Course: Educational Assessment and Evaluation (8602)
Semester: Spring, 2021
ASSIGNMENT No. 2

Q.1 What is the relationship between validity and reliability of test.

It might seem that validity is one of those concepts reserved for foundational or "basic" research projects. But that is simply not the case. Validity should be of concern to anyone who is making inferences and decisions based on some type of data. And the more profound the consequences of those inferences and decisions, the more important validity becomes. As teachers and instructors, the inferences that we make about our students' learning and the decisions we then make about facilitating their learning carry with them potentially deep consequences. For example, we might infer (based on data) that a student has not mastered a concept, which is then reflected in their assigned grade, which could ultimately have consequences for course completion, continuation of study in the degree, and graduation. Therefore we need to make sure that our inferences are sound, and that the decisions we make which follow from these inferences are well supported.

My goal in this post is to convince you that assessment validity should be of concern to everyone who teaches. Some backing for this assertion follows. We need to:

- ensure that we are making sound inferences about our students' learning of the target concepts and content so that we can help guide their future learning.
- help develop alignment between our own assessment of student learning and those made (inferred) by external assessments (e.g., large-scale assessments such as NAEP, PISA, ACT, SAT, GRE, or other external assessments such as Concept Inventories).
- contribute to a culture which views teaching as a complex, highly skilled, and professional endeavor.

Before going any further, let us agree that assessment and testing are not dirty words. Both are an essential part of good teaching practice. In order to teach well, we must continually assess well. While the focus of my argument in this piece is more related to summative assessments of learning, the same principles apply to formative assessment practices.

The concept of test validity (as it is referred to in the research literature) is rich and complex. Historically, validity has been conceptualized within one of three models or frameworks, or some combination thereof. These are the criterion, content, and construct models. I will briefly describe each of these before turning to a more contemporary conception of validity, that being the unified, argument-based approach.

The criterion model of validity is based on the concept that a test is valid if scores on that test correlate with some other "objective measure" of the factor being measured, such as performance on some task (Angoff, 1988). The criterion model could be applied either concurrently or in a predictive fashion (Kane, 2006). In the former, the criterion score with which test scores are correlated is collected at the same (or at least near) time with the test scores. Predictive applications involve the correlation of test scores with some future performance (e.g., grade in a subsequent course of study). In the past, predictive applications of the criterion model were

widely used in testing efforts (e.g., in the armed services), while concurrent applications were more often used in making a case for the validity of a new instrument where an existing measure was the basis for the correlation (Angoff, 1988).

The content model of validity asks if test scores "based on a sample of performance in some area of activity [can serve] as an estimate of overall skill level in that activity"(Kane, 2006, p. 19). The observed performance (test score) can be considered an appropriate estimate of overall performance in the domain if "(a) the observed performances can be considered a representative sample from the domain, (b) the performances are evaluated appropriately and fairly, and (c) the sample is large enough to control sampling error" (Guion, 1977 as cited in Kane, 2006). Content validity is concerned with the representativeness of the tasks on the test and the ability to generalize the observed scores on that test to some estimate of ability within the content domain.

Construct validity considers the construct (the characteristic that the test is designed to measure) within a larger theory, which in turn is related to other theories in a hypothetico-deductive way. Networks link these theories to each other and to observations and/or scores which can serve as bases for making inferences about the existence of that construct in an individual. These networks of theories and inferences assume that the theory is fairly well-defined, but that it admittedly only approximates reality (Cronbach & Meehl, 1955). Construct validity has been further broken down into a substantive component, a structural component, and an external component (see Kane 2006 p.20 for a brief summary of this from Loevinger 1957). The construct model was originally proposed by Cronbach and Meehl as an alternative to the criterion and content models.

By the 1970's, researchers began advocating a unified approach to validation efforts. Messick (1989) was one of the first to outline a unified approach. Using the Construct model as a basis for this unified approach, he defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13, emphasis in original). One issue with this conception is that it does not provide much guidance for the validation effort. Because so much data and evidence could be considered relevant to making a case for the validity of a test, validation could end up being a lengthy, messy process.

Presenting the idea that test validation is an evaluation, Cronbach (1988) proposed the idea of a validity argument. He defined this argument as an evaluation of the proposed uses and interpretations of test scores. Describing the traditional trinity of validity conceptions (criterion, content, and construct) as "strands within a cable of validity argument," Cronbach emphasized the need to play devil's advocate in the development of a persuasive validity argument. The argument should not only seek to confirm, but also to falsify and contribute to revision — especially for a "young" instrument, such as that presented in this study.

A very approachable summary of this unified conception of validation and a guide for structuring validation efforts is presented in latest edition of the Standards for Educational and Psychological Testing (American Educational Research Association, et al., 2014). In keeping with Cronbach's conception of the validity

argument, the Standards define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Also emphasized is the idea that it is the score interpretations themselves that are evaluated in a validity argument — not the test itself. The implications of this idea are clear: if test scores are used or interpreted for a purpose other than the one being validated, then a new validity argument must be crafted. As stated above, one potential complication with this concept of validity is that the validation process can become overwhelming. A vast amount of evidence could be brought to bear in supporting test use and score interpretation, and evaluation of that interpretation in light of that evidence could be complex. What is needed is a structure for guiding the validity argument, and for allocating resources during the development of such an argument.

The Standards provide such a structure. They begin by calling for an articulation of the proposed score interpretations and test use. The notion of a construct is central to this model — the proposed score interpretation is to be articulated in terms of the construct of measurement. Following the proposed use and interpretation is an explication of a set of propositions which support the proposed score interpretations. It is these propositions which provide the structure for the validity argument, as they guide the collection of evidence needed to build the argument. Again in keeping with Cronbach's conceptions, the Standards state that the identification of these propositions can be facilitated by playing devil's advocate, and considering alternative or rival hypotheses.

1. **State (for yourself) how your test will be used, and how you will interpret the test scores.** And importantly, be able to defend this statement to others. If someone were to ask you "why do you give students a final exam?" what more could you say beyond "to assign a grade"? By understanding and being able to communicate your purposes for testing, you are better framing your assessment practices within your teaching.
2. **Ensure that the content of your summative assessments is aligned with your learning objectives for that unit.** This might seem obvious, but you also might be surprised when you examine your objectives and assessments. It's easy to get sidetracked by important concepts that are outside of your stated objectives. Perform this alignment check frequently. As we tweak objectives and assessments (often separately), things can get out of whack. Of course, this assumes that you have well-written and appropriate learning objectives in place.
3. **Ensure that your students are interpreting your assessment items in the way that you meant for them to be interpreted.** If you write an item intended to test a students's ability to apply Newton's Second Law, can you be sure that performance on that item is indicative of that construct, and not the student's ability to recall a memorized algorithm? You can investigate this by simply asking students to describe how they solved the problem, either in separate, think-aloud settings with a few students, or as an open-response prompt following the test item.

4. **Ensure that the test is fair for all of your students.** Do you use cultural contexts in your test items that may not be familiar to some of your students? For example, we often use sports as a context for physics test items, but many students are not familiar with baseball. Further, are some groups of students (e.g., females, English-language learners, students of color) systematically responding to an item or set of items in a different way than students of the same ability from another group? If so, your test may be biased and therefore not fair. One way to investigate this is to simply disaggregate test item performance by subgroup.
5. **Be able to relate your students' test scores to a meaningful, qualitative characterization of ability or understanding.** This is much easier said than done. But you should be able to discuss and defend what a score of 85/100 means with respect to meeting the objectives tested by the assessment. And if you set some cut score (e.g., 65% for passing), be able to defend why that cut score was chosen. This is, in many ways, the most difficult part of educational measurement. Translating scores into interpretable locations on a continuum of understanding is no small task. However, not attempting to do so makes score meaningless. But meaningless scores still have consequences for the student (and teacher), so due attention must be given here.

Q.2 Define a scoring criteraia for essay type test items for 8th grade?

Multiple-choice questions

Advantages

- Quick and easy to score, by hand or electronically
- Can be written so that they test a wide range of higher-order thinking skills
- Can cover lots of content areas on a single exam and still be answered in a class period

Disadvantages

- Often test literacy skills: "if the student reads the question carefully, the answer is easy to recognize even if the student knows little about the subject" (p. 194)
- Provide unprepared students the opportunity to guess, and with guesses that are right, they get credit for things they don't know
- Expose students to misinformation that can influence subsequent thinking about the content
- Take time and skill to construct (especially good questions)

True-false questions

Advantages

- Quick and easy to score

Disadvantages

- Considered to be "one of the most unreliable forms of assessment" (p. 195)
- Often written so that most of the statement is true save one small, often trivial bit of information that then makes the whole statement untrue
- Encourage guessing, and reward for correct guesses

Short-answer questions

Advantages

- Quick and easy to grade
- Quick and easy to write

Disadvantages

- Encourage students to memorize terms and details, so that their understanding of the content remains superficial

Essay questions

Advantages

- Offer students an opportunity to demonstrate knowledge, skills, and abilities in a variety of ways
- Can be used to develop student writing skills, particularly the ability to formulate arguments supported with reasoning and evidence

Disadvantages

- Require extensive time to grade
- Encourage use of subjective criteria when assessing answers
- If used in class, necessitate quick composition without time for planning or revision, which can result in poor-quality writing

Questions provided by test banks

Advantages

- Save instructors the time and energy involved in writing test questions
- Use the terms and methods that are used in the book

Disadvantages

- Rarely involve analysis, synthesis, application, or evaluation (cross-discipline research documents that approximately 85 percent of the questions in test banks test recall)
- Limit the scope of the exam to text content; if used extensively, may lead students to conclude that the material covered in class is unimportant and irrelevant

We tend to think that these are the only test question options, but there are some interesting variations. The article that promoted this review proposes one: Start with a question, and revise it until it can be answered with one word or a short phrase. Do not list any answer options for that single question, but attach to the exam an alphabetized list of answers. Students select answers from that list. Some of the answers provided may be used more than once, some may not be used, and there are more answers listed than questions. It's a ratcheted-up version of matching. The approach makes the test more challenging and decreases the chance of getting an answer correct by guessing.

Multiple choice questions (MCQ) is a significant form of **Objective Assessment.** In this tool of evaluating the learners are asked to choose only correct answer out of the choices from the list. The concept of MCQs was

developed by Benjamin D. Wood and its popularity increased in mid 20th century when scanners and data processing concept were developed to check the results. Some of the major features of MCQs test are as follows:

- MCQs notably lower marking time and analysis of individual question is more feasible.
- They have high reliability, validity and manageability.
- They are suitable for use in many different subject matter areas and can be used to measure a great variety of educational objectives.
- MCQs are flexible to various levels of learning outcomes from simple recall of content to more complex levels such as student ability to examine facts, understanding concepts and principles.
- MCQs test extent of understanding across a much wider range of content.

A true and false test consists of a statement that requires a true or false response. These test are factual based rather than opinion-oriented, and are designed to quickly and efficiently test learner knowledge about a particular idea or concept.

Q.3 Write a note on Mean, Median and Mode. Also dicsuss their importance in interpreting test scores.

Mean

The mean of a data set is also known as the average value. It is calculated by dividing the sum of all values in a data set by the number of values.

So in a data set of 1, 2, 3, 4, 5, we would calculate the mean by adding the values (1+2+3+4+5) and dividing by the total number of values (5). Our mean then is 15/5, which equals 3.

Disadvantages to the mean as a measure of central tendency are that it is highly susceptible to outliers (observations which are markedly distant from the bulk of observations in a data set), and that it is not appropriate to use when the data is skewed, rather than being of a normal distribution.

Median

The median of a data set is the value that is at the middle of a data set arranged from smallest to largest.

In the data set 1, 2, 3, 4, 5, the median is 3.

In a data set with an even number of observations, the median is calculated by dividing the sum of the two middle values by two. So in: 1, 2, 3, 4, 5, 6, the median is (3+4)/2, which equals 3.5.

The median is appropriate to use with ordinal variables, and with interval variables with a skewed distribution.

Mode

The mode is the most common observation of a data set, or the value in the data set that occurs most frequently.

The mode has several disadvantages. It is possible for two modes to appear in the one data set (e.g. in: 1, 2, 2, 3, 4, 5, 5, both 2 and 5 are the modes).

The mode is an appropriate measure to use with categorical data.

Test percentile scores are just one type of test score you will find on your child's testing reports from school. Percentile scores are almost always reported on major achievement tests that are taken by your child's entire class. These scores will also be found on individual diagnostic test reports.

Test percentile scores are important for making decisions about your child's education, especially when considering a special education program. Understanding these scores can help you gain a clearer picture of your child's abilities and help you spot areas where they may need extra assistance. In some cases, specific scores on an exam may be required in order to receive specialized assistance or to gain admission to certain programs.

Percentile Rank Scores vs. Percentage Scores

It is important to understand how a percentile rank score differs from a percentage score. The two terms seem similar, but they have very different meanings.

Percentage Scores

Most parents and students are familiar with percentage scores. These are the results you remember getting when you took a test in school.

Percentile scores on teacher-made tests and homework assignments are developed by dividing the student's raw score on their work by the total number of points possible. So, for example, if they got 8 points out of a possible 10, their percentile score would be 0.8, or 80 percent.

Such scores are an indicator of how well a student performed on a particular assignment or test. However, they do not provide information about how the student compares to others in their peer group.

Percentile Ranks

Percentile rank scores, on the other hand, allow for comparing students to their peer group.[1] These scores are often used on what are known as **norm-referenced tests**. Such tests allow parents and educators to compare an individual child's score to the scores of other children in the same age group.

Unlike the percentage scores, percentile ranks are not an indication of how many questions your child answered correctly, or what your child does or does not know. Instead, the scores indicate how well your child did relative to other students who have also taken the test (i.e., how his skill level compares to that of their peers).

Percentile rank scores on norm-referenced or standardized tests are calculated differently than percentage scores, and the calculations are typically included in test manuals or calculated with scoring software.[2]

Percentile ranks are often expressed as a number between 1 and 99, with 50 being the average. So if a student scored a percentile rank of 87, it would mean that they performed better than 87 percent of the other students in his norm group.

Examples of Percentile Rank Scores

It can be helpful to look at how these percentile scores are sometimes used on educational assessments.

- On many tests that are nationally norm-referenced intelligence tests, a standard score of 100 is equal to the 50th percentile. Students scoring at this level on the test are well within the average range.

- The SAT is an example of a standardized test that provides a score percentile. Often used as part of the college admissions process, a score of 1200 or higher (or the 75th percentile) is considered a good score. This number indicates that 75 percent of students scored at or below 1200, while 25 percent of students scored above 1200.[3]
- If you take a cognitive abilities test and score in the 85th percentile, it would indicate that your score is better than 85 percent of people who also took the same test.

How Percentile Rank Scores Are Used

Several other types of standard scores may also appear on test reports. A single test may provide percentile rank scores for different domains such as reading comprehension, verbal ability, and reasoning as well as an aggregate score.

These scores are often used for assessment purposes and may be utilized to make educational decisions. Low percentile scores, for example, may indicate that a child needs specialized assistance in a particular area.

Such tests can help educators spot specific needs that should be addressed and make early intervention possible. Percentile ranks may also be used to determine if a child qualifies for specialized assistance or admission to a specific educational program.

Q.4 Write the procedure of arising letter grades to test score.

Few issues have created more controversy among educators than those associated with grading and reporting student learning. Despite the many debates and multitudes of studies, however, prescriptions for best practice remain elusive. Although teachers generally try to develop grading policies that are honest and fair, strong evidence shows that their practices vary widely, even among those who teach at the same grade level within the same school.

In essence, grading is an exercise in professional judgment on the part of teachers. It involves the collection and evaluation of evidence on students' achievement or performance over a specified period of time, such as nine weeks, an academic semester, or entire school year. Through this process, various types of descriptive information and measures of students' performance are converted into grades or marks that summarize students' accomplishments. Although some educators distinguish between grades and marks, most consider these terms synonymous. Both imply a set of symbols, words, or numbers that are used to designate different levels of achievement or performance. They might be letter grades such as A, B, C, D, and F; symbols such as &NA;+, &NA;, and &NA;−; descriptive words such as Exemplary, Satisfactory, and Needs Improvement; or numerals such as 4, 3, 2, and 1. Reporting is the process by which these judgments are communicated to parents, students, or others.

Grading and reporting are relatively recent phenomena in education. In fact, prior to 1850, grading and reporting were virtually unknown in schools in the United States. Throughout much of the nineteenth century most schools grouped students of all ages and backgrounds together with one teacher in one-room

schoolhouses, and few students went beyond elementary studies. The teacher reported students' learning progress orally to parents, usually during visits to students' homes.

As the number of students increased in the late 1800s, schools began to group students in grade levels according to their age, and new ideas about curriculum and teaching methods were tried. One of these new ideas was the use of formal progress evaluations of students' work, in which teachers wrote down the skills each student had mastered and those on which additional work was needed. This was done primarily for the students' benefit, since they were not permitted to move on to the next level until they demonstrated their mastery of the current one. It was also the earliest example of a narrative report card.

With the passage of compulsory attendance laws at the elementary level during the late nineteenth and early twentieth centuries, the number of students entering high schools increased rapidly. Between 1870 and 1910 the number of public high schools in the United States increased from 500 to 10,000. As a result, subject area instruction in high schools became increasingly specific and student populations became more diverse. While elementary teachers continued to use written descriptions and narrative reports to document student learning, high school teachers began using percentages and other similar markings to certify students' accomplishments in different subject areas. This was the beginning of the grading and reporting systems that exist today.

The shift to percentage grading was gradual, and few American educators questioned it. The practice seemed a natural by-product of the increased demands on high school teachers, who now faced classrooms with growing numbers of students. But in 1912 a study by two Wisconsin researchers seriously challenged the reliability of percentage grades as accurate indicators of students' achievement.

In their study, Daniel Starch and Edward Charles Elliott showed that high school English teachers in different schools assigned widely varied percentage grades to two identical papers from students. For the first paper the scores ranged from 64 to 98, and the second from 50 to 97. Some teachers focused on elements of grammar and style, neatness, spelling, and punctuation, while others considered only how well the message of the paper was communicated. The following year Starch and Elliot repeated their study using geometry papers submitted to math teachers and found even greater variation in math grades. Scores on one of the math papers ranged from 28 to 95–a 67-point difference. While some teachers deducted points only for a wrong answer, many others took neatness, form, and spelling into consideration.

These demonstrations of wide variation in grading practices led to a gradual move away from percentage scores to scales that had fewer and larger categories. One was a three-point scale that employed the categories of Excellent, Average, and Poor. Another was the familiar five-point scale of Excellent, Good, Average, Poor, and Failing, (or A, B, C, D, and F). This reduction in the number of score categories served to reduce the variation in grades, but it did not solve the problem of teacher subjectivity.

To ensure a fairer distribution of grades among teachers and to bring into check the subjective nature of scoring, the idea of grading based on the normal probability, bell-shaped curve became increasingly popular. By this method, students were simply rank-ordered according to some measure of their performance or proficiency. A

top percentage was then assigned a grade of A, the next percentage a grade of B, and so on. Some advocates of this method even specified the precise percentages of students that should be assigned each grade, such as the 6-22-44-22-6 system.

Grading on the curve was considered appropriate at that time because it was well known that the distribution of students' intelligence test scores approximated a normal probability curve. Since innate intelligence and school achievement were thought to be directly related, such a procedure seemed both fair and equitable. Grading on the curve also relieved teachers of the difficult task of having to identify specific learning criteria. Fortunately, most educators of the early twenty-first century have a better understanding of the flawed premises behind this practice and of its many negative consequences.

In the years that followed, the debate over grading and reporting intensified. A number of schools abolished formal grades altogether, believing they were a distraction in teaching and learning. Some schools returned to using only verbal descriptions and narrative reports of student achievement. Others advocated pass/fail systems that distinguished only between acceptable and failing work. Still others advocated a mastery approach, in which the only important factor was whether or not the student had mastered the content or skill being taught. Once mastered, that student would move on to other areas of study.

At the beginning of the twenty-first century, lack of consensus about what works best has led to wide variation in teachers' grading and reporting practices, especially among those at the elementary level. Many elementary teachers continue to use traditional letter grades and record a single grade on the reporting form for each subject area studied. Others use numbers or descriptive categories as proxies for letter grades. They might, for example, record a 1, 2, 3, or 4, or they might describe students' achievement as Beginning, Developing, Proficient, or Distinguished. Some elementary schools have developed standards-based reporting forms that record students' learning progress on specific skills or learning goals. Most of these forms also include sections for teachers to evaluate students' work habits or behaviors, and many provide space for narrative comments.

Grading practices are generally more consistent and much more traditional at the secondary level, where letter grades still dominate reporting systems. Some schools attempt to enhance the discriminatory function of letter grades by adding plusses or minuses, or by pairing letter grades with percentage indicators. Because most secondary reporting forms allow only a single grade to be assigned for each course or subject area, however, most teachers combine a variety of diverse factors into that single symbol. In some secondary schools, teachers have begun to assign multiple grades for each course in order to separate achievement grades from marks related to learning skills, work habits, or effort, but such practices are not widespread.

Q.5 Discuss the difference between measures of central tendency and measures of reliability.

Measures of Central Tendency provide a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution. There are three main measures of central tendency: the mean, the median and the mode.

Mean

The mean of a data set is also known as the average value. It is calculated by dividing the sum of all values in a data set by the number of values.

So in a data set of 1, 2, 3, 4, 5, we would calculate the mean by adding the values (1+2+3+4+5) and dividing by the total number of values (5). Our mean then is 15/5, which equals 3.

Disadvantages to the mean as a measure of central tendency are that it is highly susceptible to outliers (observations which are markedly distant from the bulk of observations in a data set), and that it is not appropriate to use when the data is skewed, rather than being of a normal distribution.

Median

The median of a data set is the value that is at the middle of a data set arranged from smallest to largest.

In the data set 1, 2, 3, 4, 5, the median is 3.

In a data set with an even number of observations, the median is calculated by dividing the sum of the two middle values by two. So in: 1, 2, 3, 4, 5, 6, the median is (3+4)/2, which equals 3.5.

The median is appropriate to use with ordinal variables, and with interval variables with a skewed distribution.

Mode

The mode is the most common observation of a data set, or the value in the data set that occurs most frequently.

The mode has several disadvantages. It is possible for two modes to appear in the one data set (e.g. in: 1, 2, 2, 3, 4, 5, 5, both 2 and 5 are the modes).

The mode is an appropriate measure to use with categorical data.

a measure of the amount of measurement error associated with a test score.

- Ranges from 0.00 to 1.00
- The higher the value, the more reliable the test score
- Typically, a measure of internal consistency, indicating how well items are correlated with one another
- High reliability indicates that items are measuring the same construct (e.g., knowledge of how to calculate integrals)
- Two ways to improve test reliability: 1) increase the number of items or 2) use items with high discrimination values

Reliability Interpretation

- .90 and above Excellent reliability; at the level of the best standardized tests
- .80 - .90 Very good for a classroom test
- .70 - .80 Good for a classroom test; in the range of most. There are probably a few items that could be improved.

- .60 - .70 Somewhat low. This test should be supplemented by other measures to determine grades. There are probably some items that could be improved.
- .50 - .60 Suggests need to revise the test, unless it is quite short (ten or fewer items). The test must be supplemented by other measures for grading.
- .50 or below Questionable reliability. This test should not contribute heavily to the course grade, and it needs revision.

Distractor Evaluation

Another useful item review technique is distractor evaluation.

You should consider each distractor an important part of an item in view of nearly 50 years of research that shows that there is a relationship between the distractors students choose and total test score. The quality of the distractors influences student performance on a test item.

Although correct answers must be truly correct, it is just as important that distractors be clearly incorrect, appealing to low scorers who have not mastered the material rather than to high scorers. You should review all item options to anticipate potential errors of judgment and inadequate performance so you can revise, replace, or remove poor distractors.

One way to study responses to distractors is with a frequency table that tells you the proportion of students who selected a given distractor. Remove or replace distractors selected by a few or no students because students find them to be implausible.

Caution when Interpreting Item Analysis Results

Mehrens and Lehmann (1973) offer three cautions about using the results of item analysis:

- Item analysis data are not synonymous with item validity. An external criterion is required to accurately judge the validity of test items. By using the internal criterion of total test score, item analyses reflect internal consistency of items rather than validity.
- The discrimination index is not always a measure of item quality. There are a variety of reasons why an item may have low discrimination power:

o extremely difficult or easy items will have low ability to discriminate, but such items are often needed to adequately sample course content and objectives.

o an item may show low discrimination if the test measures many content areas and cognitive skills. For example, if the majority of the test measures "knowledge of facts," then an item assessing "ability to apply principles" may have a low correlation with total test score, yet both types of items are needed to measure attainment of course objectives.

- Item analysis data are tentative. Such data are influenced by the type and number of students being tested, instructional procedures employed, and chance errors. If repeated use of items is possible, statistics should be recorded for each administration of each item.

Reliability refers to the consistency of a measure. Psychologists consider three types of consistency: over time (test-retest reliability), across items (internal consistency), and across different researchers (inter-rater reliability).

When researchers measure a construct that they assume to be consistent across time, then the scores they obtain should also be consistent across time. Test-retest reliability is the extent to which this is actually the case. For example, intelligence is generally thought to be consistent across time. A person who is highly intelligent today will be highly intelligent next week. This means that any good measure of intelligence should produce roughly the same scores for this individual next week as it does today. Clearly, a measure that produces highly inconsistent scores over time cannot be a very good measure of a construct that is supposed to be consistent.

Assessing test-retest reliability requires using the measure on a group of people at one time, using it again on the same group of people at a later time, and then looking at test-retest correlation between the two sets of scores. This is typically done by graphing the data in a scatterplot and computing Pearson's r. Figure 5.2 shows the correlation between two sets of scores of several university students on the Rosenberg Self-Esteem Scale, administered two times, a week apart. Pearson's r for these data is +.95. In general, a test-retest correlation of +.80 or greater is considered to indicate good reliability.