Psychological Bulletin, 1955, **52**, 194–216.

#035

Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores

Paul E. Meehl & Albert Rosen[1]

In clinical practice, psychologists frequently participate in the making of vital decisions concerning the classification, treatment, prognosis, and disposition of individuals. In their attempts to increase the number of correct classifications and predictions, psychologists have developed and applied many psychometric devices, such as patterns of test responses as well as cutting scores for scales, indices, and sign lists. Since diagnostic and prognostic statements can often be made with a high degree of accuracy purely on the basis of actuarial or experience tables (referred to hereinafter as *base rates*), a psychometric device, to be efficient, must make possible a greater number of correct decisions than could be made in terms of the base rates alone. The efficiency of the great majority of psychometric devices reported in the clinical psychology literature is difficult or impossible to evaluate for the following reasons:

1. Base rates are virtually never reported. It is, therefore, difficult to determine whether or not a given device results in a greater number of correct decisions than would be possible solely on the basis of the rates from previous experience. When, however, the base rates can be estimated, the reported claims of efficiency of psychometric instruments are often seen to be without foundation.
2. In most reports, the distribution data provided are insufficient for the evaluation of the probable efficiency of the device in other settings where the base rates are markedly different. Moreover, the samples are almost always too small for the determination of optimal cutting lines for various decisions.
3. Most psychometric devices are reported without cross-validation data. If a psychometric instrument is applied solely to the criterion groups from which it was developed, its reported validity and efficiency are likely to be spuriously high, especially if the criterion groups are small.
4. There is often a lack of clarity concerning the type of population in which a psychometric device can be effectively applied.
5. Results are frequently reported only in terms of significance tests for differences between groups rather than in terms of the number of correct decisions for individuals within the groups.

The purposes of this paper are to examine current methodology in studies of predictive and concurrent validity (APA, Committee on Test Standards 1954), and to present some methods for the evaluation of the efficiency of psychometric devices as well as for the improvement in the interpretations made from such devices. Actual studies reported in the literature will be used for illustration wherever possible. It should be emphasized that these particular illustrative studies of common practices were chosen simply because they

---

[1] The senior author carried on his part of this work in connection with his appointment to the Minnesota Center for the Philosophy of Science at the University of Minnesota. The junior author was on the Neuropsychiatric Service, VA Hospital, Minneapolis, Minnesota, and in the Divisions of Psychiatry and Clinical Psychology of the University of Minnesota Medical School.

contained more complete data than are commonly reported, and were available in fairly recent publications.

Importance of Base Rates

Danielson and Clark (1954) have reported on the construction and application of a personality inventory which was devised for use in military induction stations as an aid in detecting those men who would not complete basic training because of psychiatric disability or AWOL recidivism. One serious defect in their article is that it reports cutting lines which have not been cross validated. Danielson and Clark state that inductees were administered the Fort Ord Inventory within two days after induction into the Army, and that all of these men were allowed to undergo basic training regardless of their test scores.

Two samples (among others) of these inductees were selected for the study of predictive validity: (a) A group of 415 men who had made a good adjustment (Good Adjustment Group), and (b) a group of 89 men who were unable to complete basic training and who were sufficiently disturbed to warrant a recommendation for discharge by a psychiatrist (Poor Adjustment Group). The authors state that "the most important task of a test designed to screen out misfits is the detection of the (latter) group" (Danielson & Clark, 1954, p. 139). The authors found that their most effective scale for this differentiation picked up, at a given cutting point, 55% of the Poor Adjustment Group (valid positives) and 19% of the Good Adjustment Group (false positives). The overlap between these two groups would undoubtedly have been greater if the cutting line had been cross validated on a random sample from the *entire population* of inductees, but for the purposes of the present discussion, let us assume that the results were obtained from cross-validation groups. There is no mention of the percentage of all inductees who fall into the Poor Adjustment Group, but a rough estimate will be adequate for the present discussion. Suppose that in their population of soldiers, as many as 5% make a poor adjustment and 95% make a good adjustment. The results for 10,000 cases would be as depicted in Table 1.

TABLE 1

Number of Inductees in the Poor Adjustment and Good Adjustment Groups Detected By a Screening Inventory

(55% Valid Positives; 19% False Positives)

Predicted Adjustment	Actual Adjustment				Total Predicted
	Poor		Good		
	No.	%	No.	%	
Poor	275	55	1,805	19	2,080
Good	225	45	7,695	81	7,920
Total Actual	500	100	9,500	100	10,000

Efficiency in detecting poor adjustment cases. The efficiency of the scale can be evaluated in several ways. From the data in Table 1 it can be seen that if the cutting line given by

the authors were used at Fort Ord, the scale could not be used directly to "screen out misfits." If all those predicted by the scale to make a poor adjustment were screened out, the number of false positives would be extremely high. Among the 10,000 potential inductees, 2080 would be predicted to make a poor adjustment. Of these 2080, only 275, or 13%, would actually make a poor adjustment, whereas the decisions for 1805 men, or 87% of those screened out, would be incorrect.

Efficiency in prediction for all cases. If a prediction were made for every man on the basis of the cutting line given for the test, 275 + 7695, or 7970, out of 10,000 decisions would be correct. Without the test, however, every man would be predicted to make a good adjustment, and 9500 of the predictions would be correct. Thus, use of the test has yielded a drop from 95% to 79.7% in the total number of correct decisions.

Efficiency in detecting good adjustment cases. There is one kind of decision in which the Inventory can improve on the base rates, however. If only those men are accepted who are predicted by the Inventory to make a good adjustment, 7920 will be selected, and the outcome of 7695 of the 7920, or 97%, will be predicted correctly. This is a 2% increase in hits among predictions of "success." The decision as to whether or not the scale improves on the base rates sufficiently to warrant its use will depend on the cost of administering the testing program, the administrative feasibility of rejecting 21% of the men who passed the psychiatric screening, the cost to the Army of training the 225 maladaptive recruits, and the intangible human costs involved in psychiatric breakdown.

Populations to which the scale is applied. In the evaluation of the efficiency of any psychometric instrument, careful consideration must be given to the types of populations to which the device is to be applied. Danielson and Clark (1954, p. 138) have stated that "since the final decision as to disposition is made by the psychiatrist, the test should be classified as a screening adjunct." This statement needs clarification, however, for the efficiency of the scale can vary markedly according to the different ways in which it might be used as an adjunct.

It will be noted that the test was administered to men who were already in the Army, and not to men being examined for induction. The reported validation data apply, therefore, specifically to the population of *recent inductees*. The results might have been somewhat different if the population tested consisted of *potential inductees*. For the sake of illustration, however, let us assume that there is no difference in the test results of the two populations.

An induction station psychiatrist can use the scale cutting score in one or more of the following ways, that is, he can apply the scale results to a variety of populations. (a) The psychiatrist's final decision to accept or reject a potential inductee may be based on both the test score and his usual interview procedure. The population to which the test scores are applied is, therefore, *potential inductees interviewed by the usual procedures for whom no decision was made*. (b) He may evaluate the potential inductee according to his usual procedures, and then consult the test score *only if* the tentative decision is to reject. That is, a decision to accept is final. The population to which the test scores are applied is *potential inductees tentatively rejected by the usual interview procedures*. (c) An alternative procedure is for the psychiatrist to consult the test score only if the tentative decision is to accept, the population being *potential inductees tentatively accepted by the usual interview procedures*. The decision to reject is final. (d) Probably the commonest proposal for the use of tests as screening adjuncts is that the more skilled and costly psychiatric evaluation should be made only upon the test positives, that is, inductees classified by the test as good risks are not interviewed, or are subjected only to a very short and superficial interview. Here the population is *all potential inductees*, the test being used to make either a *final* decision to "accept" or a decision to "examine."

Among these different procedures, how is the psychiatrist to achieve maximum effectiveness in using the test as an adjunct? There is no answer to this question from the available data, but it can be stated definitely that the data reported by Danielson and Clark apply only to the third procedure described above. The test results are based on a selected group of men *accepted* for induction and not on a random sample of potential inductees. If the scale is used in any other way than the third procedure mentioned above, the results may be considerably inferior to those reported, and, thus, to the use of the base rates without the test.[2]

The principles discussed thus far, although illustrated by a single study, can be generalized to any study of predictive or concurrent validity. It can be seen that many considerations are involved in determining the efficiency of a scale at a given cutting score, especially the base rates of the subclasses within the population to which the psychometric device is to be applied. In a subsequent portion of this paper, methods will be presented for determining cutting points for maximizing the efficiency of the different types of decisions which are made with psychometric devices.

Another study will be utilized to illustrate the importance of an explicit statement of the base rates of population subgroups to be tested with a given device. Employing an interesting configural approach, Thiesen (1952) discovered five Rorschach patterns, each of which differentiated well between 60 schizophrenic adult patients and a sample of 157 gainfully employed adults. The best differentiator, considering individual patterns or number of patterns, was Pattern A, which was found in 20% of the patients' records and in only .6% of the records of normals. Thiesen concludes that if these patterns stand the test of cross validation, they might have "clinical usefulness" in early detection of a schizophrenic process or as an aid to determining the gravity of an initial psychotic episode (Thiesen, 1952, p. 369). If by "clinical usefulness" is meant efficiency in a clinic or hospital for the diagnosis of schizophrenia, it is necessary to demonstrate that the patterns differentiate a higher percentage of schizophrenic patients from *other diagnostic groups* than could be correctly classified without any test at all, that is, solely on the basis of the rates of various diagnoses in any given hospital. If a test is to be used in differential diagnosis among psychiatric patients, evidence of its efficiency for this function cannot be established solely on the basis of discrimination of diagnostic groups from normals. If by "clinical usefulness" Thiesen means that his data indicate that the patterns might be used to detect an early schizophrenic process among nonhospitalized gainfully employed adults, he would do better to discard his patterns and use the base rates, as can be seen from the following data.

Taulbee and Sisson (1954) cross validated Thiesen's patterns on schizophrenic patient and normal samples, and found that Pattern A was the best discriminator. Among patients, 8.1% demonstrated this pattern, and among normals, none had this pattern. There are approximately 60 million gainfully employed adults in this country, and it has been estimated that the rate of schizophrenia in the general population is approximately .85% (Anastasi & Foley, 1949, p. 558). The results for Pattern A among a population of 10,000 gainfully employed adults would be as shown in Table 2. In order to detect 7 schizophrenics, it would be necessary to test 10,000 individuals.

In the Neurology service of a hospital a psychometric scale is used which is designed to differentiate between patients with psychogenic and organic low back pain (Hanvik, 1949). At a given cutting point, this scale was found to classify each group with approximately 70% effectiveness upon cross validation, that is, 70% of cases with no organic findings scored above an optimal cutting score, and 70% of surgically verified organic cases scored

[2] Goodman (1953) has discussed this same problem with reference to the supplementary use of an index for the prediction of parole violation.

below this line. Assume that 90% of all patients in the Neurology service with a primary complaint of low back pain are in fact "organic." Without any scale at all the psychologist can say every case is organic, and be right 90% of the time. With the scale the results would be as shown in Section A of Table 3. Of 10 psychogenic cases, 7 score above the line; of 90 organic cases, 63 score below the cutting line. If every case above the line is called psychogenic, only 7 of 34 will be classified correctly, or about 21%. Nobody wants to be right only one out of five times in this type of situation, so that it is obvious that it would be imprudent to call a patient psychogenic on the basis of this scale. Radically different results occur in prediction for cases below the cutting line. Of 66 cases 63, or 95%, are correctly classified as organic. Now the psychologist has increased his diagnostic hits from 90 to 95% on the condition that he labels only cases falling below the line, and ignores the 34% scoring above the line.

TABLE 2

Number of Persons Classified as Schizophrenic and Normal
by a Test Pattern Among a Population of Gainfully Employed Adults
(8.1% valid positives; 0.0% false positives)

Classification by Test	Criterion Classification				Total Classified by Test
	Schizophrenia		Normal		
	No.	%	No.	%	
Schizophrenia	7	8.1	0	0	7
Normal	78	91.9	9,915	100	9,993
Total in class	81	100	9,915	100	10,000

TABLE 3

Number of Patients Classified as Psychogenic and Organic on a Low Back Pain Scale
Which Classifies Correctly 70% of Psychogenic and Organic Cases

Classification by Scale	Actual Diagnosis		Total Classified by Scale
	Psychogenic	Organic	
A. Base Rates in Population Tested: 90% Organic; 10% Psychogenic			
Psychogenic	7	27	34
Organic	3	63	66
Total diagnosed	10	90	100
B. Base Rates in Population Tested: 90% Psychogenic; 10% Organic			
Psychogenic	63	3	66
Organic	27	7	34
Total diagnosed	90	10	100

In actual practice, the psychologist may not, and most likely will not, test every low back pain case. Probably those referred for testing will be a select group, that is, those who the neurologist believes are psychogenic because neurological findings are minimal or absent.

This fact changes the population from "all patients in Neurology with a primary complaint of low back pain," to "all patients in Neurology with a primary complaint of low back pain *who are referred for testing*." Suppose that a study of past diagnoses indicated that of patients with minimal or absent findings, 90% were diagnosed as psychogenic and 10% as organic. Section B of Table 3 gives an entirely different picture of the effectiveness of the low back pain scale, and new limitations on interpretation are necessary. Now the scale correctly classifies 95% of all cases above the line as psychogenic (63 of 66), and is correct in only 21% of all cases below the line (7 of 34). In this practical situation the psychologist would be wise to refrain from interpreting a low score.

From the above illustrations above it can be seen that the psychologist in interpreting a test and in evaluating its effectiveness must be very much aware of the population and its subclasses and the base rates of the behavior or event with which he is dealing at any given time.

It may be objected that no clinician relies on just one scale but would diagnose on the basis of a configuration of impressions from several tests, clinical data and history. We must, therefore, emphasize that the preceding single-scale examples were presented for simplicity only, but that the main point is not dependent upon this "atomism." *Any complex configurational procedure in any number of variables, psychometric or otherwise, eventuates in a decision*. Those decisions have a certain objective success rate in criterion case identification; and for present purposes we simply treat the decision function, whatever its components and complexity may be, as a single variable. It should be remembered that the literature does not present us with cross-validated methods having hit rates much above those we have chosen as examples, regardless of how complex or configural the methods used. So that even if the clinician approximates an extremely complex configural function "in his head" before classifying the patient, for purposes of the present problem this complex function is treated as the scale. In connection with the more general "philosophy" of clinical decision making see Bross (1953) and Meehl (1954/1996).

Applications of Bayes' Theorem

Many readers will recognize the preceding numerical examples as essentially involving a principle of elementary probability theory, the so-called "Bayes' Theorem." While it has come in for some opprobrium on account of its connection with certain pre-Fisherian fallacies in statistical inference, as an algebraic statement the theorem has, of course, nothing intrinsically wrong with it and it does apply in the present case. One form of it may be stated as follows: If there are k antecedent conditions under which an event of a given kind may occur, these conditions having the antecedent probabilities $P_1, P_2, \ldots P_k$ of being realized, and the probability of the event upon each of them is $p_1, p_2, p_3, \ldots p_k$; then, given that the event is observed to occur, the probability that it arose on the basis of a specified one, say j, of the antecedent conditions is given by

$$P_{j(o)} = \frac{P_j p_j}{\sum_{i=1}^{k} P_i p_i}.$$

The usual illustration is the case of drawing marbles from an urn. Suppose we have two urns, and the urn-selection procedure is such that the probability of our choosing the first urn is $1/10$ and the second $9/10$. Assume that 70% of the marbles in the first urn are black, and 40% of those in the second urn are black. I now (blindfolded) "choose" an urn and then, from it, I choose a marble. The marble turns out to be black. What is the probability that I drew from the first urn?

$$P_1 = .10 \qquad P_2 = .90$$
$$p_1 = .70 \qquad p_2 = .40$$

Then

$$P_{1(b)} = \frac{(.10)(.70)}{(.10)(.70)+(.90)(.40)} = .163.$$

If I make a practice of inferring under such circumstances that an observed *black marble* arose from the first urn, I shall be correct in such judgments, in the long run, only 16.3% of the time. Note, however, that the "test item" or "sign" *black marble* is correctly "scored" in favor of Urn No. 1, since there is a 30% difference in black marble rate between it and Urn No. 2. But this considerable disparity in symptom rate is overcome by the very low base rate ("antecedent probability of choosing from the first urn"), so that inference to first-urn origin of black marbles will actually be wrong some 84 times in 100. In the clinical analogue, the urns are identified with the subpopulations of patients to be discriminated (their antecedent probabilities being equated to their base rates in the population to be examined), and the black marbles are test results of a certain ("positive") kind. The proportion of black marbles in one urn is the valid positive rate, and in the other is the false positive rate. Inspection and suitable manipulations of the formula for the common two-category case, viz.,

$$P_{d(o)} = \frac{Pp_1}{Pp_1 + Qp_2}$$

$P_{d(o)}$ = Probability that an individual is diseased, given that his observed test score is positive
P = Base rate of actual positives in the population examined
$P + Q$ = 1
p_1 = Proportion of diseased identified by test ("valid positive" rate)
q_1 = $1 - p_1$
p_2 = Proportion of nondiseased misidentified by test as being diseased ("false positive" rate)
q_2 = $1 - p_2$

yields several useful statements. Note that in what follows we are operating entirely with exact population parameter values; that is, sampling errors are not responsible for the dangers and restrictions set forth. See Table 4.

1. In order for a positive diagnostic assertion to be "more likely true than false," the ratio of the positive to the negative base rates in the examined population must exceed the ratio of the false positive rate to the valid positive rate. That is,

$$\frac{P}{Q} > \frac{p_2}{p_1}.$$

If this condition is not met, the attribution of pathology on the basis of the test is more probably in error than correct, *even though the sign being used is valid* (i.e., $p_1 \neq p_2$).

Example: If a certain cutting score identifies 80% of patients with organic brain damage (high scores being indicative of damage) but is also exceeded by 15% of the nondamaged sent for evaluation, in order for the psychometric decision "brain damage present" to be more often true than false, the ratio of actually braindamaged to nondamaged cases among all seen for testing must be at least one to five (.19).

TABLE 4

Definition of Symbols

Diagnosis from Test	Actual Diagnosis	
	Positive	Negative
Positive	p_1 Valid positive rate (Proportion of positives called positive)	p_2 False positive rate (Proportion of negatives called positive)
Negative	q_1 False negative rate (Proportion of positives called negative)	q_2 Valid negative rate (Proportion of negatives called negative)
Total with actual diagnosis	$p_1 + q_1 = 1.0$ (Total positives)	$p_2 + q_2 = 1.0$ (Total negatives)

Note.— For simplicity, the term "diagnosis" is used to denote the classification of any kind of pathology, behavior, or event being studied, or to denote "outcome" if a test is used for prediction. Since horizontal addition (e.g., $p_1 + p_2$) is meaningless in ignorance of the base rates, there is no symbol or marginal total for these sums. *All values are parameter values*.

Piotrowski has recommended that the presence of 5 or more Rorschach signs among 10 "organic" signs is an efficient indicator of brain damage. Dorken and Kral (1952), in cross validating Piotrowski's index, found that 63% of organics and 30% of a mixed, nonorganic, psychiatric patient group had Rorschachs with 5 or more signs. Thus, our estimate of $p_2 / p_1 = .30 / .63 = .48$, and in order for the decision "brain damage present" to be correct more than one-half the time, the proportion of positives (P) in a given population must exceed .33 (i.e., $P / Q > .33 / .67$). Since few clinical populations requiring this clinical decision would have such a high rate of brain damage, especially among psychiatric patients, the particular cutting score advocated by Piotrowski will produce an excessive number of false positives, and the positive diagnosis will be more often wrong than right. Inasmuch as the base rates for any given behavior or pathology differ from one clinical setting to another, *an inflexible cutting score should not be advocated for any psychometric device*. This statement applies generally—thus, to indices recommended for such diverse purposes as the classification or detection of deterioration, specific symptoms, "traits," neuroticism, sexual aberration, dissimulation, suicide risk, and the like. When P is small, it may be advisable to explore the possibility of dealing with a restricted population within which the base rate of the attribute being tested is higher. This approach is discussed in an article by Rosen (1954) on the detection of suicidal patients in which it is suggested that an attempt might be made to apply an index to sub-populations with higher suicide rates.

2. If the base rates are equal, the probability of a positive diagnosis being correct is the ratio of valid positive rate to the sum of valid and false positive rates. That is,

$$p_{d(o)} = \frac{p_1}{p_1 + p_2} \quad \text{if } P = Q = ½.$$

Example: If our population is evenly divided between neurotic and psychotic patients the condition for being "probably right" in diagnosing psychosis by a certain method is simply that the psychotics exhibit the pattern in question more frequently than the neurotics. This is the intuitively obvious special case; it is often misgeneralized to justify use of the test in

those cases where base-rate asymmetry ($P \neq Q$) counteracts the ($p_1 - p_2$) discrepancy, leading to the paradoxical consequence that *deciding on the basis of more information can actually worsen the chances of a correct decision*. The apparent absurdity of such an idea has often misled psychologists into behaving as though the establishment of "validity" or "discrimination," that is, that $p_1 \neq p_2$, indicates that a procedure should be used in decision making.

Example: A certain test is used to select those who will continue in outpatient psychotherapy (positives). It correctly identifies 75% of these good cases but the same cutting score picks up 40% of the poor risks who subsequently terminate against advice. Suppose that in the past experience of the clinic 50% of the patients terminated therapy prematurely. Correct selection of patients can be made with the given cutting score on the test 65% of the time, since $p_1 / (p_1 + p_2) = .75 / (.75 + .40) = .65$. It can be seen that the efficiency of the test would be exaggerated if the base rate for continuation in therapy were actually .70, but the efficiency were evaluated solely on the basis of a research study containing equal groups of continuers and noncontinuers, that is, if it were assumed that $P = .50$.

3. In order for the hits in the entire population which is under consideration to be increased by use of the test, the base rate of the more numerous class (called here positive) must be less than the ratio of the valid negative rate to the sum of valid negative and false negative rates. That is, unless

$$P < \frac{q_2}{q_1 + q_2},$$

the making of decisions on the basis of the test will have an adverse effect. An alternative expression is that $(P / Q) < (q_2 / q_1)$ when $P > Q$, that is, the ratio of the larger to the smaller class must be less than the ratio of the valid negative rate to the false negative rate. When $P < Q$, the conditions for the test to improve upon the base rates are:

$$Q < \frac{p_1}{p_1 + p_2}$$

and

$$\frac{Q}{P} < \frac{p_1}{p_2}.$$

Rotter, Rafferty, and Lotsof (1954) have reported the scores on a sentence completion test for a group of 33 "maladjusted" and 33 "adjusted" girls. They report that the use of a specified cutting score (not cross validated) will result in the correct classification of 85% of the maladjusted girls and the incorrect classification of only 15% of the adjusted girls. It is impossible to evaluate adequately the efficiency of the test unless one knows the base rates of maladjustment (P) and adjustment (Q) for the population of high school girls, although there would be general agreement that $Q > P$. Since $p_1 / (p_1 + p_2) = .85 / (.85 + .15) = .85$, the overall hits in diagnosis with the test will not improve on classification based solely on the base rates unless the proportion of adjusted girls is less than .85. Because the reported effectiveness of the test is spuriously high, the proportion of adjusted girls would no doubt have to be considerably less than .85. Unless there is good reason to believe that the base rates are similar from one setting to another, it is impossible to determine the efficiency of a test such as Rotter's when the criterion is based on ratings unless one replicates his research, including the criterion ratings, with a representative sample of each new population.

4. In altering a sign, improving a scale, or shifting a cutting score, the increment in valid positives per increment in valid positive *rate* is proportional to the positive base rate; and

analogously, the increment in valid negatives per increment in valid negative *rate* is proportional to the negative base rate. That is, if we alter a sign the net improvement in over-all hit rate is

$$H'_T - H_T = \Delta p_1 P + \Delta q_2 Q,$$

where H_T = original proportion of hits (over-all) and H'_T = new proportion of hits (over-all).

5. A corollary of this is that altering a sign or shifting a cut will improve our decision making if, and only if, the ratio of *improvement* Δp_1 in valid positive rate to *worsening* Δp_2 in false negative rate exceeds the ratio of actual negatives to positives in the population.

$$\frac{\Delta p_1}{\Delta p_2} > \frac{Q}{P}.$$

Example: Suppose we improve the intrinsic validity of a certain "schizophrenic index" so that it now detects 20% more schizophrenics than it formerly did, at the expense of only a 5% increase in the false positive rate. This surely looks encouraging. We are, however, working with an outpatient clientele only 1/10th of whom are actually schizophrenic. Then, since

$$\Delta p_1 = .20 \qquad P = .10$$
$$\Delta p_2 = .05 \qquad Q = .90$$

applying the formula we see that

$$\frac{.20}{.05} \not> \frac{.90}{.10}.$$

that is, the required inequality does not hold, and the routine use of this "improved" index will result in an increase in the proportion of erroneous diagnostic decisions.

In the case of any pair of unimodal distributions, this corresponds to the principle that the optimal cut lies at the intersection of the two distribution envelopes (Horst, 1941, pp. 271-272).

Manipulation of Cutting Lines for Different Decisions

For any given psychometric device, no one cutting line is maximally efficient for clinical settings in which the base rates of the criterion groups in the population are different. Furthermore, different cutting lines may be necessary for various decisions within the same population. In this section, methods are presented for manipulating the cutting line of any instrument in order to maximize the efficiency of a device in the making of several kinds of decisions. Reference should be made to the scheme presented in Table 5 for understanding of the discussion which follows. This scheme and the methods for manipulating cutting lines are derived from Duncan, Ohlin, Reiss, and Stanton (1953).

A study in the prediction of juvenile delinquency by Glueck and Glueck (1950) will be used for illustration. Scores on a prediction index for 451 delinquents and 439 non-delinquents (1950, p. 261) are listed in Table 6. If the Gluecks' index is to be used in a population with a given juvenile delinquency rate, cutting lines can be established to maximize the efficiency of the index for several decisions. In the following illustration, a delinquency rate of .20 will be used. From the data in Table 6, optimal cutting lines will be determined for maximizing the proportion of correct predictions, or hits, for all cases (H_T), and for maximizing the proportion of hits (H_P) among those called delinquent (positives) by the index.

TABLE 5

Symbols to Be Used in Evaluating the Efficiency of a Psychometric Device in Classification or Prediction

Diagnosis from Test	Actual Diagnosis		Total Diagnosed from Test
	Positive	Negative	
Positive	NPp_1 (Number of valid positives)	NQp_2 (Number of false positives)	$NPp_1 + NQp_2$ (Number of test positives)
Negative	NPq_1 (Number of false negatives)	NQq_2 (Number of valid negatives)	$NPq_1 + NQq_2$ (Number of test negatives)
Total with actual diagnosis	NP (Number of actual positives)	NQ (Number of actual negatives)	N (Total number of cases)

Note.— For simplicity, the term "diagnosis" is used to denote the classification of any kind of pathology, behavior, or event being studied, or to denote "outcome" if a test is used for prediction. "Number" means *absolute frequency*, not *rate* or probability.

In the first three columns of Table 6, "*f*" denotes the number of delinquents scoring in each class interval, "*cf*" represents the cumulative frequency of delinquents scoring above each class interval (e.g., 265 score above 299), and p_1 represents the proportion of the total group of 451 delinquents scoring above each class interval. Columns 4, 5, and 6 present the same kind of data for the 439 nondelinquents.

TABLE 6

Prediction Index Scores for Juvenile Delinquents and Nondelinquents and Other Statistics for Determining Optimal Cutting Lines for Certain Decisions in a Population with a Delinquency Rate of .20

Prediction Index Score	*Delinquents*			*Nondelinquents*			$1-p_2$	$.2p_1$	$.8p_2$	$.8q_2$	Pp_1+Qq_2	Pp_1+Qp_2	Pp_1/R_P
			cf / 451			*cf* / 439							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	f	*cf*	p_1	*f*	*cf*	p_2	q_2	Pp_1	Qp_2	Qq_2	H_T	R_P	H_P
400+	51	51	.1131	1	1	.0023	.9977	.0226	.0018	.7982	.821	.024	.926
350-399	73	124	.2749	8	9	.0205	.9795	.0550	.0164	.7836	.839	.071	.770
300-349	141	265	.5876	23	32	.0729	.9271	.1175	.0583	.7417	.859	.176	.668
250-299	122	387	.8581	70	102	.2323	.7677	.1716	.1858	.6142	.786	.357	.480
200-249	40	427	.9468	68	170	.3872	.6128	.1894	.3098	.4902	.680	.499	.379
150-199	19	446	.9889	102	272	.6196	.3804	.1978	.4957	.3043	.502	.694	.285
<150	5	451	1.0000	167	439	1.0000	.0000	.2000	.8000	.0000	.200	1.0000	.200

Note.— Frequencies in columns 1 and 4 are from Glueck and Glueck (1950, p. 261)

Maximizing the number of correct predictions or classifications for all cases. The proportion of correct predictions or classifications (H_T) for any given cutting line is given by the formula, $H_T = Pp_1 + Qq_2$. Thus, in column 11 of Table 6, labelled H_T, it can be seen that the best cutting line for this decision would be between 299 and 300, for 85.9% of all predictions would be correct if those above the line were predicted to become delinquent and all those below the line nondelinquent. Any other cutting line would result in a smaller proportion of correct predictions, and, in fact, any cutting line set lower than this point

would make the index inferior to the use of the base rates, for if all cases were predicted to be nondelinquent, the total proportion of hits would be .80.

Maximizing the number of correct predictions or classifications for positives. The primary use of a prediction device may be for *selection* of (a) students who will succeed in a training program, (b) applicants who will succeed in a certain job, (c) patients who will benefit from a certain type of therapy, and the like. In the present illustration, the index would most likely be used for detection of those who are likely to become delinquents. Thus, the aim might be to maximize the number of hits only within the group predicted by the index to become delinquents (predicted positives = $NPp_1 + NQp_2$). The proportion of correct predictions for this group by the use of different cutting lines is given in column 13, labelled H_P. Thus, if a cutting line is set between 399 and 400, one will be correct over 92 times in 100 if predictions are made *only* for persons scoring above the cutting line. The formula for determining the efficiency of the test when only positive predictions are made is $H_P = Pp_1 / (Pp_1 + Qp_2)$.

One has to pay a price for achieving a very high level of accuracy with the index. Since the problem is to select potential delinquents so that some sort of therapy can be attempted, the proportion of this selected group in the total sample may be considered as a selection ratio. The selection ratio for positives is $R_P = Pp_1 + Qp_2$, that is, predictions are made only for those above the cutting line. The selection ratio for each posssible cutting line is shown in column 12 of Table 6, labelled R_P. It can be seen that to obtain maximum accuracy in selection of delinquents (92.6%), predictions can be made for only 2.4% of the population. For other cutting lines, the accuracy of selection and the corresponding selection ratios are given in Table 6. The worker applying the index must use his own judgment in deciding upon the level of accuracy and the selection ratio desired.

Maximizing the number of correct predictions or classifications for negatives. In some selection problems, the goal is the selection of negatives rather than positives. Then, the proportion of hits among all predicted negative for any given cutting line is $H_N = Qq_2 / (Qq_2 + Pq_1)$, and the selection ratio for negatives is $R_N = Pq_1 + Qq_2$.

In all of the above manipulations of cutting lines, it is essential that there be a large number of cases. Otherwise, the percentages about any given cutting line would be so unstable that very dissimilar results would be obtained on new samples. For most studies in clinical psychology, therefore, it would be necessary to establish cutting lines according to the decisions and methods discussed above, and then to cross validate a specific cutting line on new samples.

The amount of shrinkage to be expected in the cross validation of cutting lines cannot be determined until a thorough mathematical and statistical study of the subject is made. It may be found that when criterion distributions are approximately normal and large, cutting lines should be established in terms of the normal probability table rather than on the basis of the observed p and q values found in the samples. In a later section dealing with the selection ratio we shall see that it is sometimes the best procedure to select all individuals falling above a certain cutting line and to select the others needed to reach the selection ratio by choosing at random below the line; or in other cases to establish several different cuts defining *ranges* within which one or the opposite decision should be made.

Decisions based on score intervals rather than cutting lines. The Gluecks' data can be used to illustrate another approach to psychometric classification and prediction when scores for large samples are available with a relatively large number of cases in each score interval. In Table 7 are listed frequencies of delinquents and nondelinquents for prediction index score intervals. The frequencies for delinquents are the same as those in Table 6,

whereas those for nondelinquents have been corrected for a base rate of .20 by multiplying each frequency in column 4 of Table 6[3] by

$$4.11 = \frac{(.80)}{(.20)}\,\frac{(451)}{(439)}.$$

[frequencies corrected from typos in original publication.—LJY]

TABLE 7

Percentage of Delinquents (D) and Nondelinquents (ND) in Each Prediction Index Score Interval in a Population in Which the Delinquency Rate is .20*

Prediction Index Score Interval	No. of D	No. of ND	*Total of D and ND*	% of D in Score Interval	% of ND in Score Interval	% of D and ND in Score Interval
400+	51	4	55	92.7	7.3	100
350-399	73	33	106	68.9	31.1	100
300-349	141	95	236	59.7	40.3	100
250-299	122	288	410	29.8	70.2	100
200-249	40	279	319	12.5	87.5	100
150-199	19	419	438	4.3	95.7	100
<150	5	686	691	.7	99.3	100
Total	451	1804	2255			

*Modification of Table XX-2 from Glueck and Glueck (1950, p. 261).

Table 7 indicates the proportion of delinquents and nondelinquents among all juveniles who fall within a given score interval when the base rate of delinquency is .20. It can be predicted that of those scoring 400 or more, 92.7% will become delinquent; of those scoring between 350 and 399, 68.9% will be delinquent, and so forth. Likewise, of those scoring between 200 and 249, it can be predicted that 87.5% will not become delinquent. Since 80% of predictions will be correct without the index if all cases are called nondelinquent, one would not predict nondelinquency with the index in score intervals over 249. Likewise, it would be best not to predict delinquency for individuals in the intervals under 250 because 20% of predictions will be correct if the base rate is used.

It should be emphasized that there are different ways of quantifying one's clinical errors, and they will, of course, not all give the same evaluation when applied in a given setting. "Percentage of valid positives" (= p_1) is rarely if ever meaningful without the correlated "percentage of false positives" (= p_2), and clinicians are accustomed to the idea that we pay for an increase in the first by an increase in the second, whenever the increase is achieved not by an improvement in the test's intrinsic validity but by a shifting of the cutting score. But the two quantities p_1 and p_2 do not define our over-all hit frequency, which depends also upon the base rates P and Q. The three quantities p_1, p_2, and P do, however, contain all the information needed to evaluate the test with respect to any given sign or cutting score that yields these values. Although p_1, p_2, and P contain the relevant information, other forms of it may be of greater importance. No two of these numbers, for example, answer the obvious question most commonly asked (or vaguely implied) by psychiatrists when an inference is made from a sign, viz., "How sure can you be on the basis of that sign?" The

[3] The Gluecks' Tables XX-2, 3, 4, 5 (1950, pp. 261-262), and their interpretations therefrom are apt to be misleading because of their exclusive consideration of approximately equal base rates of delinquency and nondelinquency. Reiss (1951), in his review of the Gluecks' study, has also discussed their use of an unrepresentative rate of delinquency.

answer to this eminently practical query involves a probability different from any of the above, namely, the *inverse* probability given by Bayes' formula:

$$H_P = \frac{Pp_1}{Pp_1 + Qp_2}.$$

Even a small improvement in the hit frequency to $H'_T = Pp_1 + Qq_2$ over the $H_T = P$ attainable without the test may be adjudged as worthwhile when the increment ΔH_T is multiplied by the N examined in the course of one year and is thus seen to involve a dozen lives or a dozen curable schizophrenics. On the other hand, the simple fact that an actual *shrinkage* in total hit rate may occur seems to be unappreciated or tacitly ignored by a good deal of clinical practice. One must keep constantly in mind that numerous diagnostic, prognostic, and dynamic statements can be made about almost all neurotic patients (e.g., "depressed," "inadequate ability to relate," "sexual difficulties") or about very few patients (e.g., "dangerous," "will act out in therapy," "suicidal," "will blow up into a schizophrenia"). A psychologist who uses a test sign that even cross validates at $p_1 = q_2 = 80\%$ to determine whether "depression" is present or absent, working in a clinical population where practically everyone is fairly depressed except a few psychopaths and old-fashioned hysterics, is kidding himself, the psychiatrist, and whoever foots the bill.

"Successive-Hurdles" Approach

Tests having low efficiency, or having moderate efficiency but applied to populations having very unbalanced base rates ($P \gg Q$) are sometimes defended by adopting a "crude initial screening" frame of reference, and arguing that certain other procedures (whether tests or not) can be applied to the subset identified by the screener ("successive hurdles"). There is no question that in some circumstances (e.g., military induction, or industrial selection with a large labor market) this is a thoroughly defensible position. However, as a general rule one should examine this type of justification critically, with the preceding considerations in mind. Suppose we have a test which distinguishes brain-tumor from non-brain-tumor patients with 75% accuracy and no differential bias ($p_1 = q_2 = .75$). Under such circumstances the test hit rate H_T is .75 regardless of the base rate. If we use the test in making our judgments, we are correct in our diagnoses 75 times in 100. But suppose only one patient in 10 actually has a brain tumor, we will drop our over-all "success" from 90% (attainable by diagnosing "No tumor" in all cases) to 75%. We do, however, identify 3 out of 4 of the real brain tumors, and in such a case it seems worth the price. The "price" has two aspects to it: We take time to give the test, and, having given it, we call many "tumorous" who are not. Thus, suppose that in the course of a year we see 1000 patients. Of these, 900 are non-tumor, and we erroneously call 225 of these "tumor." To pick up (100) (.75) = 75 of the tumors, *all* 100 of whom would have been called tumor-free using the base rates alone, we are willing to mislabel 3 times this many as tumorous who are actually not. Putting it another way, whenever we say "tumor" on the basis of the test, the chances are 3 to 1 that we are mistaken. When we "rule out" tumor by the test, we are correct 96% of the time, an improvement of only 6% in the confidence attachable to a negative finding over the confidence yielded by the base rates.[4]

[4] Improvements are expressed throughout this article as *absolute* increments in percentage of hits, because: (a) This avoids the complete arbitrariness involved in choosing between original hit rate and miss rate as starting denominator; and (b) for the clinician, the person is the most meaningful unit of gain, rather than a proportion *of* a proportion (especially when the reference proportion is very small).

Now, picking up the successive-hurdles argument, suppose a major decision (e.g., exploratory surgery) is allowed to rest upon a second test which is infallible but for practically insuperable reasons of staff, time, and the like, cannot be routinely given. We administer Test 2 only to "positives" on (screening) Test 1. By this tactic we eliminate all 225 false positives left by Test 1, and we verify the 75 valid positives screened in by Test 1. The 25 tumors that slipped through as false negatives on Test 1 are, of course, not picked up by Test 2 either, because it is not applied to them. Our total hit frequency is now 97.5%, since the only cases ultimately misclassified out of our 1000 seen are these 25 tumors which escaped through the initial sieve Test 1. We are still running only 7.5% above the base rate. We have had to give our short-and-easy test to 1000 individuals and our cumbersome, expensive test to 300 individuals, 225 of whom turn out to be free of tumor. But we have located 75 patients with tumor who would not otherwise have been found.

Such examples suggest that, except in "life-or-death" matters, the successive-screenings argument merely tends to soften the blow of Bayes' Rule in cases where the base rates are very far from symmetry. Also, if Test 2 is not assumed to be infallible but only highly effective, say 90% accurate both ways, results start looking unimpressive again. Our net false positive rate rises from zero to 22 cases miscalled "tumor," and we operate 67 of the actual tumors instead of 75. The total hit frequency drops to 94.5%, only 4.5% above that yielded by a blind guessing of the modal class.

The Selection Ratio

Straightforward application of the preceding principles presupposes that the clinical decision maker is free to adopt a policy solely on the basis of maximizing hit frequency. Sometimes there are external constraints such as staff time, administrative policy, or social obligation which further complicate matters. It may then be impossible to make all decisions in accordance with the base rates, and the task given to the test is that of selecting a subset of cases which are decided in the direction opposite to the base rates but will still contain fewer erroneous decisions than would ever be yielded by opposing the base rates without the test. If 80% of patients referred to a Mental Hygiene Clinic are recoverable with intensive psychotherapy, we would do better to treat everybody than to utilize a test yielding 75% correct predictions. But suppose that available staff time is limited so that we *can* treat only half the referrals. The Bayes-type injunction to "follow the base rates when they are better than the test" becomes pragmatically meaningless, for it directs us to make decisions which we cannot implement. The imposition of an *externally* imposed selection ratio, not determined on the basis of any maximizing or minimizing policy but by non-statistical considerations, renders the test worthwhile.

Prior to imposition of any arbitrary selection ratio, the fourfold table for 100 referrals might be as shown in Table 8. If the aim were simply to minimize total errors, we would predict "good" for each case and be right 80 times in 100. Using the test, we would be right only 75 times in 100. But suppose a selection ratio of .5 is externally imposed. We are then forced to predict "poor" for half the cases, even though this "prediction" is, in any given case, likely to be wrong. (More precisely, we handle this subset *as if* we predicted "poor," by refusing to treat.) So we now select our 50 to-be-treated cases from among those 65 who fall in the "test-good" array, having a frequency of $60/65 = 92.3\%$ hits among those selected. This is better than the 80% we could expect (among those selected) by choosing half the total referrals at random. Of course we pay for this, by making many "false negative" decisions; but these are necessitated, whether we use the test or not, by the fact that the selection ratio was determined without regard for hit maximization but by external considerations. Without the test, our false negative rate q_1 is 50% (i.e., 40 of the 80 "good"

cases will be called "poor"); the test reduces the false negative rate to 42.5% (= 34/80), since 15 cases from above the cutting line must be selected at random for inclusion in the not-to-be-treated group below the cutting line [i.e., 20 + (60/65)15 = 34]. Stated in terms of correct decisions, without the test 40 out of 50 selected for therapy will have a good therapeutic outcome; with the test, 46 in 50 will be successes.

TABLE 8
Actual and Test-Predicted Therapeutic Outcome

Test Prediction	Therapeutic Outcome		
	Good	Poor	Total
Good	60	5	65
Poor	20	15	35
Total	80	20	100

Reports of studies in which formulas are developed from psychometrics for the prediction of patients' continuance in psychotherapy have neglected to consider the relationship of the selection ratio to the specific population to which the prediction formula is to be applied. In each study the population has consisted of individuals who were *accepted for therapy* by the usual methods employed at an outpatient clinic, and the prediction formula has been evaluated *only* for such patients. It is implied by these studies that the formula would have the same efficiency if it were used for the *selection* of "continuers" from all those *applying* for therapy. Unless the formula is tested on a random sample of applicants who are allowed to enter therapy without regard to their test scores, its efficiency for selection purposes is unknown. The reported efficiency of the prediction formula in the above studies pertains only to its use in a population of patients who have already been selected for therapy. There is little likelihood that the formula can be used in any practical way for further selection of patients unless the clinic's therapists are carrying a far greater load than they plan to carry in the future.

The use of the term "selection" (as contrasted with "prediction" or "placement") ought not to blind us to the important differences between industrial selection and its clinical analogue. The incidence of false negatives—of potential employees screened out by the test who would actually have made good on the job if hired—is of little concern to management except as it costs money to give tests. Hence the industrial psychologist may choose to express his aim in terms of minimizing the false positives, that is, of seeing to it that the job success *among those hired* is as large a rate as possible. When we make a clinical decision to treat or not to treat, we are withholding something from people who have a claim upon us in a sense that is much stronger than the "right to work" gives a job applicant any claim upon a particular company. So, even though we speak of a "selection ratio" in clinical work, it must be remembered that those cases *not selected* are patients about whom a certain kind of important negative decision is being made.

For any *given* selection ratio, maximizing total hits is always equivalent to maximizing the hit rate for either type of decision (or minimizing the errors of either, or both, kinds), since cases shifted from one cell of the table have to be exactly compensated for. If *m* "good" cases that were correctly classified by one decision method are incorrectly classified by another, maintenance of the selection ratio entails that *m* cases correctly called "poor" are also miscalled "good" by the new method. Hence an externally imposed selec-

tion ratio eliminates the often troublesome value questions about the relative seriousness of the two kinds of errors, since they are unavoidably increased or decreased at exactly the same rate.

If the test yields a score or a continuously varying index of some kind, the values of p_1 and p_2 are not fixed, as they may be with "patterns" or "signs." Changes in the selection ratio, R, will then suggest shifting the cutting scores or regions on the basis of the relations obtaining among R, P, and the p_1, p_2 combinations yielded by various cuts. It is worth special comment that, in the case of continuous distributions, the optimum procedure is *not* always to move the cut until the total area truncated = NR, selecting all above that cut and rejecting all those below. Whether this "obvious" rule is wise or not depends upon the distribution characteristics. We have found it easy to construct pairs of distributions such that the test is "discriminating" throughout, in the sense that the associated cumulative frequencies q_1 and q_2 maintain the same direction of their inequality everywhere in the range, that is,

$$\frac{1}{N_2}\int_{-\infty}^{x_i} f_2(x)\,dx > \frac{1}{N_1}\int_{-\infty}^{x_i} f_1(x)\,dx \quad \text{for all } x_i;$$

yet in which the hit frequency given by a single cut at R is inferior to that given by first selecting with a cut which yields $N_c < NR$, and then picking up the remaining $(NR - N_c)$ cases at random below the cut. Other more complex situations may arise in which different types of decisions should be made in different regions, actually reversing the policy as we move along the test continuum. Such numerical examples as we have constructed utilize continuous, unimodal distributions, and involve differences in variability, skewness, and kurtosis not greater than those which arise fairly often in clinical practice. Of course the utilization of any very complicated pattern of regions requires more stable distribution frequencies than are obtainable from the sample sizes ordinarily available to clinicians.

It is instructive to contemplate some of the moral and administrative issues involved in the practical application of the preceding ideas. It is our impression that a good deal of clinical research is of the "So what?" variety, not because of defects in experimental design such as inadequate cross validation but because it is hard to see just what are the useful changes in decision making which could reasonably be expected to follow. Suppose, for example, it is shown that "duration of psychotherapy" is 70% predictable from a certain test. Are we prepared to propose that those patients whose test scores fall in a certain range should not receive treatment? If not, then is it of any real advantage therapeutically to "keep in mind" that the patient has 7 out of 10 chances of staying longer than 15 hours, and 3 out of 10 chances of staying less than that? We are not trying to poke fun at research, since presumably almost any lawful relationship stands a chance of being valuable to our total scientific comprehension some day. But many clinical papers are ostensibly inspired by practical aims, and can be given theoretical interpretation or fitted into any larger framework only with great difficulty if at all. It seems appropriate to urge that such "practical"-oriented investigations should be really *practical*, enabling us to see how our clinical decisions could rationally be modified in the light of the findings. It is doubtful how much of current work could be justified in these terms.

Regardless of whether the test validity is capable of improving on the base rates, there are some prediction problems which have practical import only because of limitations in personnel. What other justification is there for the great emphasis in clinical research on "prognosis," "treatability," or "stayability"? The very formulation of the predictive task as "maximizing the number of hits" already presupposes that we intend *not* to treat some cases; since if we treat all comers, the ascertainment of a bad prognosis score has no practical effect other than to discourage the therapist (and thus hinder therapy?). If

intensive psychotherapy could be offered to all veterans who are willing to accept referral to a VA Mental Hygiene Clinic, would it be licit to refuse those who had the poorest outlook? Presumably not. It is interesting to contrast the emphasis on prognosis in clinical psychology with that in, say, cancer surgery, where the treatment *of choice* may still have a very low probability of "success," but is nevertheless carried out on the basis of that low probability. Nor does this attitude seem unreasonable, since no patient would refuse the best available treatment on the ground that even it was only 10% effective. Suppose a therapist, in the course of earning his living, spends 200 hours a year on nonimprovers by following a decision policy that also results in his unexpected success with one 30-year-old "poor bet." If this client thereby gains $16 \times 365 \times 40 = 233{,}600$ hours averaging 50% less anxiety during the rest of his natural life, it was presumably worth the price.

These considerations suggest that, with the expansion of professional facilities in the behavior field, the prediction problem will be less like that of industrial *selection* and more like that of *placement*. "To treat or not to treat" or "How treatable" or "How long to treat" would be replaced by "What *kind* of treatment?" But as soon as the problem is formulated in this way, the external selection ratio is usually no longer imposed. Only if we are deciding between such alternatives as classical analysis and, say, 50-hour interpretative therapy would such personnel limitations as can be expected in future years impose an arbitrary R. But if the decision is between such alternatives as short-term interpretative therapy, Rogerian therapy, Thorne's directive therapy, hypnotic retraining, and the method of tasks (Herzberg, 1945; Salter, 1949; Wolpe, 1952), we could "follow the base rates" by treating every patient with the method known to have the highest success frequency among patients "similar" to him. The criteria of similarity (class membership) will presumably be multiple, both phenotypic and genotypic, and will have been chosen because of their empirically demonstrated prognostic relevance rather than by guesswork, as is current practice. Such an idealized situation also presupposes that the selection and training of psychotherapists will have become socially realistic so that therapeutic personnel skilled in the various methods will be available in some reasonable proportion to the incidence with which each method is the treatment of choice.

How close are we to the upper limit of the predictive validity of personality tests, such as was reached remarkably early in the development of academic aptitude tests? If the now-familiar 2/3 to 3/4 proportions of hits against even-split criterion dichotomies are already approaching that upper limit, we may well discover that for many decision problems the search for tests that will significantly better the base rates is a rather unrewarding enterprise. When the criterion is a more circumscribed trait or symptom ("depressed," "affiliative," "sadistic," and the like), the difficulty of improving upon the base rates is combined with the doubtfulness about how valuable it is to have such information with 75% confidence anyhow. But this involves larger issues beyond the scope of the present paper.

Availability of Information on Base Rates

The obvious difficulty we face in practical utilization of the preceding formulas arises from the fact that actual quantitative knowledge of the base rates is usually lacking. But this difficulty must not lead to a dismissal of our considerations as clinically irrelevant. In the case of many clinical decisions, chiefly those involving such phenotypic criteria as overt symptoms, formal diagnosis, subsequent hospitalization, persistence in therapy, vocational or marital adjustment, and the numerous "surface" personality traits which clinicians try to assess, *the chief reason for our ignorance of the base rates is nothing more subtle than our failure to compute them*. The file data available in most installations having a fairly stable source of clientele would yield values sufficiently accurate to permit minimum and max-

imum estimates which might be sufficient to decide for or against use of a proposed sign. It is our opinion that this rather mundane taxonomic task is of much greater importance than has been realized, and we hope that the present paper will impel workers to more systematic efforts along these lines.

Even in the case of more subtle, complex, and genotypic inferences, the situation is far from hopeless. Take the case of some such dynamic attribution as "strong latent dependency, which will be anxiety-arousing as therapy proceeds." If this is so difficult to discern *even during intensive therapy* that a therapist's rating on it has too little reliability for use as a criterion, it is hard to see just what is the value of guessing it from psychometrics. If a skilled therapist cannot discriminate the personality characteristic after considerable contact with the patient, it is at least debatable whether the characteristic makes any practical difference. On the other hand, if it can be reliably judged by therapists, the determination of approximate base rates again involves nothing more complex than systematic recording of these judgments and subsequent tabulation. Finally, "clinical experience" and "common sense" must be invoked when there is nothing better to be had. Surely if the q_1/q_2 ratio for a test sign claiming validity for "difficulty in accepting inner drives" shows from the formula that the base rate must not exceed .65 to justify use of the sign, we can be fairly confident in discarding it for use with *any* psychiatric population! Such a "backward" use of the formula to obtain a maximum useful value of P, in conjunction with the most tolerant common-sense estimates of P from daily experience, will often suffice to answer the question. If one is really in complete ignorance of the limits within which P lies, then obviously no rational judgment as to the probable efficiency of the sign can be made.

Estimation versus Significance

A further implication of the foregoing thinking is that the exactness of certain small sample statistics, or the relative freedom of certain nonparametric methods from distribution assumptions, has to be stated with care lest it mislead clinicians into an unjustified confidence. When an investigator concludes that a sign, item, cutting score, or pattern has "validity" on the basis of small sample methods, he has rendered a certain very broad null hypothesis unplausible. To decide, however, whether this "validity" warrants clinicians in using the test is (as every statistician would insist) a further and more complex question. To answer this question, we require more than knowledge that $p_1 \neq p_2$. We need in addition to know, with respect to each decision for which the sign is being proposed, whether the appropriate inequality involving p_1, p_2, and P is fulfilled. More than this, since we will usually be extrapolating to a somewhat different clinical population, we need to know whether altered base rates P' and Q' will falsify these inequalities. To do this demands *estimates* of the test parameters p_1 and p_2, the setting up of confidence belts for their difference $p_1 - p_2$ rather than the mere proof of their nonidentity. Finally, if the sign is a cutting score, we will want to consider shifting it so as to *maintain* optimal hit frequency with new base rates. The effect upon p_1 and p_2 of a contemplated movement of a critical score or band requires a knowledge of distribution form such as only a large sample can give.

As is true in all practical applications of statistical inference, nonmathematical considerations enter into the use of the numerical patterns that exist among P, p_1, p_2, and R. But "pragmatic" judgments initially require a separation of the several probabilities involved, some of which may be much more important than others in terms of the human values associated with them. In some settings, over-all hit rate is all that we care about. In others, a redistribution of the hits and misses even without much total improvement may concern us. In still others, the proportions p_1 and q_2 are of primary interest; and, finally, in some

instances the confrontation of a certain increment in the absolute frequency (NPp_1) of one group identified will outweigh all other considerations.

Lest our conclusions seem unduly pessimistic, what constructive suggestions can we offer? We have already mentioned the following: (a) Searching for subpopulations with different base rates; (b) successive-hurdles testing; (c) the fact that even a very small *percentage* of improvement may be worth achieving in certain crucial decisions; (d) the need for systematic collection of base-rate data so that our several equations can be applied. To these we may add two further "constructive" comments. First, test research attention should be largely concentrated upon behaviors having base rates nearer a 50-50 split, since it is for these that it is easiest to improve on a base-rate decision policy by use of a test having moderate validity. There are, after all, a large number of clinically important traits which do not occur "almost always" or "very rarely." Test research might be slanted more toward them; the current popularity of *Q*-sort approaches should facilitate the growth of such an emphasis, by directing attention to items having a reasonable "spread" in the clinical population. Exceptions to such a research policy will arise, in those rare domains where the pragmatic consequences of the alternative decisions justify focusing attention almost wholly on maximizing Pp_1, with relative neglect of Qp_2. Secondly, we think the injunction "quit wasting time on noncontributory psychometrics" is really constructive. When the clinical psychologist sees the near futility of predicting rare or near-universal events and traits from test validities incapable of improving upon the base rates, his clinical time is freed for more economically defensible activities, such as research which will improve the parameters p_1 and p_2; and for *treating* patients rather than uttering low-confidence prophecies or truisms about them (in this connection see Meehl, 1954/1996, pp. vii [1996, p. xv], 7, 127-128). It has not been our intention to be dogmatic about "what is worth finding out, how often." We do suggest that the clinical use of patterns, cutting scores, and signs, or research efforts devoted to the discovery of such, should always be evaluated in the light of the simple algebraic fact discovered in 1763 by Mr. Bayes.

Summary

1. The practical value of a psychometric sign, pattern, or cutting score depends jointly upon its intrinsic validity (in the usual sense of its discriminating power) and the distribution of the criterion variable (base rates) in the clinical population. Almost all contemporary research reporting neglects the base-rate factor and hence makes evaluation of test usefulness difficult or impossible.

2. In some circumstances, notably when the base rates of the criterion classification deviate greatly from a 50 percent split, use of a test sign having slight or moderate validity will result in an *increase* of erroneous clinical decisions.

3. Even if the test's parameters are precisely known, so that ordinary cross-validation shrinkage is not a problem, application of a sign within a population having these same test parameters but a different base rate may result in a marked change in the proportion of correct decisions. For this reason validation studies should present trustworthy information respecting the criterion distribution in addition to such test parameters as false positive and false negative rates.

4. Establishment of "validity" by exact small sample statistics, since it does not yield accurate information about the test parameters (a problem of estimation rather than significance), does not permit trustworthy judgments as to test usefulness in a new population with different or unknown base rates.

5. Formulas are presented for determining limits upon relations among (a) the base rates, (b) false negative rate, and (c) false positive rate which must obtain if use of the test sign is to improve clinical decision making.

6. If, however, external constraints (e.g., available staff time) render it administratively unfeasible to decide all cases in accordance with the base rates, a test sign may be worth applying even if following the base rates *would* maximize the total correct decisions, were such a policy possible.

7. Trustworthy information as to the base rates of various patient characteristics can readily be obtained by file research, and test development should (other things being equal) be concentrated on those characteristics having base rates nearer .50 rather than close to .00 or 1.00.

8. The basic rationale is that of Bayes' Theorem concerning the calculation of so-called "inverse probability."

REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurements Used in Education, Joint Committee (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51, 201-238.

Anastasi, A., & Foley, J. P. (1949). *Differential psychology* (Rev. Ed.). New York: Macmillan.

Bross, I. D. J. (1953). *Design for decision*. New York: Macmillan.

Danielson, J. R., & Clark, J. H. (1954). A personality inventory for induction screening. *Journal of Clinical Psychology*, 10, 137-143.

Dorken, H., & Kral, A. (1952). The psychological differentiation of organic brain lesions and their localization by means of the Rorschach test. *American Journal of Psychiatry*, 108, 764-770.

Duncan, O. D., Ohlin, L. E., Reiss, A. J., & Stanton, H. R. (1953). Formal devices for making selection decisions. *American Journal of Sociology*, 58, 573-584.

Glueck, S., & Glueck, E. (1950). *Unraveling juvenile delinquency*. Cambridge, MA: Harvard University Press.

Goodman, L. A. (1953). The use and validity of a prediction instrument. I. A reformulation of the use of a prediction instrument. *American Journal of Sociology*, 58, 503-509.

Hanvik, L. J. (1949). Some psychological dimensions of low back pain. Unpublished doctor's thesis, University of Minnesota.

Herzberg, A. (1945). *Active psychotherapy*. New York: Grune & Stratton.

Horst, P. (Ed.) (1941). The prediction of personality adjustment. *Social Science Research Coun.[Counseling?] Bulletin*, No. 48, 1-156.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press. Reprinted with new Preface, 1996, by Jason Aronson, Northvale, NJ. Reprinted 2013 by Echo Point Books

Reiss, A. J. (1951). Unraveling juvenile delinquency. II. An appraisal of the research methods. *American Journal of Sociology*, 57, 115-120.

Rosen, A. (1954). Detection of suicidal patients: An example of some limitations in the prediction of infrequent events. *Journal of Consulting Psychology*, 18, 397-403.

Rotter, J. B., Rafferty, J. E., & Lotsof, A. B. (1954). *Journal of Consulting Psychology*, 18, 105-111.

Salter, A. (1950). *Conditioned reflex therapy*. New York: Creative Age Press.

Taulbee, E. S., & Sisson, B. D. (1954). Rorschach pattern analysis in schizophrenia: A cross-validation study. *Journal of Clinical Psychology*, 10, 80-82.

Thiesen, J. W. (1952). A pattern analysis of structural characteristics of the Rorschach test in schizophrenia. *Journal of Consulting Psychology*, 16, 365-370.

Wolpe, J. Objective psychotherapy of the neuroses. *South African Medical Journal*, 26, 825-839.

pdf by LJY, July 2015