording in visual threshold experiments. *Journal of the Ophthalmology Society of America, 44,* 322–326.

Brauer, D., McNeil, M. R., Collins, M. J., Deal, J. L., Duffy, J. R., & Keith, R. (1988). Discriminating normal, aphasic and right hemisphere performance with the PICA. [Abstract] *ASHA,* 1988, *30*(10), 173.

Brauer, D., McNeil, M. R., Duffy, J. R., Keith R. L., & Collins M. J. (1990). The differentiation of normal from aphasic performance using PICA discriminant function scores. In T. E. Prescott (Ed.), *Clinical Aphasiology* (Vol. 18, pp. 117–129). Austin, TX: PRO-ED.

Darley, F. L. (1982). *Aphasia.* Philadelphia, PA: W. B. Sanders.

Dorfman, D. D., & Alf, E., Jr. (1969). Maximum likelihood estimates of parameters of signal-detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology, 6,* 487–496.

Duffy, J. R., & Keith, R. (1980). Performance of non-brain injured adults on the PICA: Descriptive data and a comparison to patients with aphasia. *Aphasia-Apraxia-Agnosia. 2*(2), 1–30.

Gescheider, G. A. (1985). *Psychophysics: Method, theory, and application* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics.* Huntington, NY: Krieger. (Original work published in 1966, New York: Wiley).

Halpern, H., Darley, F. L., & Brown J. R. (1973). Differential language and neurological characteristics in cerebral involvement. *Journal of Speech and Hearing Disorders, 38,* 162–173.

McNeil, M. R., Brauer, D., & Prescott, T. E. (1988). Discriminating normal, aphasic and right hemisphere performance with the RTT. [Abstract] *ASHA, 30:* (10), 161.

McNeil, M. R., & Prescott, T. E. (1978). *Revised Token Test.* Austin, TX: PRO-ED.

Porch, B. E. (1971). *Porch Index of Communicative Ability.* Palo Alto: Consulting Psychologists Press.

Porch, B. E., Frieden, T., & Porec, J. (1977). Objective differentiation of aphasic versus nonorganic patients. Paper presented to the International Neuropsychology Association, Santa Fe, NM.

Swets, J. A. (1959). Indices of signal detectability obtained with various psychophysical procedures. *Journal of the Acoustical Society of America, 31,* 511–513.

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

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

Wertz, R. T., & Rosenbek, J. C. (1971). Appraising apraxia of speech. *Journal of the Colorado Speech and Hearing Association, 5,* 18–36.