DOCUMENT RESUME

ED 053 201 TM 000 733

AUTHOR	Perry, Dallis
TITLE	Interpreting Standardized Test Scores.
INSTITUTION	Minnesota Univ., St. Paul. Student Counseling Bureau.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
REPORT NO	TR-8000
PUB DATE	71
NOTE	57p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Criterion Referenced Tests, Evaluation, Expectancy Tables, Grade Equivalent Scores, Measurement, Norm Referenced Tests, *Prediction, Profile Evaluation, Raw Scores, *Scores, Scoring, Standard Error of Measurement, *Standardized Tests, Testing, *Test Interpretation, Test Reliability, *Test Validity, True Scores

ABSTRACT

Principles of test administration, test validity, and accuracy of measurement underlying interpretation of standardized test scores in educational administration, instruction, and guidance are presented. Types of norm-referenced score transformations, including percentiles, standard scores, and grade equivalents, and of criterion referenced scores, including content scales, predicted scores, and expectancies, are described; and their applications are illustrated. Special attention is given to multi-scores tests and the representation of their scores as profiles and similarity indexes. (Author)

ED053201

THE CENTER FOR THE STUDY OF ORGANIZATIONAL PERFORMANCE AND HUMAN EFFECTIVENESS

University of Minnesota
Minneapolis, Minnesota

Office of Naval Research Contract
ONR N00014-68-A-0141-0003

TM 000 733

INTERPRETING STANDARDIZED TEST SCORES

Dallis Perry

Technical Report No. 8000

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

Student Counseling Bureau
University of Minnesota

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

Interpreting Standardized Test Scores

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Dallis K. Perry

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
1971	47	7

8a. CONTRACT OR GRANT NO.

N00014-68-A-014?-0003

b. PROJECT NO.

NR 151-323

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

8000

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

Approved for public release; distribution is unlimited

11. SUPPLEMENTARY NOTES

A revised version of this paper will appear as a chapter in Minnesota Test Norms and Expectancy Tables, published by the Minn. Dept. of Education

12. SPONSORING MILITARY ACTIVITY

Personnel & Training Research Programs
Office of Naval Research (code 458)
Department of the Navy
~~Arlington, Virginia 22217~~

13. ABSTRACT

Principles of test administration, test validity, and accuracy of measurement underlying interpretation of standardized test scores in educational administration, instruction, and guidance are presented. Types of norm-referenced score transformations, including percentiles, standard scores, and grade equivalents, and of criterion referenced scores, including content scales, predicted scores, and expectancies, are described; and their applications are illustrated. Special attention is given to multi-scores tests and the representation of their scores as profiles and similarity indexes.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Security Classification

A-31408

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Test Measurement Evaluation Guidance Score Norm Profile Expectancy Centour Similarity						

INTERPRETING STANDARDIZED TEST SCORES

by

Dallis K. Perry

Student Counseling Bureau

University of Minnesota

INTERPRETING STANDARDIZED TEST SCORES

Dallis Perry

Student Counseling Bureau, University of Minnesota

Uses of Tests

Standardized tests are used to assist in making a wide variety of educational decisions:

Which students should be selected for Special Program A?

What educational and vocational plans are reasonable for Student B?

For what level of instruction in mathematics is Student C ready?

Has Class D made the expected progress in science?

How successful is the new social studies curriculum in School E?

Test scores provide just one of many kinds of information that must be evaluated and integrated to answer these questions. The ways in which such information is used in educational administration, instruction, and guidance is the subject of such disciplines as educational evaluation, teaching methodology, and counseling and is beyond the scope of this discussion; but, before we use test results in any way, we must understand what information is contained in the test scores--what it is they do and do not tell us.

Cronbach's (1970) definition of a test as "a systematic procedure for observing a person's behavior and describing it with the aid of a numerical scale or category system" is perhaps as satisfactory as any. The tests with which we are concerned are standardized tests--standardized both with respect to the presentation of the stimuli (items) that elicit the behavior that is observed and with respect to the reference data by which the numerical results are interpreted. The score that results directly from a test operation is ordinarily an arbitrary and quite meaningless figure. A considerable portion of test technology is concerned with procedures for transforming such "raw" scores to scales that "build in" significance through their relation to one or more kinds of reference information. Two general classes of transformed scores are "norm-referenced" scores, which indicate relative standing in comparison with a specified reference group, and "criterion-referenced" scores, which relate test performance to the kind of behavior exhibited by or expected from, the examinee. Underlying the interpretation of both kinds of scores are the concepts of validity and accuracy of measurement and the assumption that the tests have been presented to students in a standard manner. The following sections discuss test administration circumstances and the concepts of measurement accuracy, validity, norm-reference, and criterion-reference as they influence the interpretation of standardized test scores.

Test Administration

Fundamental to a standardized test is the equivalence of test content from one student to another, which makes possible comparison of scores. It is essential that this standardization not be compromised by special instructions, assistance, or failures in test security that may effectively alter the content in unknown ways for some students. Testing conditions cannot, of course, be identical for all examinees, but they should be comparable in every way possible. Because most educational tests are regarded as measures of maximum performance, each student must have an opportunity to do his best. Satisfactory physical conditions of lighting, heating, ventilation, space, and work surfaces are assumed, as well as rigid adherence to specified directions and time limits. Equally important, and much more difficult to control, are the internal conditions that each student brings to the test. If a test score is to represent maximum performance, the effort and therefore the motivation to do well on the test must be comparable to that expected in the situations to which the test score is related. Test manuals and administration directions give little attention to pretest preparation and instruction of examinees. A clear explanation of the purpose and significance of the tests, without resorting to exhortation, is preferable to presentation of the tests as a required but unexplained task. Motivation cannot be completely standardized, of course, and the counselor or teacher with specific knowledge of each student as well as of the testing situation can best judge whether a given test score should be accepted at face value, regarded with extra caution, or disregarded completely because of the circumstances in which it was obtained.

Accuracy of Measurement

No single test score is completely representative of the "universe" of behavior for a person. A test score is based on a sample of behavior, and scores based on different samples can be expected to vary. Interpretation of the score must take into account the amount of such variation to be expected under given circumstances. This variation is usually expressed as "error of measurement", considered to be the difference between the observed score and a hypothetical "true" score consisting of the mean of a very large number of measurements of the same kind on the same person.

Standard Error of Measurement

The standard deviation of the distribution of measurements on a person, of which the person's true score is the mean (or equivalently the standard deviation of the differences between true and observed scores), is called the standard error of measurement (s.e.m.) for that person. Although the s.e.m.

on a test varies from one person to another, in practice the average s.e.m. over a sample of persons is determined as an estimate of the s.e.m. on the test for each person.

The s.e.m. of a test indicates the extent to which a person's scores obtained by repeated measurement of the same kind would vary around the person's true score. It may be pictured as shown in Fig. 1. Within the range of $\pm$ 1 s.e.m. from a person's true score will fall 68% of his obtained scores,

---

Figure 1 about here

---

and 95% will fall within $\pm$ 2 s.e.m.. If the s.e.m. is 3, for example, the probability is 68% that any observed score is within 3 points of the true score.

The s.e.m. of a test is important because it emphasizes that an observed test score is just an estimate and not a precisely determined number, and at the same time it quantifies the dependability of the score. Test scores are sometimes reported as ranges or bands, typically extending 1 s.e.m. above and below the observed score, with or without the observed score indicated. Although the interpretation of such ranges is difficult to specify precisely in probability terms, they have the advantage of emphasizing to users the limits of score dependability.

In evaluating scores on a test with reference to its s.e.m., two points should be considered:

1. The reported s.e.m. is an estimate of the <u>average</u> s.e.m. for all persons who take the test. Individuals differ in their variability as well as in their true scores, so the actual s.e.m. is not the same for all persons. The s.e.m. of a well-constructed test should not be correlated with test scores, but in practice persons near the extremes of a distribution are less likely to

be measured accurately than those near the middle. This situation may arise, for example, if the test has insufficient "ceiling" so that differences among the more able students cannot be detected, or if it is so difficult for some students that they respond randomly or by excessive guessing.

2. Different procedures used to estimate the s.e.m. of a test ascribe different sources of observed score variance to error. It is important to keep in mind the sources of variance represented in the s.e.m., and, therefore, the generalizability of the score. Internal consistency procedures (Kuder-Richardson formulas, split-half, odd-even) or alternate form correlations generally include as error that variance due to sampling of test content and that due to momentary factors that differentially influence performance during a single testing occasion. Factors that would differentially affect scores on another occasion are ascribed to "true" score variance. Retesting at a different time with the same instrument leads to the inclusion of differences due to testing occasions, but not differences due to content sampling, in the error variance estimate.

Reliability Coefficients

As Fig. 1 implies, the error variance ordinarily is much smaller than the total variance on a test. If it were not--if all the variance were error variance--there would be no true score variance and the test would have no value. Interpretation of the s.e.m. of a test depends in part on how much smaller than total score variance it is. An s.e.m. of 3 has quite different significance for a test with a standard deviation of 4 than for a test with a standard deviation of 40. The variance of a group of test scores is composed of the error variance plus the true score variance, or

$$S^2_{\text{observed}} = S^2_{\text{true}} + S^2_{\text{error}} \qquad (1)$$

The relationship of these variances is usually expressed as the ratio of true score variance to total score variance, called the reliability coefficient,

$$r = S_T^2 / S_0^2 \qquad (2)$$

Because true score variance, and therefore total observed variance, is a function of the heterogeneity of the group being measured, a reliability coefficient reflects both group and test characteristics, whereas the error component of scores on a test, (s.e.m. squared) is regarded as a characteristic of the test, fixed for all groups. In interpretation of an individual test score the s.e.m. most directly indicates the confidence that can be placed in the score, but the stability of the score with respect to an entire group of scores, as indicated by the reliability coefficient, also should be known. Given the standard deviation of the group in question one can, from (1) and (2) above, compute either s.e.m. or r from the other according to the familiar formulas

$$\text{s.e.m.} = S_o\sqrt{1-r} \qquad (3)$$

$$r = 1 - \frac{(\text{s.e.m.})^2}{S_o^{\ 2}} \qquad (4)$$

Internal consistency reliability of the Minnesota Scholastic Aptitude Test (MSAT) was found to be .93 (Layton, no date), which indicates, according to formulas (3) and (4), a s.e.m. about one-fourth as large ($\sqrt{.07}$ = .26) as the standard deviation of 13.8, or about 3.7. Referring to the MSAT norms in Table 1 we find that, if, for example, a student's "true" score is at the 71st percentile (RS=44), about two-thirds of the time in repeated testing his observed MSAT score would be between the 63rd and 70th percentiles. He would obtain a score below the 54th percentile less than 3% of the time.

The relationship of these variances is usually expressed as the ratio of true score variance to total score variance, called the reliability coefficient,

$$r = S_T^2 / S_O^2 \qquad (2)$$

Because true score variance, and therefore total observed variance, is a function of the heterogeneity of the group being measured, a reliability coefficient reflects both group and test characteristics, whereas the error component of scores on a test, (s.e.m. squared) is regarded as a characteristic of the test, fixed for all groups. In interpretation of an individual test score the s.e.m. most directly indicates the confidence that can be placed in the score, but the stability of the score with respect to an entire group of scores, as indicated by the reliability coefficient, also should be known. Given the standard deviation of the group in question one can, from (1) and (2) above, compute either s.e.m. or r from the other according to the familiar formulas

$$\text{s.e.m.} = S_O\sqrt{1-r} \qquad (3)$$

$$r = 1 - \frac{(\text{s.e.m.})^2}{S_O^{\,2}} \qquad (4)$$

Internal consistency reliability of the Minnesota Scholastic Aptitude Test (MSAT) was found to be .93 (Layton, no date), which indicates, according to formulas (3) and (4), a s.e.m. about one-fourth as large ($\sqrt{.07}$ = .26) as the standard deviation of 13.8, or about 3.7. Referring to the MSAT norms in Table 1 we find that, if, for example, a student's "true" score is at the 71st percentile (RS=44), about two-thirds of the time in repeated testing his observed MSAT score would be between the 63rd and 70th percentiles. He would obtain a score below the 54th percentile less than 3% of the time.

Validity

The most critical information underlying the interpretation of test scores is how well the scores measure the characteristic the test is being used to measure, i.e., how valid is the test for the purpose to which it is being put. Because a test may be used for many purposes, it may have many validities and even several different kinds of validity. Different kinds of validity are generally classified into three categories: content validity, criterion-related validity, and construct validity.

Content Validity

When a test is used to determine a person's current knowledge or performance in a domain represented by the test, evidence of how well the test actually represents the domain is required to establish the content validity of the test. Such evidence usually takes the form of an analysis of the domain into subdivisions, description of the subdivisions, and identification of the items related to each subdivision. In educational achievement tests such subdivisions usually correspond to educational objectives. It is important that both subject matter content and process be included in the analysis and description of the test.

Establishment of a test's content validity requires demonstration not only of what the test does measure but also of what it does not measure. Extraneous factors that are measured by a test but are not conceptually a part of its content lower its content validity. Two of the most common such influences are reading skill and working speed, because so many achievement tests are composed of written items and are given with time limits.

The careful analysis and description of the measurement domain which characterize the establishment of content validity distinguish it from "face validity", which refers to the superficial appearance, or even name, of a test.

Motivation may be better if test items appear to examinees to be relevant to the purposes of testing; therefore, face validity may be desirable, but it is not the same as content validity.

Criterion-related Validity

When a test is used to predict a specific kind of performance other than that measured by the test itself, evidence is required that the test scores are indeed related to the other, criterion, performance. Such evidence is most commonly presented in the form of a coefficient of correlation between test and criterion scores.

Clearly, a test has as many validities as criteria. Thus the median correlation of MSAT scores with grades of freshmen in Minnesota colleges is .43, which demonstrates its validity as a measure of scholastic aptitude; but the coefficients in individual colleges vary from .10 to .76.

Adequate evidence of criterion-related validity requires not only a validity coefficient of sufficient size to be useful but also a criterion measure that truly represents the behavior or performance to be predicted. School marks or grades are the most commonly used educational criteria, and tests validated against such measures must be used with awareness of the limited scope of relevant behavior represented in the criterion. Nevertheless, because grades do represent a significant aspect of achievement and one that may be critical in determining continuation and completion of an educational program, correlation of test scores with grades is an important and meaningful indication of validity.

Construct Validity

Criterion-related validity is invaluable for use of a test to aid in reaching a decision, e.g., choice of college, regarding a specific course of action, the outcome of which can be measured in some way, e.g., by subsequent course grades. However, we cannot expect that tests will have been specifically validated against criteria for all decisions of all students who may be aided by a better understanding of their capabilities and characteristics as measured by tests. For effective counseling use of tests to help understand students and to help students understand themselves we must know "what the test measures", apart from its prediction of behavior in specific situations. Evidence of the meaning of test scores in terms of the psychological characteristics, or constructs, represented by the scores is termed "construct validity". Such evidence may take the form of analysis of the content of the test, synthesis of criterion-related validity coefficients, correlations with other tests, factor analysis, differences or similarities of scores of specified groups (e.g., age or educational levels), item analysis, observation of test-taking behavior, and influence of training or experience on scores. As with evidence of content validity, demonstration of what the test does not measure is as important as demonstration of what it does measure.

Interpretation of the Differential Aptitude Tests (DAT) for counseling secondary school students, for example, depends largely on construct validity. Although the DAT manual reports more than 5,000 predictive validity coefficients, few counselors will have such evidence available for their students and for criteria specifically relevant for their students. Focusing on the Mechanical Reasoning (MR) test we find by examining the items that they deal with gears, levers, pulleys, the application of forces, and similar principles that are part of the content of physical mechanics. The items are presented pictorially, with verbal questions about the pictures, so the test requires some reading ability; but the questions and the words in them are short and should be easily understood. Correlations of about .5 to .6 with the Verbal Reasoning test and with various intelligence tests indicate that MR is measuring something different than verbal ability, and item analyses of the very similar Mechanical Reasoning Test indicate that it is measuring a general mechanical ability, not separate "levers ability", "gears ability", etc. (Cronbach, 1970). On the average MR correlates higher with high school grades in science than in other subjects (although it is not the best DAT predictor of science grades), and it was found to be an effective predictor of vocational school performance of machine shop students but not of auto mechanics students. Girls' scores on the test tend to be substantially lower and less reliable than boys' scores and to have higher correlations with grades in "unrelated" high school courses such as English and social science, suggesting that the test functions somewhat differently for the two sexes. Because MR is a revision of earlier Mechanical Comprehension Tests, evidence that scores on the latter are related to evaluations of training and job performance of various jobs concerned with machinery supports the construct validity of MR. Finally, MR scores are correlated about .4 with mechanical and scientific interests of boys as measured by the Kuder Preference Record and negligibly with other interests. Again, the relationships for girls are lower. Taken together the evidence briefly summarized above supports the notion that MR measures a meaningful characteristic of students, one that is appropriately labeled "mechanical reasoning", is not the same as general intelligence, and is important in certain scientific and mechanical pursuits though not in every activity labeled "mechanical".

Establishment of construct validity in a different domain is illustrated by the development of the Academic Achievement (AACH) scale for the Strong Vocational Interest Blank (Campbell and Johansson, 1966). This scale was developed by selecting SVIB items that significantly differentiated between students with high grades in college and high school and those with low grades. The scale correlated about .35 with high school and college grade averages in a cross-validation sample drawn from the same population as that on which the scale was constructed and also in a 25-year-old sample of college freshmen tested in the 1930's. Low correlations with MSAT scores show that the scale is not just another measure of scholastic aptitude, and the AACH score adds slightly to the multiple correlation of HSR and MSAT with college GPA. In 10-year and 25-year follow-up groups the scale showed substantial differences between students who dropped out of college, and, in order, those who obtained BA, MA, and Ph D degrees. Scores were found to increase until about age 28 and then remain relatively stable. Examination of the item content indicates that items scored positively represent scientific, aesthetic, and intellectual activities, whereas those scored negatively involve sales, business, and manual skills. AACH scores of occupational groups are ranked very much like the average educational levels of the groups, with scientists (biologists, mathematicians, psychiatrists, physicists) at the top and policemen, forest service men, pilots, and office workers at the bottom. Scores of outstanding persons in 10 occupations showed similar differences, with outstanding composers, novelists, astronauts, and psychologists scoring high and outstanding life insurance salesmen, military men, and football coaches scoring low. In summary the AACH scale appears to measure interest in activities that lead to getting good grades and continuing in school, but it is not a measure of scholastic aptitude as such nor a predictor of success within occupations.

Norm-Referenced Scores

A norm-referenced score indicates an individual's standing in comparison with a standard reference group of persons who have taken the same test. In the interpretation of norm-referenced scores both the nature of the score transformation and the nature of the reference group must be considered.

Score Transformations

The most commonly used norm-referenced scores are percentiles, standard scores, and grade equivalents.

Percentiles. Percentile scores indicate relative standing in a group in very much the way rank ordering does, and they are often called percentile ranks. Because the meaning of a given rank depends on the size of the group ranked, percentiles adjust for group size by, in effect, indicating the equivalent of rank order in a standard group of 100 scores. The concept of rank order and the analogy of "a ladder with 100 rungs" are easy to understand, and percentiles are much used because of the ease with which their meaning can be communicated. The most likely misunderstanding of percentiles is an interpretation of them as indicating "percent correct", and in reporting test results to students and parents it is important to insure that this interpretation is not made.

A distribution of percentile scores from a group comparable to that on which the percentile norms are based will be rectangular, that is, will have about the same number of cases at each score. There will be, for example, about the same number of scores at the 98th percentile as at the 50th. Because there are far more cases near the middle of the raw score distribution than near the extremes, a small raw score change results in a much larger percentile change near the middle than near the extremes. This tendency to accentuate differences among mid-range scores and de-emphasize differences among extreme scores is a major disadvantage of percentiles.

Standard scores. This disadvantage is avoided by standard scores, in which differences are proportional to raw score differences. Standard scores are anchored at the mean of the norm group distribution, with units proportional to the standard deviation of the norm group distribution. The basic standard score transformation (z-score) is made by subtracting the mean from each score and dividing the remainder by the standard deviation, producing a score with mean of zero and standard deviation of 1. Because the fractional and negative scores produced by the z-score transformation are inconvenient, transformations that assign more units to the standard deviation and a positive score to the mean are usually used for score interpretation. Some standard score transformations commonly encountered are:

Score	Mean	S.D.	Relation to z
Stanine	5	2	2z + 5
ITED, ACT	15	5	5z + 15
T-Score	50	10	10z + 50
GATB	100	20	20z + 100
CEEB	500	100	100z + 500

Because standard score differences are proportional to raw score differences, comparisons of scores in different parts of the distribution are less subject to misinterpretation than comparisons of percentiles; and standard scores can be manipulated mathematically to obtain meaningful averages, correlations, etc. The meaning of a standard score is not immediately clear, however, without some understanding of its relation to a normal distribution of scores. Fig. 2 pictures this relationship for several standard score scales as well as for percentiles.

Figure 2 about here

Grade equivalents. Whereas a percentile or a standard score indicates the location of a score within one specified norm group distribution, a grade-equivalent score identifies a specific score distribution for which the obtained score is the median. The score distribution is for students at a particular grade level. For example, if the grade equivalent for a raw score of 38 is 4.0, 38 would be the median score of the norm group of beginning 4th-graders. Decimal parts are added to represent fractions of a 10-month school year, so that a grade equivalent of 4.2, for example, represents the median of students tested at the end of the second month of the 4th grade. Although there is a hypothetical norm group for each separate grade equivalent, in practice only a few levels are tested within the range of grade equivalents reported. A junior high school achievement test might be normed on students tested in the middle of the seventh (7.5), eighth (8.5) and ninth (9.5) grades, for example. Intermediate grade equivalents are determined by interpolation, and equivalents below the lowest group tested and above the highest group tested are determined by extrapolation.

Because grade equivalents are especially convenient for measuring progress, and because the significance of the score that is "built in" in the form of reference to educational levels seems especially easy to understand, grade equivalents are widely used. They have some disadvantages, however, that should cause users to interpret them with special caution. Although the meaning of a grade equivalent of 6.6 for a student in the middle of the 6th grade is clear, the meaning of the same score for a student in the middle of the 4th grade is less clear because we have no guidance as to whether such a deviation from the "expected" score is rare and significant or common. Certainly the two scores represent different kinds of achievement and have quite different meanings for the two students. Because students do not progress at the same rate in different subjects nor in the same subject at different levels, comparisons across subjects are difficult to interpret. At the high school level, where students are not taught every subject every year, grade equivalents have largely been abandoned for this reason. Finally, grade equivalents seem especially likely to be misinterpreted as performance standards. It seems easier to accept the notion that, on the average, half the students in the class must be below the 50th percentile than that half must be below "grade level".

Perhaps the simplest source of misunderstanding of a test score to be guarded against is confusion among the concepts underlying the various score transformations. A score of 75, for example, might be a grade equivalent with the decimal point omitted (common practice), a percentile rank, a standard score: mean 50, or a standard score: mean 100. Knowledge and understanding of the specific transformation is obviously essential to correct interpretation of the score.

Norm Groups

Because the meaning carried by norm-referenced scores is relative standing in a defined reference group, the characteristics of the reference group are most important.

Size. The group must have adequate size to provide stable results. If the norm group is a sample from a large population, it must be large enough so that variations due to sampling are minimized. Even when the norm group can be regarded as the entire population, as, for example, with school or class norms, anomalous and possibly misleading norms may be obtained if the group is very small.

Representativeness. Adequate size does not insure that a norm group will be adequately representative of the population specified. Norm groups are frequently difficult to obtain, and it is rare that samples can be randomly selected. The factors that do influence selection are likely to cause the norm group to be unrepresentative in unknown ways. Despite the care and expense applied to the development of national norms for standardized achievement batteries, the norms for different batteries are likely to be quite different. State norms may be easier to develop and more meaningful, but unless testing programs are mandated by the state, variations in testing practices among schools will make the development of representative norms difficult. "User norms", which are based on all the students from a defined population who happen to have taken the test, should be especially suspect.

Currency. Norms must be representative not only at the time they are developed but also at the time they are used. Norms that are not current may be misleading because they do not reflect educational and occupational changes.

Appropriateness. Given technical soundness in the form of adequate size, representativeness, and currency of a norm group, it is also important to consider the appropriateness of a norm group both for the student and for the decisions to be made. The student may be currently a member of the populations represented by some norms, so their appropriateness for the student is assured. A 9th-grade student who has taken the Lorge-Thorndike Intelligence Test (LTIT) and the Iowa Test of Educational Development (ITED) is a member of the populations represented by local school, state, and national norms for each test, all of which are appropriate for him. For decisions about his educational experiences in the immediate future, the local norms would be most appropriate because they indicate how he compares with his classmates in various areas. For longerrange planning the state norms, because they represent the students with whom he would most likely be compared in other high schools or post-high school institutions, would be more helpful. National as well as state norms might be used in evaluating how well the school's educational program achieved in various domains the kind of educational development expected for students with ability levels like those in the school.

For example, Alice's LTIT Verbal and Non-verbal scores of 59 and 52 put her at the 73rd percentile according to 9th-grade state norms, indicating an above-average student. On local norms for her school, however, these scores are at the 99th and 93rd percentiles, respectively, which suggest that she is likely to move much more rapidly than most of her classmates and may require special material to enable her to apply her abilities appropriately. In another school Brian's 9th-grade LTIT scores of 60 and 51 give him local percentile scores of 49 and 46, indicating an average student who should progress with the rest of the class. His percentiles of 75 and 70 on state norms, however, show above average ability, suggesting that his educational program should be one that will support many possible post-high school options.

Some norms represent populations of which the student is only potentially, not currently, a member. The MSAT norms in Table 1 are of both types. Each student who takes the test is clearly a member of the high school junior norm group, but only potentially a Minnesota college freshman. Similarly, technical school norms for scores on the General Aptitude Test Battery (GATB) and Minnesota Vocational Interest Inventory (MVII) (Pucel and Nelson, 1970a, 1970b) represent applicants who successfully completed various training programs. Such norms indicate not only relative standing in the norm population, but also whether it is reasonable to consider the student as a member of the population in the first place. According to Table 1 Cathy's MSAT score of 32 is average (53rd percentile) among high school juniors and also among Minnesota junior college freshmen (51st percentile) somewhat below average among state college freshmen (35th percentile), and substantially below average among liberal arts college freshmen (11th percentile). Nevertheless, Cathy clearly is a potential member of any of these groups, and it is reasonable to explore additional information about all three types of college. Douglas' MSAT score of 20, however, giving him a liberal arts college percentile of 1, indicates not only that Douglas' chances of successful performance in most Minnesota liberal arts colleges are quite low but also that his more specific estimates of performance in such colleges (see "Criterion-Referenced Scores") may not be applicable to Douglas because he is quite unlike the populations on which they are based. He is, however, a potential member of the junior college population (12th percentile), and performance estimates based on this group would be meaningful. It is important to note that, although members of such norm groups are identified after they become members of the defined population, their status at the time they were tested was the same as that of the students to whom the norms are applied. Thus the Minnesota college freshmen norm groups were tested as high school juniors, and the vocational program graduates were tested as applicants for the programs. Some norms, such as those often reported for employees in various occupations, are based on groups of persons already in the defined population at the time they are tested. In applying such norms to persons who are only potential members of the norm population, the influence on the test results of status at time of testing must be considered.

Multi-Score Tests

Profiles. Although the principles of test interpretation apply whether there is a single score or several, additional considerations are involved in tests or test batteries that produce multiple scores. Such scores are commonly presented on profiles, which offer a convenient means of displaying several items of information. A test profile is simply a graphic representation of several scores on comparable scales. Fig. 3 is an example of one such profile, showing Edwin's percentile scores on the DAT plotted as vertical bars extending above or below the midpoint of the score range for each test. Profiles are often prepared also with adjacent scores connected to each other, rather than to the midpoint of the scales, with straight lines, as in Fig. 4. The key word in the definition of a test profile is "comparable". It is inappropriate to profile raw scores because there is no basis for comparing raw scores on one test with those on another. The raw scores must be transformed to scales with comparable units, such as percentiles or standard scores. Furthermore, the transformations for all tests must be based on the same norm group. The provision of such comparability was an important objective and is now a basic feature of standardized batteries of aptitude and achievement tests.

Figure 3 about here

Difference scores. Because profiles do make score comparison easy, it is important to guard against over-interpretation of the differences that appear. The concept of error of measurement is especially important in evaluating differences in scores because the measurement errors cumulate, making the differences less reliable than the separate scores. In psychometric terms the standard error of the difference, $S.E._D$ is given by

$$S.E._D = \sqrt{S_{e_1}^2 + S_{e_2}^2} \qquad (5)$$

where S_{e_1} and S_{e_2} are the standard errors of the two tests whose scores are being compared. If the two standard errors are equal, formula (5) indicates that $S. E._D$ is about 1.4 times the standard errors of the individual tests. Computation of $S.E._D$ is cumbersome, and test publishers commonly offer convenient guides to the significance of score differences. When scores are reported as percentile bands, as on School and College Ability Tests and Sequential Tests of Educational Progress, bands that do not overlap are regarded as representing reliably different true scores. The manual for the High School Stanford Achievement Test (SAT) includes a table of standard errors of difference for each pair of tests in the battery, which should be consulted in evaluating SAT profiles. The reported $S. E._D$ of 5 for Spelling and Numerical Competence, for example, indicates that only one-third of the time would differences as large as 5 be obtained if the true scores for these abilities are equal, and only 5% of the time would differences as large as 10 be obtained.

Nearly all of the SAT $S.E._D$'s range from 4 to 6, although a few are as small as 3. Standardized tests used for individual student diagnosis and guidance should generally have reliabilities close to .9, which will provide $S.E._D$'s of about half a standard deviation (5 points on the SAT standard score scale). The profile for the DAT is printed with 1 inch=1 S.E. = 2 $S.E._D$ (approximately), so that differences of one inch or more correspond to a critical ratio of 2 (5 percent significance level) and may be regarded as significantly different. It is suggested that differences of one-half inch be interpreted if confirmed by other evidence. Comparison of Edwin's DAT scores in Fig. 3 with the 50th percentile reference line indicates that his scores are generally low, only the score on Mechanical Reasoning reaching the average level. Of the individual scores, Mechanical Reasoning is significantly different from all except perhaps Space Relations; whereas the other, despite their apparent differences, are sufficiently similar that differences among them should not be emphasized.

Profile applications. Profiles conveniently display both the overall level of a student's scores and areas of strength or weakness. Thus Frank's 11th-grade ITED scores in Fig. 4 show generally superior performance, with special strength in mathematics and some weakness in English expression, literature, and vocabulary. The scores provide a basis for discussion with Frank of his high school program for the remainder of his junior and senior year and of his post-high school plans. The counselor may wish to suggest that Frank concentrate on improving his communication skills in preparation for college work. Fig. 4 illustrates another use of profiles, namely for examining change. Frank's performance is very consistent from the 9th- to the 11th-grade, except for a fairly sizable improvement in his social studies score. This change may reflect an unusual course sequence in Frank's case, or perhaps the development of new interests.

Figure 4 about here

A test profile is a convenient way to summarize group as well as individual test performance. Overall performance of a school or class can be evaluated in comparison with the norm-group average, and strengths and weaknesses can be noted in the same way as with individual scores. Similarly the scores of the same group at two different times or of two different groups at the same time can be plotted on one profile to facilitate group comparisons and reveal changes. Special care must be taken in evaluating the magnitude of group differences in terms of score scales based on individuals, because the mean scores of groups are much less variable than individual scores. Whereas an individual percentile score of 60 differs rather inconsequentially from the midpoint of the norm group, a group mean at the 60th percentile is likely to be extremely high in comparison with other groups. Precise interpretation of such differences requires norms of group means.

To learn more about the nature of group differences revaled by the profile it may be helpful to examine the distributions of scores for individual tests. Fig. 5 shows 9th-grade percentile scores for the state norm group and the local percentiles for one school plotted against raw scores on the SAT-HS English Test. (Either percentile scores or cumulative percentages can be used, but both groups must be represented in the same way.) The school's average score is somewhat below the state mean, but the graph shows that this difference appears almost entirely in the lower part of the score distribution. This evidence does not explain the lower mean score, of course. One possibility is that the curriculum or the instruction is such that insufficient attention has been given to the less able students. An equally tenable hypothesis is that the English achievement scores reflect a similar distribution of learning ability of the students in the school. This hypothesis could be checked by examining scores of the same students on a general intelligence test such as LTIT in comparison with state norms.

Figure 5 about here

Similarity indexes. We sometimes wish to compare a student's scores with each of several reference groups. This may be done either by transforming the student's scores into standard scores or percentiles based on each reference group in turn, or by displaying the reference group distributions as well as the student's performance in terms of a single norm. Vocational training program norms for the GATB and MVII (Pucel & Nelson, 1970a, 1970b) are of the latter type. As a student's scores are compared with each of several groups and similarities and differences are noted, questions of how different the student is from a given group, or which group he is most like, arise; and the multiple comparisons produce more information than even profiles can conveniently summarize. To summarize such comparisons and obtain answers to questions like those above, indexes of profile similarity are used. One such index is the centour score, developed by Rulon, Tiedeman, Tatsuoka, and Langmuir (1967). Centour scores are like the scores on a target, where the bullseye, or the <u>center</u> (not the top) of the reference group, gets a score of 100, and the rings successively further in any direction from the center get successively lower scores. A centour score of zero, like missing the target completely, corresponds to a set of test scores outside of the "test space" occupied by any score in the reference group. (In actual use centour scores are usually based on more than two test scores, and therefore more than two dimensions, and take into account not just differences in individual scores but also in score combinations. Consequently the "target" is elliptical rather than round, and multi-dimensional rather than flat.) Just as a student's percentile gives the percentage of scores in the norm group <u>lower</u> than his, the centour score gives the percentage of score combinations in the norm group "further out" than his. Like all summaries, centour scores both reveal and conceal information. A student's centour scores reveal his similarity simultaneously to a large number of reference groups in which he may be interested. At the same time they conceal the specific ways in which he is similar to and different from each of them. Centour scores of 50 for three different groups may result from a student's having all higher scores than the average for one group, all lower scores than the average for another, and some higher and some lower than the average for the third. The differences are important, and to discover them we must go back to each profile and consider it in detail.

For example, Table 2 gives the centour score representations of seven GATB aptitude scores for five students with respect to 18 vocational training groups studied by Pucel and Nelson (1970a). Greg's centour scores show little similarity to any of the vocational training programs. Examination of his aptitude scores indicates that they are all lower, some of them substantially lower, than average for students in these programs. These are not the only training programs available, of course, nor do these tests measure all important abilities. It will be necessary for the counselor to explore with Greg his possible strengths in other areas and the ways in which these strengths match possible training or job opportunities.

Table 2 about here

Helen's scores, like Greg's, are dissimilar to those of graduates of all 18 programs, but the reason is quite different in her case. Most of her aptitude scores are quite high in comparison with the vocational school population. Helen may want to start with a more academic program, perhaps in a junior college, where she would have an opportunity more gradually to narrow her focus on a career program or a college transfer curriculum.

Although none of Irene's centour scores is high, she does have several--Agri-technology, Clerical training, Cosmetology--that suggest a careful look at these fields. Her weakest ability, according to the aptitude scores, is in working with numbers (which also influences the G score). Neither the centour scores nor the aptitude scores provide any information about the relative importance of this weakness for various occupations, but both the "construct validity" of numerical ability and the lower mean N score of the Cosmetology students suggest that it may be less significant in the Cosmetology program than in either of the other two.

In contrast to the other students, Jerry's scores fall in the area where all the training groups overlap. As a result, all of his centour scores are high, including several that are very high. Although the high centour scores provide some guidance, Jerry's ability pattern fits well into all the training groups, and other considerations than his abilities will likely determine his choice.

The pattern of Karen's scores is similar to Irene's, but all of her aptitude scores are higher, and this difference is reflected in higher centour scores in more areas. In addition to clerical and cosmetology training, practical nursing and secretarial training offer good possibilities.

It is important to note that similarity indexes, like all norm-referenced scores, do not in themselves indicate the likelihood of behavior of any kind other than that required by the tests themselves. To predict from the test scores to behavior in other situations we must rely on information about test validity, which is not introduced or represented by the norming process.

Interest profiles. Interest profiles are a special case of score representation by profile. Because of the way occupational scales are constructed, the practice has developed of norming each scale on its own occupational group, rather than on a single standard reference group for all scales. On the SVIB and MVII the scores are standard scores with an occupational group mean of 50 and S.D. of 10; on the Kuder Occupational Interest Survey the scores are, in effect, correlations between the students' responses and those of each reference group. Such profiles must be interpreted somewhat differently from those based on a single norm group. To provide a comparable reference point the SVIB and MVII profiles show the mid-third range of scores for a standard men-in-general group on each scale. These considerations do not apply to the Basic

Scales of the SVIB or the Homogeneous Scales of the MVII, which in each case are all normed on a single reference group.

Criterion-Referenced Scores

Whereas norm-referencing procedures provide meaning to test scores in terms of relative standing in a defined group of persons, criterion-referencing provides meaning in terms of expected behavior. The behavior may be defined by the test content itself, in which case we have content scores, or by a separate (criterion) measure, in which case we have predicted scores.

Content Scores

Scores on a content-referenced scale are summaries of the behavior on the test. Rate scores (e.g. reading rate, typing speed) and percentage scores are commonly used to represent performance, but to have meaning such scores must be accompanied by definitions of the content itself. Thus we have a "reading rate of 247 wpm on passages from The Readers' Digest", or "83 percent accuracy on 2-digit by 2-digit multiplication problems". If brief descriptions do not suffice to define the content, samples or examples may be used, such as "ability to spell 77 percent of words such as ambitious, anticipate, disappoint, eligible, indefinite, liability, miniature, oblige, sympathy, treasurer". To be most useful the content referred to should be not just described but scaled, so that mastery of a specified level implies mastery of all easier levels. Such scaling is just beginning in some fields, and few standardized instruments are available that reflect it. A fundamental requirement for the use of content-refereneed scores, of course, is satisfactory content validity.

Predicted Scores

If criterion-related validity has been demonstrated, the validity relationship can be used to report test performance directly in terms of expected criterion behavior. This is usually done in the form of either criterion estimates or expectancy tables or graphs.

Criterion estimates. Given a linear relationship between a test score (or scores) and a criterion variable, as reflected by a significant validity coefficient, an individual's expected score on the criterion variable can be predicted by the corresponding regression equation. From the correlation of .60 between a college aptitude index (I) and first-term freshman grades (GPA) in one university, for example, we obtain the following equation for predicting GPA from I:

$$GPA = .74 + .02\ I \qquad (6)$$

From this equation we learn that the predicted GPA corresponding to the minimum acceptable index of 40 is 1.54.

Like any test scores predicted scores are accompanied by uncertainty. In the case of predicted scores, however, this uncertainty is caused not only by the error of measurement of the test score, but also by measurement error in the criterion and by lack of perfect correlation between the true scores of the two measures. The combination of these three sources of error usually results in considerable imprecision in prediction, and it is important that this uncertainty be recognized in interpreting predicted scores. It is usually expressed as the standard error of estimate, computed as

$$S.E._{est} = \sqrt{S_c \ 1-r^2}\ , \qquad (7)$$

where r is the validity coefficient and S_c is the criterion standard deviation, and interpreted as the standard deviation of observed criterion scores around each predicted score. Fig. 6 portrays the standard error of estimate in relation to the standard deviation of criterion scores.

In the case of the regression equation discussed above the standard error of estimate is computed from the validity coefficient and the criterion S.D. to be .60. This figure, combined with the predicted GPA obtained above, indicates that of students with an index of 40 two-thirds will obtain GPA's between .94 and 2.14 and 95% will obtain GPA's between .34 and 2.74. The importance of taking into account the error of estimate in interpreting predicted scores is indicated by the width of the range needed to provide considerable assurance that the criterion score will indeed be included in the predicted range.

Predicted scores are used, of course, not for persons whose criterion scores are known, but for a new group of individuals (e.g., applicants) who have not been measured on the criterion. The standard error of estimate does not take into account sampling error in determining the regression equation. Interpretation of a predicted score and its associated estimate of precision assumes that the score comes from the same population represented by the sample on which the regression equation was determined and that this sample is large enough to provide accurate estimates of the regression parameters for the population.

Figure 6 about here

Expectancy Tables

Instead of predicting a specific criterion score and accompanying confidence band corresponding to each test score, a common practice is to report the probability of obtaining a criterion score within certain fixed ranges or above certain points. The criterion ranges for which probabilities are given are the same for all test scores, and the probabilities usually are reported for test score ranges rather than for individual scores. The expectancy tables relating high school rank (HSR) and aptitude test score to first-term college grades given in Tables 3 and 4 are examples of this method of criterion-referenced score interpretation. These tables were produced by determining the proportions of students in each fifth of the predictor distribution who obtained a college grade average of C or better and of B or better. Application of the tables can be illustrated with the scores of Linda, who has always done above average but not outstanding work in school (HSR=63) and has been developing a serious interest in art, in which she seems to have some talent. She wants a "good, general education" and plans to obtain it at the liberal arts college of the state university, which she can attend while living at home. Her aptitude test score of 36 is consistent with her high school record (junior percentile=69), and is sufficient to enter the university (college percentile=58). Linda's HSR is in the 60-79 range of the university expectancy table (Table 3) which is clearly below average for university females (above 12% and below 59%) but indicates a reasonable probability (67%) of obtaining at least a C average. Her chances of getting a B average or better are not high (10%). Information provided by the aptitude test expectancy table is consistent. Her college percentile, in the 40-59 range, is in the lowest quarter of entering university students and shows grade probabilities nearly identical to the HSR table. Linda has been considering, besides the university, the applied arts program in a state college. According to the state college expectancy table (Table 4) Linda's scores are below average for entering freshmen here also, but not quite so far below, and her chances of getting satisfactory grades are somewhat higher (79% and 80%). Properly interpreted these data can help Linda understand some differences between the two colleges, consider the kind of program and level of intellectual challenge most appropriate for her, and stimulate her to seek further information to help her resolve the choice.

Tables 3 & 4 about here

In comparison with criterion estimates based on regression equations, expectancy tables do not require a normal bivariate distribution underlying their interpretation, and they avoid an unwarranted appearance of precision. The uncertainties associated with measurement error and degree of relationship between the variables are reflected by the probability figures themselves. However, there are important cautions to be observed in using expectancy tables, cautions which reflect the fact that the tabled figures are actually proportions of previous classes rather than probabilities of future performance. (It has been suggested that they be called experience tables rather than expectancy tables.) First, in interpreting the figures as expectancies for new students we must assume that the composition of the new classes will be the same with respect to academic ability as the classes on which the tables are based and that they will be treated the same, i.e., that grading practices will remain the same. (Theoretically, it is unnecessary to assume that class composition remains the same if absolute marking standards do not change; but, because most grading is at least partly relative, it is more realistic to expect that a marked change in class composition will change the expectancies.) Entering classes will differ somewhat from year to year; but, unless there is a definite change in policy, such as an increase in admission standards, the differences are likely to be slight enough to maintain the validity of the expectancy tables. Over a period of years, however, such changes can cumulate, so the tables must either be reasonably current or be accompanied by evidence of consistency, such as predictor and criterion distributions that remain the same from year to year, if they are to be relied on. Second, it is important that each table be based on a group large enough to provide stable proportions. Like the standard error of estimate the expectancies reflect uncertainty due to measurement and prediction error but not that due to sampling variation. The number of cases in each predictor range (i.e., each row of the tables) determines the stability of the proportions for that range. It is for this reason that predictors are grouped into just five or six categories rather than a larger number that would permit more discriminating probability estimates. Because the classes on which the percentages are based are obviously not random samples from the schools' populations of entering students, interpretation of the standard error in terms of expected variation for future classes is not possible; but it is clear that the expectancies based on small N's should be used with extra caution. Finally, expectancy tables are necessarily based on the experiences of enrolled students; and these students form populations that differ from high school seniors in ways varying from one college to another as a result of both college admissions policies and practices and students' college selection decisions. To refer a student's score to a given expectancy table it must be reasonable to consider him a potential member of the population on which the table is based. If the table shows no scores in the range containing the student's score, it is clear that the table is not applicable to him. Even if a small percentage of the class had similar predictor scores, these students were atypical of their classmates with respect to these scores; and, inasmuch as they were enrolled despite this atypicality, they are likely to be atypical in unknown ways of students with similar scores. Thus, not only expectancies based on small N's, but also those based on small proportions of the class, should be viewed with caution.

Consider, for example, Michael's HSR of 36. The expectancy table for the University indicates that Michael's chances of obtaining passing grades (57%) or a B average or better (11%) are slightly larger than those of boys with HSR's in the range of 40-59. The first explanation to be considered for anomalies of this kind in the tables is a small number of cases, but in this case the N of about 70 (4% of 1981) should be sufficient to avoid fluctuations of this size merely because of sampling error. As noted above, students who enroll in a college despite very low predictor scores are likely to have special strengths in other areas or high scores on other predictors. Unless Michael has such strengths he would be unwise to rely too heavily on the tabled expectancies.

When predictions of the same criterion are made from more than one predictor, the results will not always agree. Norman is thinking of going to the state college, and referral of his aptitude test percentile of 40 to the expectancy table indicates that his chances of obtaining passing grades on the average are 70%, but according to his HSR of 39 his chances of getting a C average are only 30%. Which is correct? Part of the discrepancy may be ascribed to the fact that Norman's scores are at the upper edge of one interval and at the lower edge of the other. The coarse grouping results in some inaccuracy. Thus Norman's chances for a C average are undoubtedly more like those of a student whose HSR is 20, which is in Norman's interval with 30% probability. Some interpolation of probabilities may be made to adjust for this phenomenon, but even with such adjustments Norman's two predictions are discrepant. To determine which is more valid, Norman should consider with his counselor such information as whether special problems or responsibilities, which would not affect his college work, have held his high school grades down; whether his other test scores confirm the ability indicated by the aptitude scores or suggest that it is singularly high; whether Norman's academic motivation and study habits have changed in such a way as to give him a better chance of success in college than his high school grades indicate.

As the considerations above suggest, the expectancy tables do not in themselves decide whether or not a student should attend a given college. The same probability of success that leads one student to choose a college may lead another to look elsewhere. A 30% chance of success may encourage one student, whereas a 70% chance may discourage another. Nor should the tables be used to "shop" for a college by seeking to identify the college in which the student has the best chance of obtaining good grades. But they do provide information, suggest additional questions, and supply some answers to help clarify tentative choices or narrow the field of possibilities.

Discrepancy scores. Expectancy tables may be used not only to help reach decisions about the future but also to help explain the past. In the latter application, comparison of actual performance with expectancies based on previous scores may aid a counselor in understanding that performance. Quite different explanations of a student's failing grades, and different courses of action, may be indicated if his probability of a passing average were, say 17%, than if it were 70%.

Expectancy tables especially intended for this kind of interpretation, rather than prediction, of performance are sometimes provided for combinations of ability and achievement test scores. The manual for the SAT High School Battery presents quartile scores for each achievement test based on the distributions of scores for students in each stanine on the Otis Gamma Mental Ability Test. Orley's standard score of 57 on the English test puts him well above average (national norms) for 11th-graders in general, but more than three-fourths of 11th-grade students with Otis scores in the 8th stanine, as his is, score higher. This information may lead the teacher or counselor to a different interpretation of his English score than its percentile equivalent alone. Because the interest in expectancy tables of this kind is on the discrepancy between the ability and achievement scores, they are discussed here under the heading of "discrepancy scores"; but in reality such expectancy tables do not give criterion-referenced scores at all. Neither the ability test nor the achievement test is a criterion. The ability test, rather, is used to divide the norm group into more homogeneous subgroups so that more specific norms can be provided. Emphasis on the norm-referenced character of this kind of information may help to avoid reification of score differences into concepts such as "underachiever" and "overachiever". At the very least it is important to be aware of the differences between criterion-referenced and norm-referenced expectancy tables. Thorndike (1967) has pointed out a paradox in connection with the latter, namely that their value depends on the existence of moderate, rather than very high or very low, relationships between ability and achievement scores. If the relationship is very low, of course, achievement norms for low-ability students will not be appreciably different than those for high-ability students; and subdivision of the norm group will be useless. If the relationship is extremely high, on the other hand, the tests will be measuring much the same thing; and discrepancies between scores on the two instruments will be due largely to measurement error and not subject to meaningful interpretation. For prediction purposes, of course, the higher the relationship represented in an expectancy table, the more helpful is the information.

References

Campbell, D. P., and Johansson, C. B. "Academic interests, scholastic achievements and eventual occupation". Journal of Counseling Psychology, 1966, 13, 416-424.

Cronbach, L. J. Essentials of Psychological Testing (3rd ed.) New York: Harper and Row, 1970.

Layton, W. L. Construction of a short form of the Ohio State University Psychological Examination. Student Counseling Bureau, University of Minnesota.

Pucel, D. J., and Nelson, H. F. General Aptitude Test Battery Training Success Norms. Department of Industrial Education, University of Minnesota, February 1970 (a).

Pucel, D. J., and Nelson, H. F. Minnesota Vocational Interest Inventory Training Success Norms. Department of Industrial Relations, University of Minnesota, May 1970 (b).

Rulon, P. J., Tiedeman, D. J., Tatsuoka, M. M., and Langmuir, C. R. Multivariate Statistics for Personnel Classification. New York: Wiley, 1967.

Thorndike, R. L., Expectancy tables--sense and nonsense. Paper presented to the 17th Annual Conference of the Minnesota Statewide Testing Programs, Minneapolis, September 16, 1967.

TABLE 1

Minnesota Scholastic Aptitude Test Norms for
High School Juniors and Entering College Freshmen
1968

Percentile	Four-yr. Lib.Arts	U of M Four-yr Coll	State Colleges	Juniors Colleges	HS Juniors
99	68-75	67-75	61-75	60-75	64-75
98	67	66	59-60	58-59	61-63
95	65	64	56	53	57
90	63	61	52	49	52
80	58	56	46	43	45
75	56	54	44	41	42
70	54	52	43	39	39
60	51	48	39	35	35
50	47	45	36	32	31
40	44	42	34	29	27
30	40	39	31	26	24
25	39	38	29	25	22
20	37	36	27	23	20
10	31	32	24	19	16
5	27	27	20	16	14
2	21-23	20-23	16-17	13	11
1	0-20	0-19	0-15	0-12	0-10

TABLE 2

Aptitude and Centour Scores for Five Students

Centours	Greg	Helen	Irene	Jerry	Karen
1. Aircraft Mechanics	0	0	1	50	1
2. Agri-Technology	0	0	21	39	7
3. Automotives	3	0	12	82	20
4. Electronics	0	2	1	86	9
5. Carpentry	0	0	0	68	1
6. Farm Equipment Mech	0	0	2	82	5
7. Machine Shop	0	0	1	82	5
8. Mech Drafting	0	0	0	90	4
9. Power Home Elect	1	0	4	81	7
10. Printing, Graphics	4	1	2	82	12
11. Welding	7	0	6	68	11
12. Accounting	0	3	6	63	29
13. Clerical	0	2	25	64	68
14. Cosmetology	0	3	24	44	71
15. Data Processing	0	3	3	60	27
16. Practical Nursing	0	16	12	68	74
17. Sales	0	0	4	72	34
18. Secretarial	0	20	10	48	70
Aptitudes					
1. General	70	124	78	113	107
2. Verbal	78	139	96	100	104
3. Numerical	54	117	81	107	107
4. Spatial	97	117	94	137	101
5. Form Perception	84	129	107	111	140
6. Clerical Perception	100	129	115	118	139
7. Motor	82	103	101	111	132

TABLE 3

State University Expectancy Table
for First-Term Grade Average

FEMALES

	High School Rank N=1971			Aptitude Test N=1990		
		Chances in 100 of a freshman obtaining an average grade of:			Chances in 100 of a freshman obtaining an average grade of:	
%ile	% of class	C or Higher	B or Higher	% of class	C or Higher	B or Higher
90-99	35	92	47	34	90	44
80-89	24	80	18	19	79	24
60-79	29	67	10	25	71	14
40-59	10	56	7	17	65	8
20-39	2	47	9	5	54	3
1-19		*	-	1	55	9

MALES

	High School Rank N=1781			Aptitude Test N=1812		
		Chances in 100 of a freshman obtaining an average grade of:			Chances in 100 of a freshman obtaining an average grade of:	
%ile	% of Class	C or Higher	B or Higher	% of class	C or Higher	B or Higher
90-99	23	88	45	27	82	39
80-89	22	74	20	18	73	20
60-79	34	62	10	31	64	13
40-59	17	50	7	20	55	8
20-39	4	57	11	5	54	8
1-19		*	-		*	-

* the number of students in this cell is not large enough to produce a reliable percentage

\- no students in this cell

TABLE 4

State College Expectancy Table
for First-Term Grade Average

FEMALES

	High School Rank N=989			Aptitude Test N=940		
		Chances in 100 of a freshman obtaining an average grade of:			Chances in 100 of a freshman obtaining an average grade of:	
%ile	% of class	C or Higher	B or Higher	% of class	C or Higher	B or Higher
80-99	53	92	40	36	92	43
60-79	30	79	17	24	75	18
40-59	12	47	6	20	80	29
20-39	5	32	3	14	60	8
1-19		-	-	6	64	-

MALES

	High School Rank N=1067			Aptitude Test N=1029		
		Chances in 100 of a freshman obtaining an average grade of:			Chances in 100 of a freshman obtaining an average grade of:	
%ile	% of class	C or Higher	B or Higher	% of class	C or Higher	B or Higher
80-99	42	90	43	28	91	47
60-79	34	73	16	24	80	25
40-59	18	57	5	25	50	16
20-39	5	30	4	16	59	6
1-19	1	25	8	6	55	6

* the number of students in this cell is not large enough to produce a reliable percentage
- no students in this cell

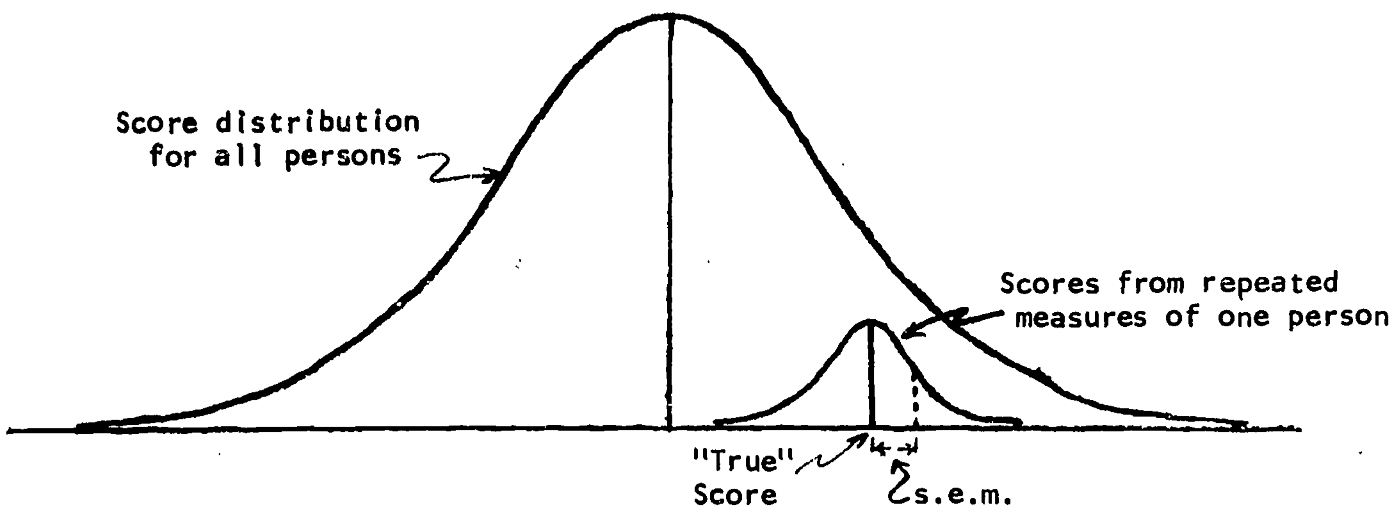

Figure 1. Standard error of measurement in relation to observed score distribution.

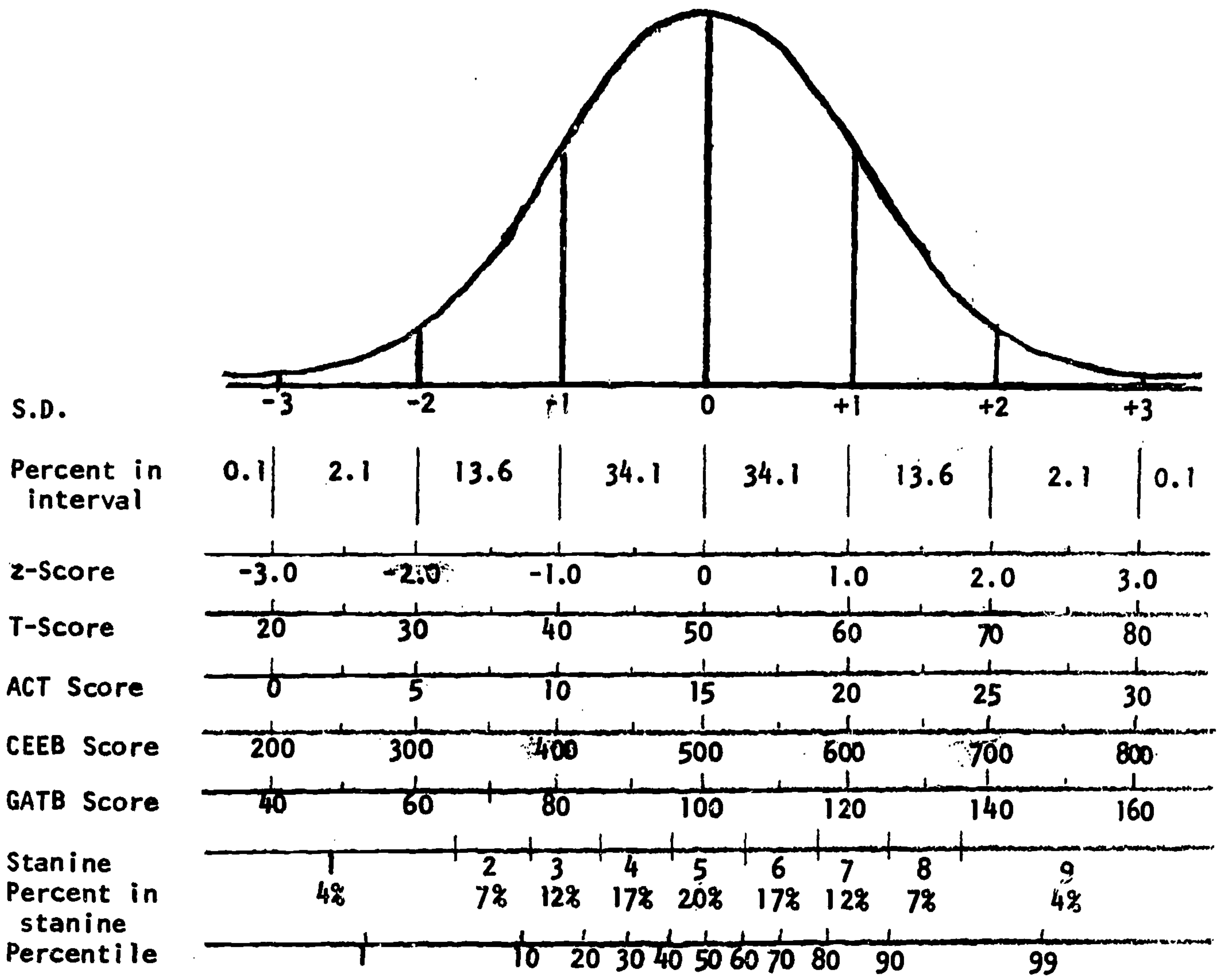

Figure 2. Common score scales and the normal distribution.

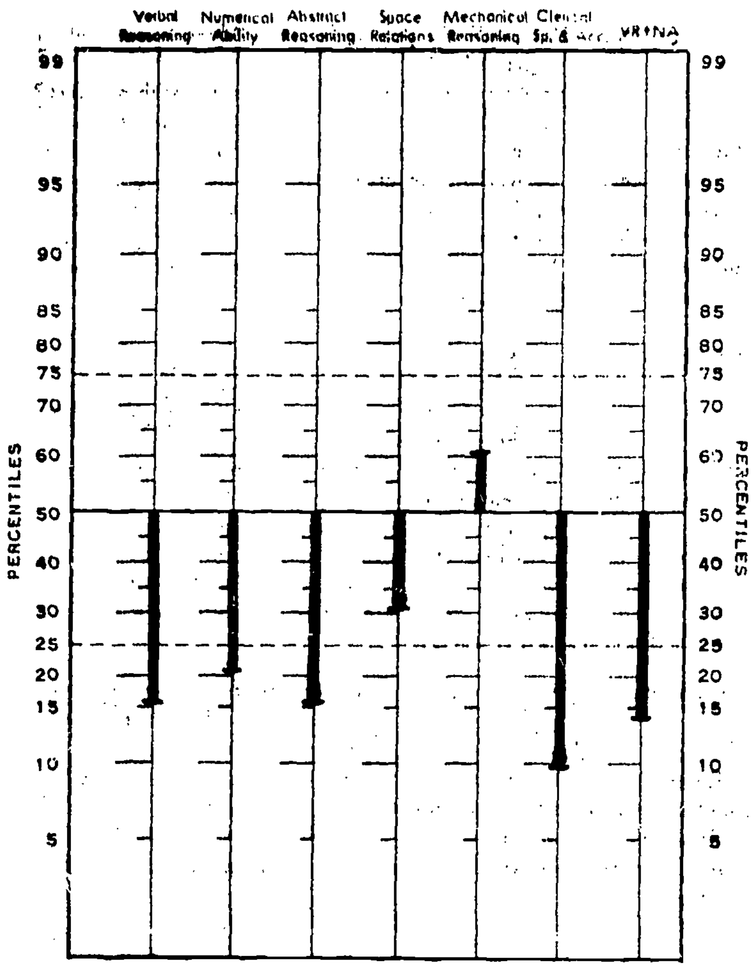

Figure 3. Edwin's DAT profile

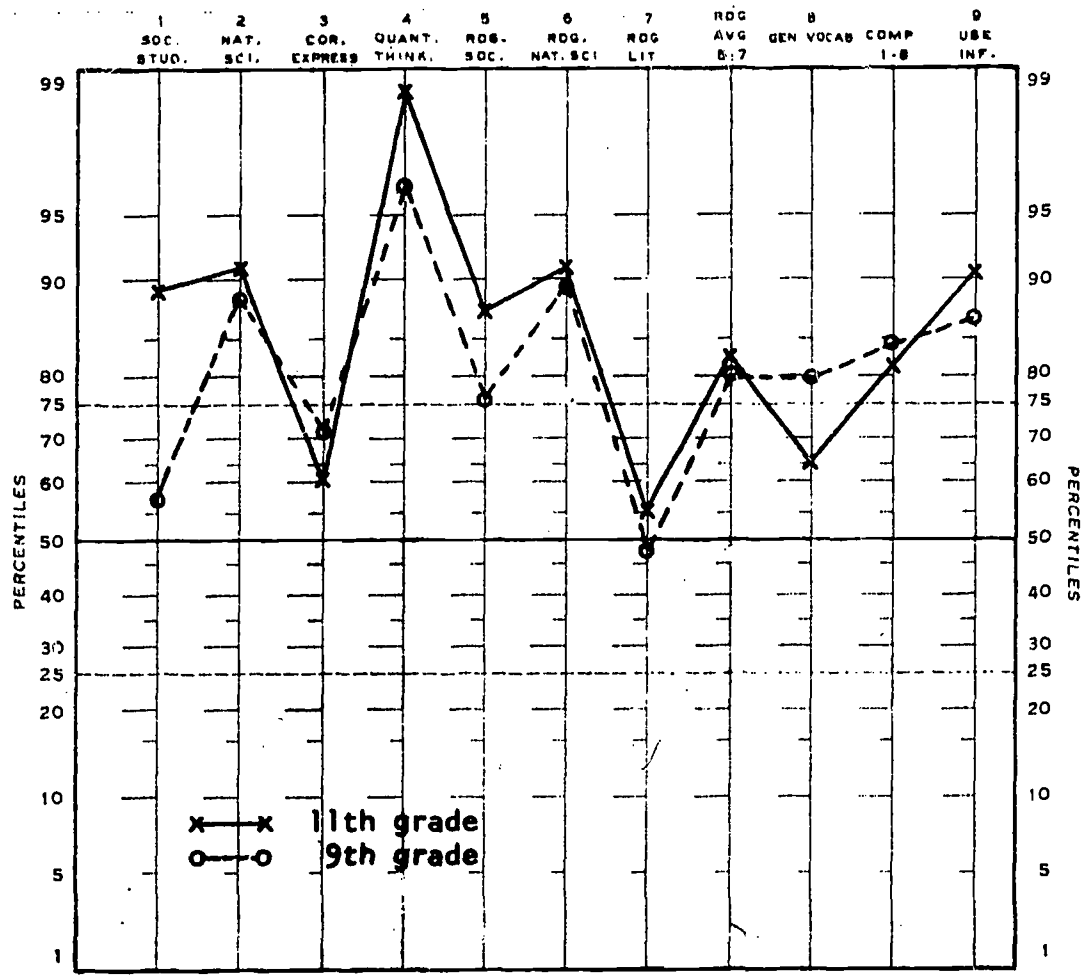

Figure 4. Frank's ITED scores

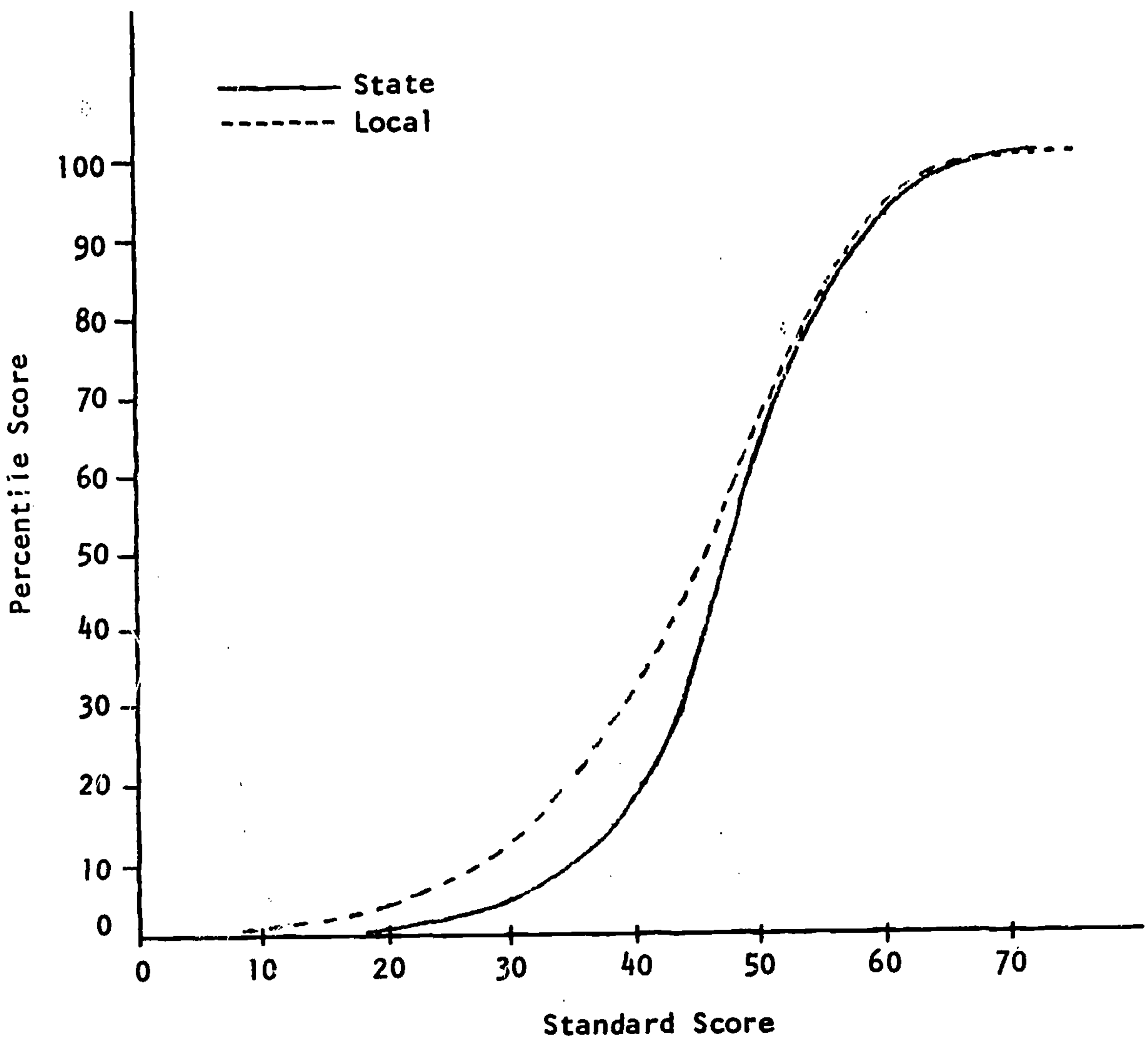

Figure 5. SAT-HS English score distributions for state and a local group.

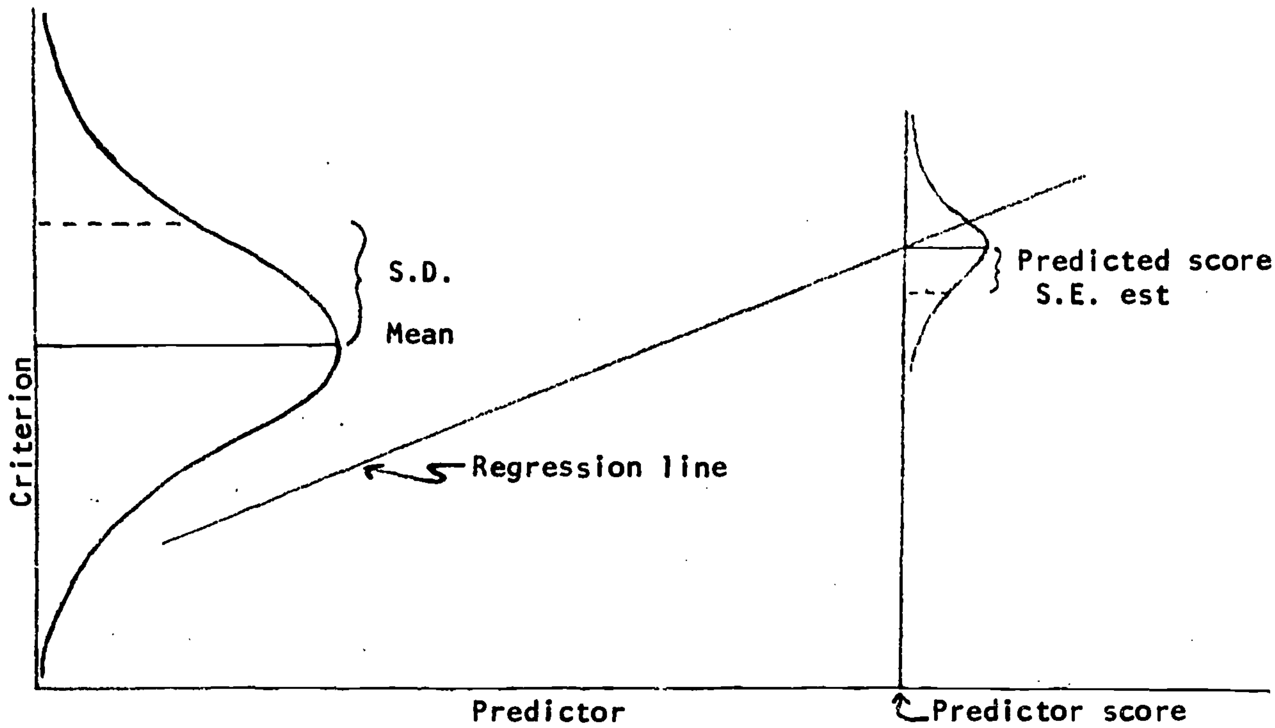

Figure 6. Relation of standard error of estimate to criterion standard deviation.

OFFICE OF NAVAL RESEARCH

PERSONNEL AND TRAINING RESEARCH PROGRAMS (Code 458)

DISTRIBUTION LIST

Contract No. N00014-68-A-0003 Contractor University of Minn. - Dunnette

4 Chief of Naval Research
Code 458
Department of the Navy
Washington, D. C. 20360

1 Director
ONR Branch Office
495 Summer Street
Boston, Massachusetts 02210

1 Director
ONR Branch Office
219 South Dearborn Street
Chicago, Illinois 60604

1 Director
ONR Branch Office
1030 East Green Street
Pasadena, California 91101

6 Director, Naval Research Laboratory
Washington, D. C. 20390
ATTN: Library, Code 2020 (ONRL)

1 Office of Naval Research
Area Office
207 West 24th Street
New York, New York 10011

1 Office of Naval Research
Area Office
1076 Mission Street
San Francisco, California 94103

6 Director
Naval Research Laboratory
Washington, D. C. 20390
ATTN: Technical Information Division

12 Defense Documentation Center
Cameron Station, Building 5
5010 Duke Street
Alexandria, Virginia 22314

1 Commanding Officer
Service School Command
U. S. Naval Training Center
San Siego, California 92133

3 Commanding Officer
Naval Personnel and Training
Research Laboratory
San Diego, California 92152

1 Commanding Officer
Naval Medical Neuropsychiatric
Research Unit
San Diego, California 92152

1 Commanding Officer
Naval Air Technical Training Center
Jacksonville, Florida 32213

1 Dr. James J. Regan, Code 55
Naval Training Device Center
Orlando, Florida 32813

1 Dr. Mortin A. Bertin
Educational Testing Service
20 Nassau Street
Princeton, New Jersey 08540

1 Technical Library
U. S. Naval Weapons Laboratory
Dahlgren, Virginia 22448

1 Research Director, Code 06
Research and Evaluation Department
U. S. Naval Examining Center
Building 2711 - Green Bay Area
Great Lakes, Illinois 60088
ATTN: C. S. Winiewicz

1 Dr. A. L. Slafkosky
Scientific Advisor (Code AX)
Commandant of the Marine Corps
Washington, D. C. 20380

1 Behavioral Sciences Department
Naval Medical Research Institute
National Naval Medical Center
Bethesda, Maryland 20014

1 Commanding Officer
Naval Medical Field Research
Laboratory
Camp Lejeune, North Carolina 28542

Distribution List - Page 2

1 Deputy Director
Office of Civilian Manpower Mgmt
Department of the Navy
Washington, D. C. 20390

1 Director
Aerospace Crew Equipment Dept
Naval Air Development Center
Johnsville
Warminster, Pennsylvania 18974

1 Chief
Naval Air Technical Training
Naval Air Station
Memphis, Tennessee 38115

1 Director
Education and Training Sciences Dept
Naval Medical Research Institute
National Naval Medical Center
Building 142
Bethesda, Maryland 20014

1 Commander
Submarine Development Group TWO
Fleet Post Office
New York, New York 09501

1 Commander
Operation Test and Evaluation Force
U. S. Naval Base
Norfolk, Virginia 23511

1 Mr. S. Friedman
Special Assistant for Research & Studies
OASN (M&RA)
The Pentagon, Room 4E794
Washington, D. C. 20350

1 Chief of Naval Operations, (op-07TL)
Department of the Navy
Washington, D. C. 20350

1 Chief of Naval Material (MAT 031M)
Room 1323, Main Navy Building
Washington, D. C. 20360

1 Mr. George N. Graine
Naval Ship Systems Command (SHIPT03H)
Department of the Navy
Washington, D. C. 20360

1 Chief
Bureau of Medicine and Surgery
Research Division (Code 713)
Department of the Navy
Washington, D. C. 20390

1 Chief
Bureau of Medicine and Surgery
Code 513
Washington, D. C. 20390

6 Technical Library (pers-11b)
Bureau of Naval Personnel
Department of the Navy
Washington, D. C. 20370

3 Personnel Research and Development Laboratory
Washington Navy Yard, Bldg. 200
Washington, D. C. 20390

1 Commandant of the Marine Corps
Headquarters, U. S. Marine Corps
Code Ao1B
Washington, D. C. 20380

1 Technical Library
Naval Ship Systems Command
Main Navy Building Room 1532
Washington, D. C. 20360

1 Technical Library Branch
Naval Ordinance Station
Indian Head, Maryland 93940

1 Library, Code 0212
Naval Postgraduate School
Monterey, California 93940

1 Technical Reference Library
Naval Medical Research Institute
National Naval Medical Center
Bethesda, Maryland 20014

1 Scientific Advisory Team (Code 71)
Staff, COMASWFORLANT
Norfolk, Virginia 23511

3 Technical Director
Personnel Research Division
Bureau of Naval Personnel
Washington, D. C. 20370

1 Deputy
Office of Civilian Manpower Management
Department of the Navy
Washington, D. C. 20390

1 Technical Library
Naval Training Device Center
Orlando, Florida 32813

1 Dr. Earl I. Jones
Director
Naval Training Research Institute
Naval Personnel & Training Research Laboratory
San Diego, California

1 Head, Personnel Measurement Staff
Captial Area Personnel Service Office - Navy
Ballston Tower #2, Room 1204
801 N. Randolph St.
Arlington, Virginia 22203

1 Director of Research
U. S. Army Armor Human Research Unit
Fort Knox, Kentucky 40121
ATTN: ATSAG-EA

1 Director
Behavioral Sciences Laboratory
U. S. Army Research Institute of Environmental Medicine
Natick, Massachusetts 01760

1 U. S. Army Behavior and Systems Research Laboratory
Commonwealth Building, Room 239
1320 Wilson Boulevard
Arlington, Virginia 22209

1 Division of Neuropsychiatry
Walter Reed Army Institute of Research
Walter Reed Army Medical Center
Washington, D. C. 20012

1 Behavioral Sciences Division
Office of Chief of Research and Development
Department of the Army
Washington, D. C. 20310

1 Commandant
U. S. Army Adjutant General School
Fort Benjamin Harrison, Indiana 46216
ATTN: ATSAG-EA

1 Dr. George S. Harker, Director
Experimental Psychology Division
U. S. Army Medical Research Laboratory
Fort Knox, Kentucky 40121

1 LTC William C. Cosgrove
USA CDC Personnel & Administrative Services Agency
Ft. Benjamin Harrison, Indiana 46216

1 Commandant
U. S. Air Force School of Aerospace Medicine
ATTN: Aeromedical Library (SMSL-4)
Brooks Air Force Base, Texas 78235

1 AFHRL (TR/Dr. G. A. Eckstrand)
Wright-Patterson Air Force Base
Ohio 45433

1 Personnel Research Division (AFHRL)
Lackland Air Force Base
San Antonio, Texas 78236

1 AFOSR (SRLB)
1400 Wilson Boulevard
Arlington, Virginia 22209

1 Headquarters, U. S. Air Force
AFPTRBD
Programs Resources and Technology Div.
Washington, D. C. 20330

1 AFHRL (HRTT/Dr. Ross L. Morgan)
Wright-Patterson Air Force Base
Ohio 45433

1 Lt. Col. John E. Dulfer
HQ, AFSC (SDEC)
Andrews Air Force Base
Washington, D. C. 20330

1 LTCOL F. R. Ratliff
Office of the Assistant Secretary of Defense (M&RU)
The Pentagon, Room 3D960
Washington, D. C. 20301

1 Dr. Ralph R. Canter
Military Manpower Research Coordinator
OASD (M&RA) M&RU
The Pentagon, Room 3D960
Washington, D. C. 20301

1 Dr. Andrew R. Molnar
Computer Innovation in Education Section
Office of Computing Activities
National Science Foundation
Washington, D. C. 20550

1 Dr. Alvin E. Goins, Exec. Secretary
Personality and Cognition Research Review Committee
Behavioral Sciences Research Branch
National Institute of Mental Health
5454 Wisconsin Avenue, Room 10A02
Chevy Chase, Maryland 20015

1 Director, National Center for Educational Research & Development
U. S. Office of Education
Dept. of Health, Education & Welfare
Washington, D. C. 20202

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (p-L)
U. S. Coast Guard Headquarters
400 Seventh Street, S.W.
Washington, D. C. 20226

1 ERIC Clearinghouse on Vocational and Technical Education
The Ohio State University
1900 Kenny Road
Columbus, Ohio 43210
ATTN: Acquisition Specialist

1 ERIC Clearinghouse on Education Media and Technology
Stanford University
Stanford, California 94304

1 Dr. Don H. Coombs, Co-Director
ERIC Clearinghouse
Stanford University
Palo Alto, California 94305

1 Dr. Richard C. Atkinson
Department of Psychology
Stanford University
Stanford, California 94305

1 Dr. Richard S. Hatch
Decision Systems Associates, Inc.
11428 Rockville Pike
Rockville, Maryland 20852

1 Director
Human Resources Research Organization
300 North Washington St.
Alexandria, Virginia 22314

1 Human Resources Research Organization
Division #1, Systems Operations
300 North Washington St.
Alexandria, Virginia 22314

1 Human Resources Research Organization
Division #3, Recruit Training
Post Office Box 5787
Presidio of Monterey, California 93940

1 Human Resources Research Organization
Division #5, Air Defense
Post Office Box 6021
Fort Bliss, Texas 79916

1 Human Resources Research Organization
Division #4, Infantry
Post Office Box 2086
Fort Benning, Georgia 31905

1 Human Resources Research Organization
Division #6, Aviation
Post Office Box 428
Fort Rucker, Alabama 36360

1 Dr. Robert J. Seidel
Human Resources Research Organization
300 North Washington St.
Alexandria, Virginia 22314

1 S. Fisher, Research Associate
Computer Facility
Graduate Center
City University of New York
33 West 42nd St.
New York, New York 10036

1 Dr. John C. Glanagan
American Institutes for Research
Post Office Box 1113
Palo Alto, California 94302

1 Dr. Robert Glaser
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, Pennsylvania 15213

1 Dr. Albert S. Glickman
American Institutes for Research
8555 Sixteenth Street
Silver Spring, Maryland 20910

1 Dr. Bert Green
Department of Psychology
Johns Hopkins University
Baltimore, Maryland 21218

1 Dr. Duncan N. Hansen
Center for Computer Assisted Instruction
Florida State University
Tallahassee, Florida 32306

1 Dr. M. D. Havron
Human Sciences Research, Inc.
Westgate Industrial Park
7710 Old Springhouse Road
McLean, Virginia 22101

1 Dr. Carl E. Helm
Department of Educational Psychology
Graduate Center
City University of New York
33 West 42nd St.
New York, New York 10036

1 Mr. Harry H. Harman
Division of Computation Sdiences
Educational Testing Service
Princeton, New Jersey 08540

1 Dr. Lee J. Cronbach
School of Education
Stanford University
Stanford, California 94305

1 Psychological Abstracts
American Psychological Association
1200 Seventeenth St. N. W.
Washington, D. C. 20036

1 Dr. Bernard M. Bass
University of Rochester
Management Research Center
Rochester, New York 14627

1 Dr. Lee R. Beach
Department of Psychology
University of Washington
Seattle, Washington 98105

1 Dr. Roger A. Kaufman
Graduate School of Leadership & Human Behavior
U. S. International University
8655 E. Pomerada Rod.
San Diego, California 92124

1 Dr. George E. Rowland
Rowland and Company, Inc.
Post Office Box 61
Haddonfield, New Jersey 08033

1 Dr. Mats Bjorkman
University of Umea
Department of Psychology
Umea 6, SWEDEN

1 Mr. Roy Ference
Room 2311
U. S. Civil Service Commission
Washington, D. C. 20415

1 Dr. Frederic M. Lord
Educational Testing Service
20 Nassau Street
Princeton, New Jersey 08540

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
Santa Barbara Research Park
6780 Cortona Drive
Goleta, California 93017

1 Dr. Stanley M. Nealey
Department of Psychology
Colorado State University
Fort Collins, Colorado 80521

1 Dr. Gabriel D. Ofiesh
Center for Educational Technology
Catholic Universtiy
4001 Harewood Road, N.E.
Washington, D.C. 20017

1 Mr. Luigi Petrullo
2431 North Edgewood Street
Arlington, Virginia 22207

1 Dr. Len Rosenbaum
Psychology Department
Montgomery College
Rockville, Maryland 20852

1 Dr. Arthur I. Siegel
Applied Psychological Services
Science Center
404 East Lancaster Avenue
Wayne, Pennsylvania 19087

1 Dr. Paul Slovic
Oregon Research Institute
Post Office Box 3196
Eugene, Oregon 97403

1 Dr. Diane M. Ramsey-Klee
R-K Research & System Design
3947 Ridgemont Drive
Malibu, Califronia 90265

1 Dr. Ledyard R. Tucker
University of Illinois
Psychology Building
Urbana, Illinoia 61820

1 Dr. John Annett
Department of Psychology
Hull University
Hull
Yorkshire, England

1 Dr. Lloyd G. Humphreys
Assistant Director for Education
National Science Foundation
Washington, D.C. 20550

1 Dr. Joseph W. Rigney
Behavioral Technology Laboratores
University of Southern California
University Park
Los Angeles, California 90007

1 Educational Testing Service
Division of Psychological Studies
Rosedale Road
Princeton, New Jersey 08540