DOCUMENT RESUME

ED 332 749 JC 910 268

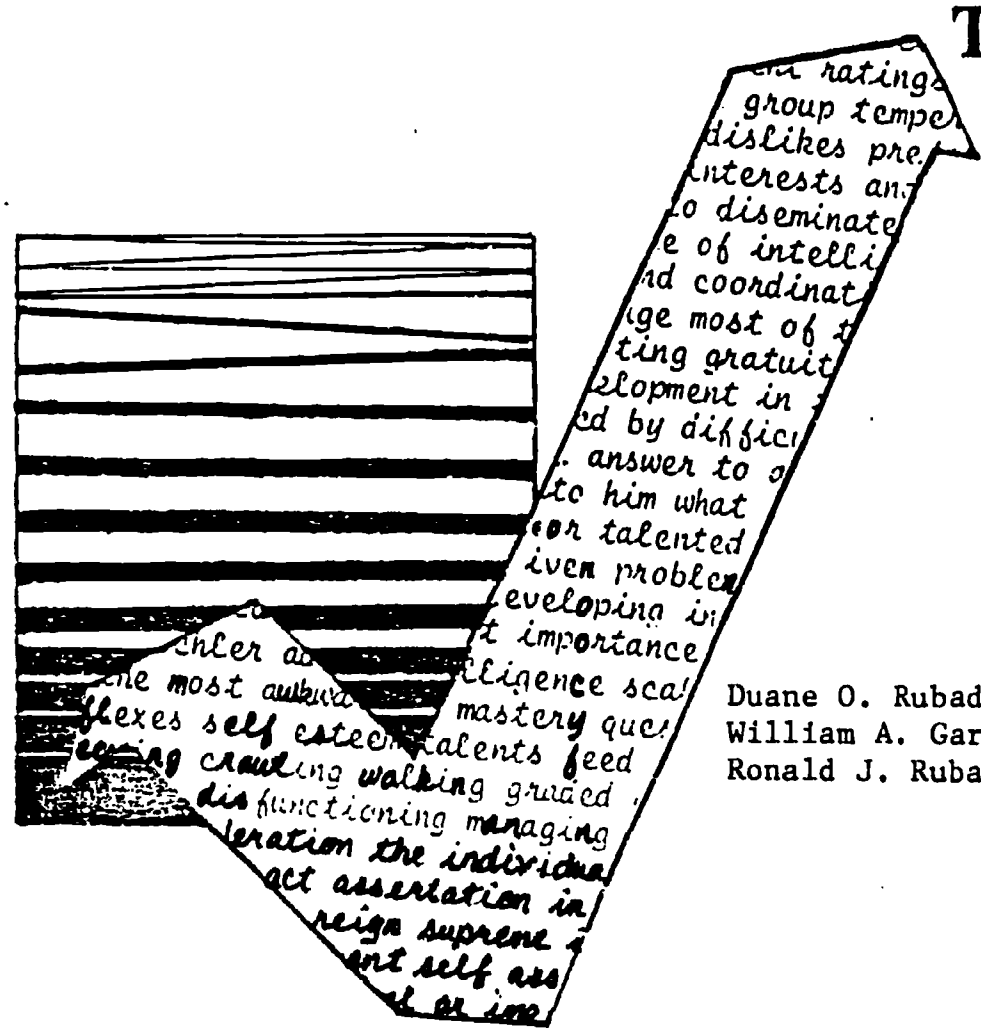
| | |
|---|---|
| AUTHOR | Rubadeau, Duane O.; And Others |
| TITLE | Standardized Testing. The Teacher's & Instructor's Guide. |
| INSTITUTION | College of New Caledonia, Prince George (British Columbia). Centre for Improved Teaching. |
| PUB DATE | 90 |
| NOTE | 31p. |
| PUB TYPE | Guides - Classroom Use - Teaching Guides (For Teacher) (052) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | Community Colleges; Elementary Secondary Education; Foreign Countries; *Standardized Tests; Teaching Guides; Testing Problems; *Test Interpretation; Test Norms; *Test Reliability; Test Results; *Test Selection; *Test Use; *Test Validity; Two Year Colleges; Two Year College Students |

ABSTRACT

Designed to assist community college practitioners in understanding and utilizing standardized test results for making instructional, guidance, and administrative decisions, this three-part guidebook examines the uses of test results, the evaluation of standardized tests, and procedures for interpreting test scores. Section I discusses the reporting of test performance, and reviews the functions of number ordering and ranking as they relate to test scoring. This section includes an examination of the nominal, ordinal, interval, and ratio measuring scales, and the data organizing procedures of two-way classification, qualitative classification, ranking, and scores expressed in uniform units. Section II reviews factors to consider when selecting a test, presents a 14-item test evaluation outline, cites sources for information about tests, and describes test evaluation criteria, including content validity, construct validity, predictive validity, face validity, reliability, standardization, and norms. The final section presents a 12-step test interpretation procedure oriented around student involvement (i.e., a procedure that allows students to apply the test results to their own educational, vocational, and personal circumstances and decisions). The 12 steps include discussing the test; inquiring about students' feelings while taking the test and their perceptions of why the test was selected; discussing norm groups; asking students to estimate their performance; discussing the use of ranges in test interpretation; providing self-estimate forms; and discussing actual scores. A list of sources for information about standardized tests is appended. (PAA)

# THE TEACHER'S & INSTRUCTOR'S GUIDE TO

# STANDARDIZED TESTING




Duane O. Rubadeau
William A. Garrett
Ronald J. Rubadeau

College of New Caledonia Press
1990

# THE TEACHER'S & INSTRUCTOR'S GUIDE TO STANDARDIZED TESTING

Duane O. Rubadeau
Co-ordinator, Centre for Improved Teaching
College of New Caledonia
Prince George, British Columbia

William A. Garrett
Director, Centre for Improved Teaching
College of New Caledonia
Prince George, British Columbia

and

Ronald J. Rubadeau
Director of Special Services
School District #23
Kelowna, British Columbia

ISBN 0-921087-18-7

Additional copies available from:

College of New Caledonia Press
3330 - 22nd Ave.
Prince George, B.C.
V2N 1P8
Phone (604) 562-2131

# TABLE OF CONTENTS


Page

Preface ..... i

Section I. The Use of Test Results ..... 1
- Measuring Scales ..... 2
  - Nominal ..... 2
  - Ordinal ..... 2
  - Interval ..... 3
  - Ratio ..... 3
- The 2-Way Classification ..... 4
- Qualitative Classification ..... 4
- Ranking ..... 4
- Scores Expressed in Uniform Units ..... 6

Section II. Evaluating Tests ..... 7
- Selecting a Test ..... 7
- Test Evaluation Outline ..... 8
- Sources of Information About Tests ..... 8
- Validity ..... 8
  - Content ..... 9
  - Construct ..... 9
  - Predictive ..... 10
  - Face ..... 11
- Reliability ..... 11
- Standardization ..... 12
  - Norms ..... 12

Section III. Interpreting Test Scores ..... 15
- Developing Student Involvement ..... 15
- An Outline of the Student Test Interpretation Procedure ... 16

Bibliography ..... 22

Appendix 1 ..... 23

# PREFACE

Teachers, instructors and other personnel in the educational setting are often faced with test results without having a clue as to what the results mean or how they are to be used. In this case, the person is at a disadvantage in how to use or interpret the test results. There are also occasions when people in the educational setting need to select a test to assess a program or a particular course without having a specific set of objectives to guide the choice of a test. Here, the person is often in a bind due to lack of information for evaluating just how "good" the test was that had been selected.

Some schools, school districts and provinces have a comprehensive testing program where all students in particular educational levels are assessed each year. Unfortunately, many teachers and instructors do not utilize this information because they do not have the background, experience, or training to use the test results.

Students at all levels in the educational system need to have more information about themselves. One source of information can be provided by utilizing valid and reliable tests. We feel there is a definite need for understandable information about testing and the interpretation of test results; hence, the development of The Teacher's and Instructor's Guide to Standardized Tests. We have tried to avoid the jargon of the field and have dealt with the statistical concepts in such a manner that a Ph.D. in mathematics is not necessary for comprehension. We think the "Guide" will help you as a test user, to understand and utilize test results for making instructional, guidance and administrative decisions.

The Guide is comprised of three sections. Section I deals with the use of test results; evaluating tests is covered in Section II; and Section III presents a procedure for interpreting test scores.

We give our sincere thanks to Linda Fieguth for her excellent work on the manuscript and Denise Chappelle for the cover design.

Duane O. Rubadeau

William A. Garrett

Ronald J. Rubadeau

July 1983.

# SECTION I

# THE USE OF TEST RESULTS

Any test that is to be administered to students, whether it is standardized or instructor-made, should be of value in the educational decision-making process. If the test is not going to provide meaningful information for you, it is a waste of money and certainly is a waste of your time.

In general, test results are used to assist in making four types of educational decisions. First, there are the instructional decisions, where the test results help you to develop a remediation or individualized educational program for a student. Second, are the guidance decisions, where the test results are given to the student to aid in making personal decisions. Administrative decisions are the third use of test results, which usually deal with curricular changes or grouping of students for instructional purposes. The fourth and final use of test results is in the area of research.

## REPORTING TEST PERFORMANCE

One of the real disasters that we have going in the fields of psychology and education is that many people assume that our tests are a lot more accurate than they really are. Another problem is the tendency of people within these two fields to want to make precise statements about a student's performance, when precise statements are not warranted on the basis of test limitations.

It appears that this confusion may result from many people not being aware of the different measuring scales that we have for recording and describing behaviour.

## MEASURING SCALES

There are four basic measuring scales which are directly related to the four primary functions of numbers:

1. First, numbers allow us to differentiate. That is, they stand for different things. For example, 7 is different from 8 and 27 is different from 14.

2. Second, numbers allow us to order in terms of greater or less. For example, 9 is greater than 7 and 14 is greater than 9.

3. Third, numbers may represent equal intervals. For example, the difference between 5 and 6 is the same as the difference between 6 and 7.

4. Fourth, and finally, numbers are used to form equal ratios. For example, 10/5 is the same as 70/35 and 21/7 is the same as 63/21.

Now, to relate these primary functions of numbers to the measuring scales.

1. Nominal Scale:

Differentiation is the only number function involved on this scale. That is, we can identify different observations or classify similar observations. Hence, we can differentiate between groups or categories, but cannot indicate the nature or degree of differences between these groups or categories.

2. Ordinal Scale:

Here we have two of the numerical functions involved - differentiation and order. As a result, we can not only identify events, but also indicate the direction of their relative standing. The ordinal scale does not, however, indicate anything about the magnitude of differences. For

example, scores on this scale would not tell us how much louder one sound is than another, just that one sound is louder.

3. Interval Scale:

This scale has three functions. It allows us to differentiate, order and because it has equal intervals, it also indicates the magnitude of differences between scores. The zero point on this scale is established arbitrarily, as for example in our measure of time. The zero point we work from is the date of Christ's presumed birth and we work back or forward from that date.

4. Ratio Scale:

The ratio scale differentiates, orders, has equal intervals and has equal ratios. An example of such a scale would be our measure of physical weight. The ratio scale is the most accurate measuring scale we have available. Perhaps the easiest way to explain this difficulty of different levels of accuracy of our measuring scales, is to look at the accuracy of some comments that are often made when a student has been assessed with a Wechsler Adult Intelligence Scale - Revised (WAIS-R).

1) Sam has a Full-Scale I.Q. of 106 on the WAIS-R.

2) According to the WAIS-R results, Sam has average ability.

3) On the WAIS-R, a measure of verbal and performance ability, Sam scored about the same level as the average person for his age.

Statement #1 is inaccurate, but is heard very commonly. The problem here, is that when specific scores are used, it implies the test has greater accuracy than it has.

Statement #2 is a bit better than statement #1, but still gives the impression that the test is more accurate than it really is.

Statement #3 is the best we can do with the information we have on Sam.

Another problem that adds to the complexity of reporting test performance is in the way we organize our data. In effect, we are dealing with the degree of refinement of our measures. There are four levels of complexity in reporting test performance that correspond to the four types of measuring scales.

1. THE 2-WAY CLASSIFICATION:

On the surface, this is the simplest, broadest and most general level of measurement. Probably the best example from the academic setting is the Pass-Fail classification. Here, the evaluation is made on the basis of an external criterion, rather than having students compete with each other for letter, or numerical grades.

2. QUALITATIVE CLASSIFCATION:

Every day, we make judgmental statements such as: "Bright", "Dull", "Mediocre", "Lovely", "Beautiful", and so on. The problem with these statements and with these types of measurements is that they are extremely vague and that differences in the meanings of the words for the individuals attempting to communicate can be phenomenal. For example, what one person might consider as bright, another person would view as a bit above average.

3. RANKING:

The individual's rank in the group is the Third level of refinement in measurement. Here, we rank individuals from most able (Rank 1) to least able (lowest rank) on a test or other tasks on which they are scored by relatively uniform standards. The main drawbacks to this procedure is that they are not an established criterion of performance and

differences in ranks are usually not equal. For example, on a 20-item spelling test given to 10 students in the Bluebird Spelling Group, the students, number of correct words, and rank in group are shown in Figure 1.

| Student | Correct Items | Rank in Group |
|---|---|---|
| A | 18 | 1 |
| B | 16 | 2 |
| C | 15 | 3 |
| D | 14 | 4.5 |
| E | 14 | 4.5 |
| F | 12 | 6 |
| G | 11 | 7 |
| H | 10 | 8 |
| I | 8 | 9 |
| J | 4 | 10 |

Figure 1.
Ranking of Students According to Performance
on a Twenty-Item Spelling Test

An examination of Figure 1 shows the uneven differences in ranks. Student G had 11 items correct for a rank of 7, Student H had 10 items correct for a rank of 8, Student I had 8 items correct (2 more wrong) but a rank of 9, and Student J had 4 items correct (far less than student I) but still had a rank of 10.

The disadvantages of this method should be obvious, yet we use this approach as the basis for much of our decision-making regarding the assignment of grades and also for reporting test score results when we convert scores into percentiles.

The advantages of the rank in group measurement are that it is accepted by parents and it is convenient.

4. SCORES EXPRESSED IN UNIFORM UNITS:

This is the most refined level of measurement. Unfortunately, only certain types of variables lend themselves to this method - e.g. Rate and Speed of Response, Weight, Height, and Money. Variables that do not lend themselves to this method of measurement include: anxiety, happiness, homosexuality and compatibility.

# SECTION II

# EVALUATING TESTS

There are times when you may want to be able to select an appropriate test for a particular purpose, and there are other times you may wish to evaluate tests already in use. We suggest you apply the following factors to each instrument you consider.

## SELECTING A TEST:

First, a test should not be selected because it happens to be popular or prominent. Unfortunately, these are frequently the only reasons that out-moded, out-dated tests continue to be used. On the other hand, a test should not be thrown out or eliminated as a possibility simply because of the publication date. Rather, a test should be judged in comparison to more recently designed instruments that serve the same function.

Second, factors such as: length of testing time, ease of scoring, ease of interpretation, and face validity are necessary, but not sufficient reasons for choosing a test. Tests should be chosen according to specific criteria to prevent superficial factors from having too great an influence in the choice.

Third, a test certainly cannot reveal information not contained in the questions and its content should not be judged from the title alone. It is only through examination of the items and rationale of the test that we can determine the validity of a test to yield the information we are trying to obtain.

There are many different procedures for doing a systematic examination of tests. The one procedure we have evolved over time requires that you answer each statement for each test that you consider using.

TEST EVALUATION OUTLINE:

1. Title of Test:
2. Author:
3. Publisher:
4. Date of Publication:
5. Cost:
6. Forms and Levels:
7. Type of Test and Purpose:
8. Time Required:
9. Brief Description of the Test:
10. Aspects Tested:
11. Adequacy of Administration, Scoring & Interpretation Procedures:
12. Norms:
13. Reliability and Validity:
14. General Evaluation:

SOURCES OF INFORMATION ABOUT TESTS:

The single best sources of information about standardized tests are the Mental Measurement Yearbooks published by Oscar Buros. Buros assists in the test selection operation by providing information on good and weak points of the tests being evaluated. In addition, two or more experts discuss each test to provide further information for judging the worth of the test. Other sources, particularly for the newer tests include: Journal of Educational Research, Perceptual and Motor Skills, Journal of Educational Measurement, School Review and the Personnel and Guidance Journal. In addition, there are the test publishers, listed in Appendix I.

VALIDITY:

A test may be considered valid if it meets the objective for which it was intended. Another way of putting it is, if a test measures what it was intended to measure, it is valid.

Basically, there are four types of validity: content, construct, predictive, and face validity. We will examine each of the four types of validity to assist you in determining whether a test is good for the purpose for which you intend to use it.

(a) Content Validity:

Content validity is primarily concerned with the subject matter or content covered by the test. Where we are primarily concerned with content validity would be on tests of achievement and also on instructor-made tests. The purpose of these two types of tests is to assess the amount of subject matter learned and/or the behavioural change that has taken place in a given course or program. The achievement and instructor-made tests should provide an adequate sample of items out of all of the possible items from which the test might have been drawn.

Content validity is determined by comparing the test items and the content covered in the course or program. Usually an outline of the content and the test items are compared to determine if the test items are appropriate and representative of material that was covered in the course or program.

(b) Construct Validity:

Construct validity is concerned with the measurement of traits or psychological constructs. The purpose of validating a test designed to measure a trait or a construct is to determine whether the trait being measured is actually the one you were trying to measure. For example, we may want to measure the trait of honesty. There are several ways we can go about determining whether or not the test items measure the trait of honesty:

1. We might consider the experts, such as psychologists, psychiatrists, teachers, counsellors and so forth, to rate the items making up the

test in terms of how appropriate they are for measuring the trait of honesty. (We are assuming they are accurate in their evaluation of the items.)

2. We might compare the scores on our test of honesty with scores that are obtained on a test that has already been validated for measuring the trait of honesty. An example of this variety of validity would be in the use of the Wechsler Adult Intelligence Scale - Revised (WAIS-R), which is a very highly accepted and respected test of intelligence. Hence, many of the tests of intelligence (a construct) are validated against the Wechsler test.

3. We also might compare scores of individuals who are assumed to rate high on the trait of honesty with individuals who are assumed to rate low on the trait. For example, the scores obtained from certain types of prison inmates might be compared to the scores obtained from priests and ministers. Assuming Jim Jones and others of his ilk are not making up the ministerial sample, the test should yield scores that distinguish between the two groups.

A point that has to be kept in mind when dealing with construct validity is that it is specific to a particular group and/or situation. Thus, what might have construct validity for one group in a particular situation may not have validity for a similar group and situation.

(c) Predictive Validity:

Predictive validity is a common type and relatively easy to explain and understand. For example, we can use an entrance examination for a particular college. The grades that the student will obtain in college are predicted on the basis of the entrance exam performance and are validated by follow-up studies to determine how accurately the grades were predicted. In this type of validity, we are assuming that the ability

that is reflected in the test performance was possessed prior to entering the college and that it is required for success in college.

(d) Face Validity

A test is said to have face validity when the items look like they measure what the test is supposed to measure. That is, when the test taker looks at the items, he/she sees them as being relevant to the purpose of the test. Thus, a test item involving bead-stringing will probably not be perceived as being part of an intelligence test.

RELIABILITY:

Any test is nothing more than a way of obtaining a sample of students' behaviour in order to estimate what his/her performance would be in a wider range of situations. For example, if we want to test a student's competence in mathematics, there is no way we could ask him/her to do all possible mathematics-type items. Rather, we select a sample of items from the many possible mathematics items. We then assume that the student will have the same degree of accuracy in items that were not on the test, as he/she had with the items that were on the test. In other words, if we have a really good selection of items from across the field of mathematics, and the student does well, we are willing to bet he/she would do well on any other type of mathematics' items too. For this assumption to hold, our test (sample of behaviour) must be valid and our test must be reliable as well. Reliability means that whatever the test measures, it measures it consistently.

The degree of consistency of a test is expressed as a reliability coefficient. The closer the coefficient (correlation) is to 1.00, the more consistent the test. For example, one way to determine the reliability of a test is to administer it twice to the same group. Then run a correlation between the two sets of scores. If the reliability coefficient is up in the .90's, you have a reliable test.

One factor that you must keep in mind is that a test must be reliable in order to be valid; however, a test may be invalid and still be reliable.

STANDARDIZATION:

A standardized test is a measuring instrument which must be administered under a standard set of conditions and scored in a predetermined manner. Further, the test results are interpreted in terms of normative group performance. The normative group was hopefully drawn as a representative sample from a specified population of a particular educational and age level. The main purpose of a standardized test is to make it possible to compare or rank students in terms of the specific behaviours sampled by the test.

It is usually not too difficult to accomplish the administration of the test under standard conditions and to score the test in a predetermined manner. The difficult task is in choosing appropriate norms.

Norms:

In effect, norms provide a yardstick with which a student's raw test score can be compared. In any normative group, half of the group is above average and half is below average. "High" and "Low" performances are viewed in relation to how far the raw test score is from the average or mean.

Norms involve a comparison of some type. For example, percentile norms make it possible to compare the achievement of an individual with many other individuals in the same grade, course or age group. Other types of norms offer similar comparisons. Our main concern here, is with the identification of the "other people" used as a basis for comparison. We refer to these "other people" as the normative sample or norm group.

One way to compare scores on a test is to convert the scores to standard or z-scores and compare them to the norm group values (see Rubadeau - A Guide to Elementary Statistics - for the computation formula for the z-score).

This method of using the z-score can also be used to compare one student's scores on several tests, as long as the Mean and Standard Deviation are known for each of the tests.

When the norm group is truly representative of all comparable individuals throughout the country, the norms are called national norms. Thus, national norms for a Grade 12 Biology test would be based on results from a sample of Grade 12 students studying Biology, selected from all Grade 12 students across Canada. If national norms are used, the sample should be truly representative of those with whom the comparisons are to be made. Unfortunately, on many tests, the norm groups come from two or three large cities in Canada, or they come from a specific geographic area. Other factors that may influence test performance besides age and education are sex, socio-economic status and language spoken. As a result, these factors should be taken into account when choosing the sample. Norms that are developed out of convenience of locality or availability of subjects tend to be suspect. You will find it well worth your while to develop local norms for these tests if you plan to use them over time.

A common norm group set of data provided with a number of tests is for a group called People-In-General. These norms are seldom useful, except for ability tests which yield I.Q. scores and for certain clinical situations. The general idea is that People-In-General norms may be misleading, where the average for a group has no meaning for a person who must perform specific tasks. For example, the skills needed by one secretary for one firm might be limited to typing and filing, while the skills of a secretary in general might include shorthand, use of various machines and receptionist duties.

Occasionally you will find that the recommended norm group may not be appropriate for meaningful interpretation of the test results. For example, math aptitude scores are usually lower for females than for males. A girl who is considering majoring in math, or who is considering a career in which math aptitude is important, will probably be competing with males. In such situations, we have found it quite helpful to compare the girls' scores with

scores obtained by males - which will be her reference group in making decisions.

# SECTION III

# INTERPRETING TEST SCORES

The test interpretation procedures that we recommend are oriented around student involvement. This procedure allows the student to apply test results to his/her own plans and to determine whether the test results are appropriate for making educational, vocational, and personal decisions. A detailed outline of the test interpretation procedure is presented to assist you in learning the technique.

## DEVELOPING STUDENT INVOLVEMENT

The first thing the student should be involved with is describing the test and determing what it measures. The student should also decide whether the information the test will yield will be of value for the kinds of questions that need to be answered. A relatively easy way to develop this type of orientation is to have the student estimate his/her score before they find out their score on the test. This requires them to think through their performances in similar situations, and also to compare their performance with that of other students of the same age and/or grade level. Most students are able to estimate their scores with amazing accuracy. In cases where there are discrepancies between the estimate and the actual performance, it is not difficult to get a discussion going to clarify the information before misunderstandings or misinterpretations occur.

In our approach, we refer to ranges of scores rather than to specific test scores. By so doing, we can take into account measurement errors inherent in the tests. For example, if we deal with the concept of I.Q., we do not talk about the student's raw I.Q. score, but rather would deal with the range of scores within which his/her score happens to fall. If the student receives a Full-

Scale I.Q. of 116 on the Wechsler Adult Intelligence Scale - Revised (WAIS-R), he/she would fall into the Above Average Range of Intellectual Functioning which runs from I.Q.'s of 110 to 119. If the student had a measured I.Q. of 96 on the WAIS-R, this would fall into the Average Range of Intellectual Functioning which runs from I.Q.'s of 90 to 109.

This procedure of using ranges coupled with verbal descriptions of expected behaviours for individuals falling within these ranges is appropriate for use with individuals, as well as with large or small groups. You can maintain confidentiality in the group quite readily, as all data is given to each student in written form, and no individual data is revealed to the group. We believe this procedure will provide the most accurate data possible, yet prevent over-interpretation or misinterpretation.

AN OUTLINE OF THE STUDENT TEST INTERPRETATION PROCEDURE:

This outline can be used as a convenient reference to keep at hand while you are going through the procedure with your students.

1. Establish Rapport With Your Students:

Begin the session by going directly to the interpretations. Don't waste time with small talk or trivia trying to make the student feel at ease. The students may perceive the delay as a stalling tactic on your part to avoid dropping the "bad news" on them. A simple greeting, followed by something like "I believe we are here to discuss your performance on the Bennett Mechanical Aptitude Test that you took last week" is probably a good way of establishing rapport with your students. This type of instructor behaviour is positive and is generally interpreted as helpful, friendly and reassuring, by the students.

2. Discuss the Test:

Here, the task is to help the students recall what the test was like. This is easily done by bringing out the test booklet and showing them the sample items. Also, bring in where and when the test was taken and who administered the test.

3. Inquire About How They Felt While Taking The Test:

This is an important area to talk about, especially if you have any doubt about the validity of the student(s)' results. For example, if the student was having a bout with the green-apple, two-step (commonly known as the flu) and had to run for the washroom every five minutes, it is quite likely that the validity of that student's results are questionable. When necessary, arrangements should be made for the student to take another form of the test at a convenient time.

It is also important that students understand that the test should provide an estimate of their "typical" behaviour. By getting into the discussion of typical behaviour at this point, you avoid having the students make excuses for their scores before learning how they scored. That is, the students decide at this time whether the testing situation was a valid one for them at the time. Now, if you have been paying attention, you also realize that you have involved your students in their first decision regarding the test results - the acceptance and use of the results, rather than not accepting and ignoring the results.

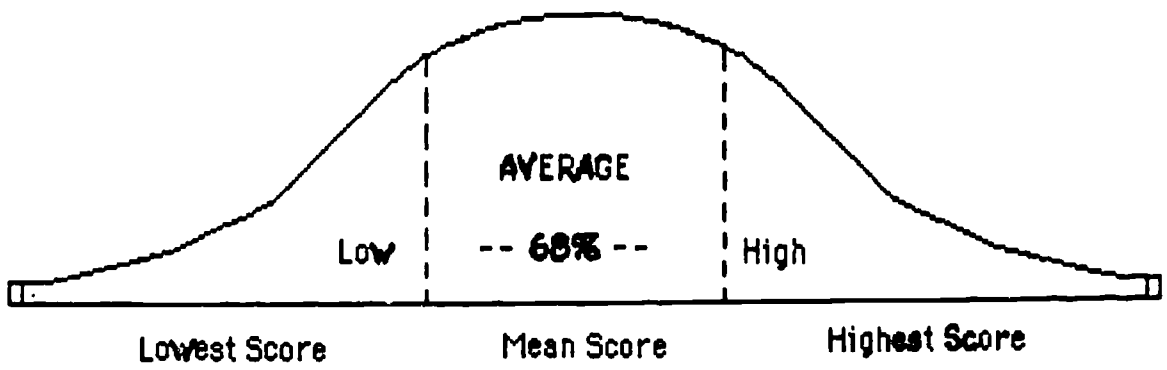
4. Inquire About Why They Think This Test Was Selected:

This provides the opportunity to talk about what kind of information the test scores give. It also gives entry into a discussion of what type of decisions can be or will be made on the basis of the test results and how this information might be used in the decision-making process.

5. Inquire About What We Know About the Data From the Norm Group With Whom Their Scores Will be Compared:

Here you have to be quite specific in stating that "your scores will be compared with those of all first-year college students in B.C." or "with all first-year Diploma Nursing students in Canada."

You will probably have to bring in such things as the normal distribution or normal curve (oddly enough, many children become familiar with this concept in about Grade 5 or 6.) Draw the distribution or have one ready as you talk, explaining that the distribution represents every score from the lowest score to the highest score. (See Rubadeau - A Guide to Elementary Statistics.)




THE NORMAL DISTRIBUTION

6. Ask Your Students That Without Knowing Anything Else About the Way a Person Scored on the Test, Where Would a Person's Score Most Likely Fall on the Normal Distribution?

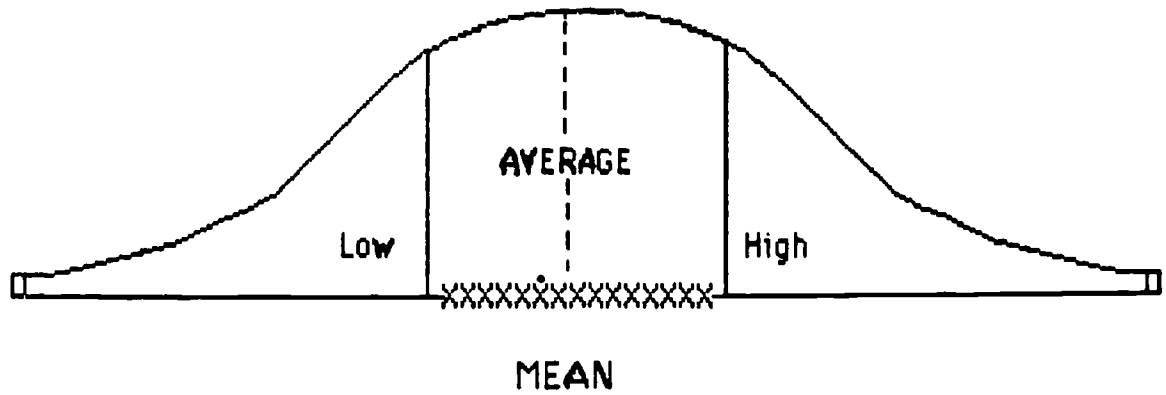
They will probably guess that the person's score would fall into the average range, as this range contains two-thirds of the scores in the distribution.

Also point out that about 15% of the scores are below the average range and 15% are above the average range of scores.

7. Set up a Hypothetical Case:

Here we are dealing with the variability of a single person's scores on the same test. The example you might use is that "when you took the test the first time, your score was exactly at the mean." "If you took the test again today, would your scores be in the exact same place on the distribution?"

The students would probably agree that their scores would not be in the same area due to a variety of factors such as: guessing differently, feel better or worse today, learned some of the answers I didn't know, and so on. As they discuss their reasons for why their scores might be different on the second testing, you can illustrate the fluctuations in the scores by adding X's to the figure of the Normal Distribution.




8. Discuss the Use of Ranges for Interpreting Test Scores:

Here, you have to explain that since the test scores obtained on successive testings vary to some extent, and we have to estimate where your "true" score would probably fall. For this reason we use a range of scores for reporting test data - e.g. - average range, below average range, above average range and so on.

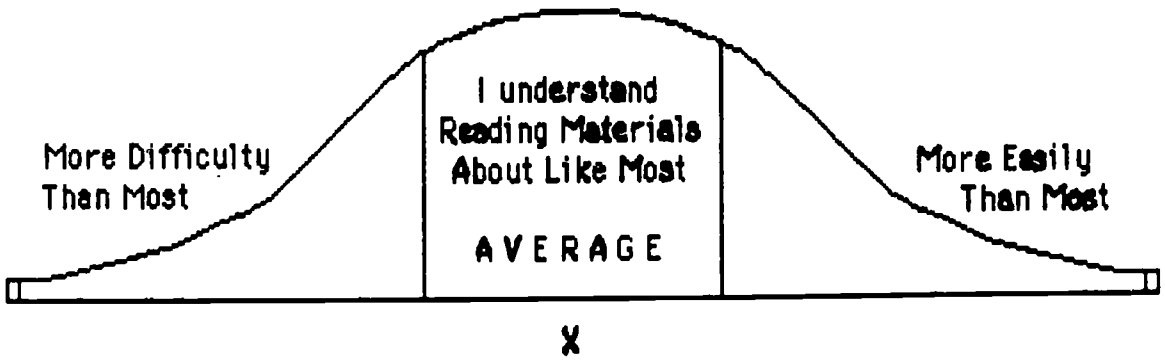
9. Hand Out a Self-Estimate Form:

The self-estimate forms are nothing more than copies of the normal distribution you have been using for illustration purposes. Ask each student to estimate his/her performance for each variable on the test by putting an "X" on the distribution for that variable. The self-estimate form should provide a space for the name of the test and a key code for the different variables.

For example:

READING DEVELOPMENT TEST

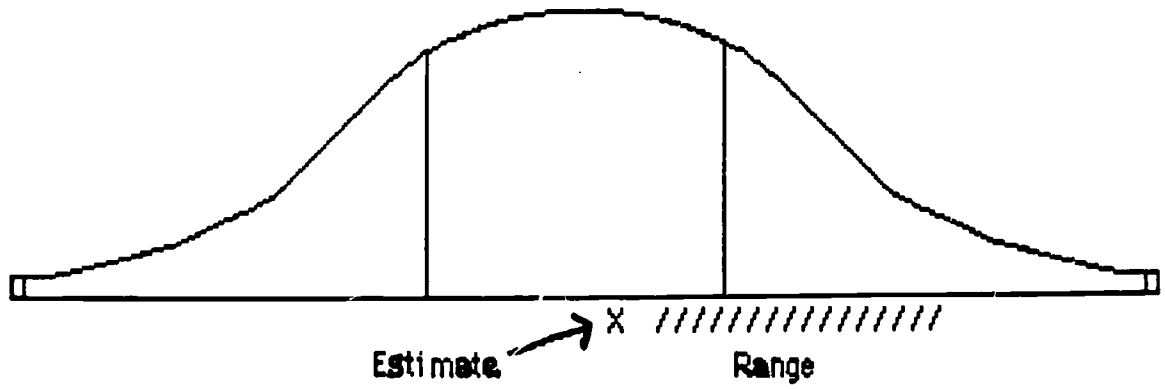
Reading Comprehension - The Ability to Understand Written Sentences and Paragraphs




10. After Completion of the Self-Estimate, Hand out Students Test Results Reported in the Form of a Range:

Now ask the students to compare their estimated scores and their obtained scores. Explain to them that their estimate is accurate if it falls anywhere within the range.

For example:




11. Inquire About How They Performed and If They Performed the Way They Expected:

As mentioned previously, students tend to be uncanny in estimating their test scores. When the estimate and the range scores are very different, there are usually one of two reasons that account for the disparity. First, either the student did not understand the definition of the variable being assessed or second, the scoring and reporting of the range of scores was in error. Hence, this technique is useful for picking up errors in reporting scores as well as being a very handy counselling device.

Other inquiries you may want to make about test scores without revealing personal data to the group are: "Did you score high on those areas you expected to score high on, and low on those you expected to score low on?" "Are your grades what you would expect for scoring the way you did on this test?" "Do the test scores support the educational and vocational choices you have made?"

12. Check to see if the Students have any Questions.

# BIBLIOGRAPHY

Aiken, L.R. Psychological Testing and Assessment, 3rd ed., Toronto: Allyn and Bacon, 1979.

Bloom, B.S., G.F. Madaus & J.T. Hastings. Evaluation to Improve Learning, New York: McGraw-Hill, 1981.

Brown, R.G. Measuring Classroom Achievement, New York: Holt, Rinehart and Winston, 1981.

Ebel, R.L. Essentials of Educational Measurement, 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1979.

Kaufman, R. & S. Thomas. Evaluation Without Fear, New York: New Viewpoint, 1980.

Noll, V.H., D.P. Scannell & R.C. Craig. Introduction to Educational Measurement, 4th ed., Boston: Houghton-Mifflin, 1979.

Salvia, J. & J.E. Ysseldyke. Assessment in Special and Remedial Education, 2nd ed., Boston: Houghton-Mifflin, 1981.

Thorndike, R.L. & E.P. Hagen. Measurement and Evaluation in Psychology and Education, 4th ed., New York: John Wiley & Sons, 1977.

# APPENDIX 1

## SOURCES OF INFORMATION ABOUT

## STANDARDIZED TESTS

Addison-Wesley Publishing Co. Inc., South Street, Reading, MA, 01867

American Guidance Service, Inc., Publishers Bldg., Circle Pines, MN, 55014

Bobbs-Merrill Co., Inc., 4300 W. 62nd St., Indianapolis, IN, 46206

Bureau of Educational Measurement, Kansas State Teachers College, Emporia, KA, 66802

Bureau of Educational Research & Service, U. of Iowa, Iowa City, IA, 52240

Bureau of Publications, Teachers College, Columbia University, N.Y., N.Y., 10027

CTB/McGraw-Hill, Del Monte Research Park, Monterey, CA, 93940

Consulting Psychologists' Press, Inc., 577 College Ave., Palo Alto, CA, 94306

Cooperative Test Division, Educational Testing Service, South Street, Reading, MA, 01867

Educational & Industrial Testing Service, P.O. Box 7234, San Diego, CA, 92107

Harcourt Brace Jovanovich, Inc., 757 Third Ave., N.Y., N.Y., 10017

Houghton-Mifflin Company, 1 Beacon St., Boston, MA, 02107

Institute for Personality & Ability Testing, 1602 Coronado Drive, Champaign, IL, 61822

Ohio Testing Services, 751 Northwest Blvd., Columbia, OH, 43212

Personnel Press, 191 Spring St., Lexington, MA, 02173

The Psychological Corporation, 757 Third Avenue, N.Y., N.Y., 10017

Scholastic Testing Service, Inc., 480 Meyer Road, Bensenville, IL, 60106

Science Research Associates, 155 N. Wacker Drive, Chicago, IL, 60606

Sheridan Psychological Services, Inc., P.O. Box 6101, Orange, CA, 92667

The Steck Co., P.O. Box 16, Austin, Texas, 78761

Western Psychological Services, 12035 Wilshire Blvd., Los Angeles, CA, 90025

IN CANADA:

Angus, Tien & Associates, Ltd., 2639 Kingsway Avenue, Port Coquitlam, B.C., V3C 1T5

Guidance Centre, 1000 Yonge Street, Toronto, Ontario, M4W 2K8

Institute of Psychological Research, Inc., 34 Fleury Street West, Montreal, Quebec, H3L 1S9