Lecture Notes

# Chapter 7: Standardized Measurement and Assessment

## Learning Objectives

* 1. Explain the meaning of measurement.
  2. Explain the different scales of measurement, including the type of information communicated by each one.
  3. Articulate the seven assumptions underlying testing and assessment.
  4. Explain the meaning of reliability.
  5. Explain the characteristics of each of the methods for computing reliability.
  6. Explain the meaning of validity and validity evidence.
  7. Explain the different methods of collecting validity evidence.
  8. Identify the different types of standardized tests and the sources of information on these tests.

## Chapter Summary

This chapter focuses on the basics of measurement and assessment. It begins by defining measurement and describing the different scales of measurement: nominal, ordinal, interval, and ratio. Testing and assessment are differentiated, and the assumptions underlying each are discussed. Qualities of good tests and assessments are discussed in terms of the constructs of reliability and validity. The different forms of reliability and validity that students will encounter with tests and research studies are detailed as are most common types of tests seen in educational settings (e.g., intelligence, personality, and educational assessment tests). Students are also referred to professional collections that review published tests (e.g., Mental Measurements Yearbook, Tests in Print) to learn more about the quality of potential tests they may use or encounter.

## Annotated Chapter Outline

1. Introduction
   1. We have talked about research. In this next series of chapters, we will focus on the foundations of research.
   2. This chapter concentrates on measurement and assessment. It discusses different ways of measuring as well as emphasizing the concepts of reliability and validity. Additionally, common tests seen in education settings are reviewed.
2. Defining Measurement
   1. **Measurement:** assigning symbols or numbers to something according to a specific set of rules.
      1. Measurement can be categorized by the type of information that is communicated by the symbols or numbers assigned to the variables of interest.
      2. When working with variables, researchers must determine how they will measure the variables they are going to use.
      3. Four levels or types or “scales of measurement.”
      4. Notes (examples of types of notes below)
      5. Discussion Question: why are different scales of measurement needed?
   2. **Nominal Scale**: A scale of measurement that uses symbols, such as words or numbers, to label, classify, or identify people or objects
      1. This is a nonquantitative measurement scale.
      2. Numbers can be used to label the categories of a nominal variable but the numbers serve only as markers not as indicators of amount or quantity (e.g., if you wanted to, you could mark the categories of the variable called “gender” with 1 = female and 2 = male).
      3. Cannot add, subtract, rank, or average nominal variable categories.
      4. Can count frequency within each category
      5. Can relate a nominal variable to other variables.
      6. Some examples of nominal-level variables are the country you were born in, college major, personality type, and experimental group (e.g., experimental group or control group)
      7. Discussion Question: Why is it useful to have nominal scales of measurement?
   3. **Ordinal Scale:** a rank-order scale of measurement
      1. This level of measurement enables one to make ordinal judgments (i.e., judgments about rank order).
      2. Used with any variable where the levels can be ranked (but you do not know if the distance between the levels is the same) is an ordinal variable.
      3. Allow you to rank individuals on some characteristic.
      4. Some examples are: order of finish position in a marathon, Billboard Top 40, and rank in class.
      5. Discussion Question: Discuss variables that can only be measured with ordinal scales.
   4. **Interval Scale:** a scale of measurement that has equal intervals of distances between adjacent numbers.
      1. This scale or level of measurement has the characteristics of rank order and equal intervals (i.e., the distance between adjacent points is the same).
      2. It does not possess an absolute zero point.
      3. Some examples are Celsius temperature, Fahrenheit temperature, and IQ scores.
      4. Absence of a true zero point: 0 °C does not mean that there is no temperature at all; in a Fahrenheit scale, it is equal to the freezing point or 32°. Zero degrees in these scales does not mean zero or no temperature.
      5. Can add, subtract, rank, and average interval-scaled scores.
      6. Discussion Question: Why is the ability to use arithmetic operations on interval-scale data important?
   5. **Ratio Scale**: A scale of measurement that has a true zero point.
      1. Highest level of quantitative measurement.
      2. True zero point.
      3. It also has all of the “lower level” characteristics (i.e., the key characteristic of each of the lower level scales) of equal intervals (interval scale), rank order (ordinal scale), and ability to mark a value with a name (nominal scale).
      4. All mathematical operations can be used with ratio scale variables.
      5. Some examples of ratio-level scales are number correct, weight, height, response time, Kelvin temperature, and annual income.
      6. Here is an example of the presence of a true zero point: if your annual income is exactly $0 then you earned no annual income at all. (You can buy absolutely nothing with $0.) Zero means zero.
      7. Discussion Question: compare and contrast the four scales of measurement.
3. Assumptions Underlying Testing and Assessment: testing and assessment are not the same thing and researchers must be aware of the differences
   1. Important definitions:
      1. **Testing:** measurement of variables
      2. **Assessment:** gathering and integrating data to make educational evaluations
      3. **Error:** the difference between true scores and observed scores
      4. **Traits**: distinguishable, relatively enduring ways in which one individual differs from another
      5. **States:** distinguishable but less enduring ways in which individuals vary
   2. All measurement involves some amount of error, but researchers try and minimize error in research and/or measurement.
   3. Table 7.3: Assumptions made by professional test developers and users
      1. Psychological traits and states exist: Traits and states are actually social constructions, but they are real in the sense that they are useful for classifying and organizing the world, they can be used to understand and predict behavior, and they refer to something in the world that we can measure.
      2. Psychological traits and states can be quantified and measured: For nominal scales, the number is used as a marker. For the other scales, the numbers become more and more quantitative as you move from ordinal scales (shows ranking only) to interval scales (shows amount but lacks a true zero point) to ratio scales (shows amount or quantity as we usually understand this concept in mathematics or everyday use of the term). Most traits and states measured in education are taken to be at the interval level of measurement.
      3. A major decision about an individual should not be made on the basis of a single test score but, rather, from a variety of different data sources. For example, different tests of intelligence tap into somewhat different aspects of the construct of intelligence. Information from several sources usually should be obtained in order to make an accurate and informed decision. For example, the idea of portfolio assessment is useful.
      4. Various sources of error are always present in testing and assessment: There is no such thing as perfect measurement. All measurement has some error. The two main types of error are random error (e.g., error due to transient factors such as being sick or tired) and systematic error (e.g., error present every time the measurement instrument is used such as an essay exam being graded by an overly easy grader). (Later when we discuss reliability and validity, you might note that unreliability is due to random error and lack of validity is due to systematic error.)
      5. Test-related attitudes and behavior can be used to predict non-test-related attitudes and behavior: The goal of testing usually is to predict behavior other than the exact behaviors required while the exam is being taken. For example, paper-and-pencil achievement tests given to children are used to say something about their level of achievement. Another paper-and-pencil test (also called a self-report test) that is popular in counseling is the MMPI (i.e., the Minnesota Multiphasic Personality Inventory). Clients’ scores on this test are used as indicators of the presence or absence of various mental disorders. The point here is that the actual mechanics of measurement (e.g., self-reports, behavioral performance, and projective) can vary widely and still provide good measurement of educational, psychological, and other types of variables. Perhaps the most important reason for giving tests is to predict future behavior. Tests provide a sample of present-day behavior. However, this “sample” is used to predict future behavior. For example, an employment test given by someone in a Personnel Office may be used as a predictor of future work behavior. Another example: the Beck Depression Inventory is used to measure depression and, importantly, to predict test taker’s future behavior (e.g., are they a risk to themselves?).
      6. With much work and continual updating, fair and unbiased tests can be developed. This requires careful construction of test items and testing of the items on different types of people. Test makers always have to be on the alert to make sure tests are fair and unbiased. This assumption also requires that the test be administered to those types of people for whom it has been shown to operate properly.
      7. Standardized testing and assessment can benefit society if the tests are developed by expert psychometricians and are properly administered and interpreted by trained professionals. Many critical decisions are made on the basis of tests (e.g., teacher competency, employability, presence of a psychological disorder, degree of teacher satisfactions, degree of student satisfaction, etc.). Without tests, the world would be much more unpredictable.
      8. Discussion Question: Evaluate each of the assumptions that underlie testing.
4. Identifying a Good Test or Assessment Procedure: good measurement is fundamental for research. If we do not have good measurement then we cannot have good research. That is why it is so important to use testing and assessment procedures that are characterized by high reliability and high validity.
   1. Overview of Reliability and Validity
      1. Reliability refers to the consistency or stability of test scores.
      2. Validity refers to the accuracy of the inferences or interpretations we make from test scores.
      3. Two most important psychometric properties to think about with a test or assessment procedure.
      4. **Systematic error:** an error that is present every time an instrument is used.
      5. Reliability is a necessary but not sufficient condition for validity (i.e., if you are going to have validity, you must have reliability but reliability in and of itself is not enough to ensure validity).
      6. Discussion Question: Why are reliability and validity important to measurement?
   2. **Reliability**: the consistency or stability of test scores.
      1. Reliability is usually determined using a correlation coefficient, and this correlational index is called a **reliability coefficient** (a correlation coefficient that is used as an index of reliability) in this context.
      2. Review point from Chapter 2 material: recall that a correlation coefficient is a measure of relationship that varies from –1 to 0 to +1 and the farther the number is from zero, the stronger the correlation. For example, minus one (–1.00) indicates a perfect negative correlation, zero indicates no correlation at all, and positive one (+1.00) indicates a perfect positive correlation. Regarding strength, –.85 is stronger than +.55, and +.75 is stronger than +.35. When you have a negative correlation, the variables move in opposite directions (e.g., poor diet and life expectancy); when you have a positive correlation, the variables move in the same direction (e.g., education and income).
      3. When looking at reliability coefficients, we are interested ONLY in the values ranging from 0 to + 1; that is, we are only interested in positive correlations. The key point here is that negative reliability coefficients mean no reliability, zero means no reliability, and +1.00 means perfect reliability.
      4. Reliability coefficients of .70 or higher are generally considered to be acceptable for research purposes. Reliability coefficients of .90 or higher are needed to make decisions that have an impact on people’s lives (e.g., the educational and clinical uses of tests).
      5. Reliability is empirically determined: we must check the reliability of test scores with specific sets of people. That is, we must obtain the reliability coefficients of interest to us.
      6. Discussion Question: Why do we only focus on positive correlation coefficients for reliability?
      7. **Test–Retest Reliability**: a measure of the consistency of scores over time
         1. It is calculated by correlating the test scores obtained at one point in time with the test scores obtained at a later point in time for a group of people.
         2. A primary issue is identifying the appropriate time interval between the two testing occasions. In too short an interval, reliability may be inflated because test takers remember the answers they gave before. In too long an interval, people change, learn new things, forget things, and develop, so reliability will be lowered.
         3. The longer the time interval between the two testing occasions, the lower the reliability coefficient tends to be.
         4. Discussion Question: Why is test–retest reliability an important type of reliability for tests?
      8. **Equivalent-forms Reliability**: the consistency of a group of individuals’ scores on alternative forms of a test measuring the same thing.
         1. It is measured by correlating the scores obtained by concurrently giving two forms of the same test to a group of people.
         2. The success of this method hinges on the equivalence of the two forms of the test.
         3. Sometimes it is difficult to get people to basically take the same test twice in a short period of time.
         4. Discussion Question: Can you think of situations where equivalent forms reliability is needed? Can you think of situations where equivalent forms reliability is not necessary?
      9. **Internal Consistency Reliability**: the consistency with which the items on a test measure a single construct.
         1. **Homogeneous tests** (a unidimensional test in which all the items measure a single construct. More interitem consistency than heterogeneous tests.
         2. Internal consistency reliability only requires one administration of the test, which makes it a very convenient form of reliability.
         3. One type of internal consistency reliability is **split-half reliability** (a measure of the consistency of the scores obtained from two equivalent halves of the same test), which involves splitting a test into two equivalent halves and checking the consistency of the scores obtained from the two halves. Value varies depending on how the test is divided into halves.
         4. A better measure of internal consistency is coefficient alpha (a formula that provides an estimate of the reliability of a homogeneous test or an estimate of the reliability of each dimension in a multidimensional test. (It is also sometimes called **Cronbach’s alpha**--a frequently used name for what Lee Cronbach called “coefficient alpha.”) The beauty of coefficient alpha is that it is readily provided by statistical analysis packages, and it can be used when test items are quantitative and when they are dichotomous (as in right or wrong).
         5. Researchers use coefficient alpha when they want an estimate of the reliability of a homogeneous test (i.e., a test that measures only one construct or trait) or an estimate of the reliability of each dimension on a multidimensional test. You will see it commonly reported in empirical research articles.
         6. Coefficient alpha will be high (e.g., greater than .70) when the items on a test are correlated with one another. But note that the number of items also affects the strength of coefficient alpha (i.e., the more items you have on a test, the higher coefficient alpha will be). This latter point is important because it shows that it is possible to get a large alpha coefficient even when the items are not very homogeneous or internally consistent.
         7. Discussion Question: Compare and contrast split-half and coefficient alpha methods of assessing internal consistency reliability.
      10. **Interscorer Reliability**: the degree of agreement or consistency between two or more scorers, judges, or raters.
          1. You could have two judges rate one set of papers. Then you would just correlate their two sets of ratings to obtain the interscorer reliability coefficient, showing the consistency of the two judges’ ratings.
          2. Sometimes raters’ scores are not very similar unless training and practice in using the rating instrument have occurred before the ratings are made.
          3. Interscorer reliability is sometimes referred to as “inter-rater reliability” or “interobserver agreement” in test manuals and research studies.
          4. Discuss Question: Describe each type of reliability discussed. Include its uses, its strength, and its weaknesses.
   3. **Validity:** the accuracy of the inferences, interpretations, or actions made on the basis of test scores.
      1. Technically speaking, it is incorrect to say that a test is valid or invalid. It is the interpretations and actions taken based on the test scores that are valid or invalid.
      2. All of the ways of collecting validity evidence are really forms of what used to be called construct validity. All that means is that in testing and assessment, we are always measuring something (e.g., IQ, gender, age, depression, and self-efficacy).
      3. **Validity evidence**: empirical evidence and theoretical rationales that support the inferences or interpretations made from test scores.
      4. **Validation:** the process of gathering evidence that supports inferences made on the basis of test scores.
      5. Look for evidence of validity as an overall construct.
      6. Table 7.6 Summary of Methods for Obtaining Validity Evidence.
         1. Discussion Question: Explain why the author wrote: “complete validation is never fully attained . . . validation therefore should be viewed as a never-ending process”
      7. Evidence Based on Content:
         1. **Content-related evidence**: validity evidence based on a judgment of the degree to which the items, task, or questions on a test adequately represent the construct domain of interest. To make a decision about content-related evidence, you should try to answer these three questions:
            1. Do the items appear to represent the thing you are trying to measure?
            2. Does the set of items underrepresent the construct’s content (i.e., have you excluded any important content areas or topics)?
            3. Do any of the items represent something other than what you are trying to measure (i.e., have you included any irrelevant items)?
      8. Evidence Based on Internal Structure
         1. Some tests are designed to measure one general construct, but other tests are designed to measure several components or dimensions of a construct. For example, the Rosenberg self-esteem scale is a 10-item scale designed to measure the construct of global self-esteem. In contrast, the Harter self-esteem scale is designed to measure global self-esteem as well as several separate dimensions of self-esteem.
         2. The use of the statistical technique called **factor analysis** (a statistical procedure that analyzes correlations among test items and tells you the number of factors present).
         3. When you examine the internal structure of a test, you can also obtain a measure of test **homogeneity** (In test validity, refers to how well the different items on a test measure the same construct or trait).
         4. The two primary indices of homogeneity are the item-to-total correlation (i.e., correlate each item with the total test score) and coefficient alpha (discussed earlier under reliability).
      9. Evidence Based on Relations to Other Variables
         1. This form of evidence is obtained by relating your test scores with one or more relevant criteria. A criterion is the standard or benchmark that you want to predict accurately on the basis of the test scores. Note that when using correlation coefficients for validity evidence we call them validity coefficients.
         2. There are several different kinds of relevant validity evidence based on relations to other variables.
         3. **Criterion-related evidence**: validity evidence based on the extent to which scores from a test can be used to predict or infer performance on some criterion such as a test or future performance.
            1. **Criterion**: the standard or benchmark that you want to predict accurately on the basis of test scores
            2. **Validity coefficient:** a correlation coefficient that is computer to provide validity evidence such as the correlation between test scores and criterion scores.
            3. **Concurrent evidence**--validity evidence based on the relationship between test scores and criterion scores obtained at the same time.
            4. **Predictive evidence**--validity evidence based on the relationship between test scores collected at one point in time and criterion scores obtained at a later time.
            5. **Convergent evidence**--validity evidence based on the relationship between the focal test scores and independent measures of the same construct. The idea is that you want your test (that you are trying to validate) to strongly correlate with other measures of the same thing.
            6. **Discriminant evidence**--evidence that the scores on your focal test are not highly related to the scores from other tests that are designed to measure theoretically different constructs. This kind of evidence shows that your test is not a measure of those other things (i.e., other constructs).
            7. **Known groups evidence**: evidence that groups that are known to differ on the construct do differ on the construct do differ on the test in the hypothesized direction. For example, if you develop a test of gender roles, you would hypothesize that women will score higher on femininity and men will score higher on masculinity. Then you would test this hypothesis to see whether you have evidence of validity.
            8. **Consequential validity:** degree to which the test is used appropriately, works well in practice, and does not have negative or abnormal social and psychological consequences.
         4. Please note that, if you and your students think we have spent a lot of time on validity and measurement, the reason is because validity is so important in empirical research. Remember, without good measurement we end up with GIGO (garbage in, garbage out).
         5. Now, to summarize these three major methods for obtaining evidence of validity, look again at Table 7.6.
         6. Discussion Question: Why are there so many types of validity evidence? Compare and contrast the information received from each source of evidence.
   4. Using Reliability and Validity Information
      1. Participants you are working with should be similar to the people who provided the data for the reliability and validity evidence.
      2. Before using any assessment procedure, look at characteristics of the **norming group** (the specific group for which the test publisher or researcher provides evidence for test validity and reliability.
      3. Look for any direct evidence of reliability and validity that you can find when evaluating others’ research.
      4. Discussion question: What would be potential examples of when reliability and validity information based on a certain norming group would not work for a participant group?
5. **Educational and Psychological Tests:** educational and psychological tests have been develop to develop most situations, characteristics, and types of performance.
   1. Intelligence Tests
      1. Intelligence has many definitions because a single prototype does not exist.
      2. Book’s definition--I**ntelligence**: the ability to think abstractly and learn readily from experience.
      3. Although the construct of intelligence is hard to define, it still has utility because it can be measured and it is related to many other constructs.
      4. Discussion Question: Do students like this definition of intelligence? What do they think intelligence is?
   2. Personality Tests: Personality is a construct similar to intelligence in that a single prototype does not exist.
      1. Personality: the relatively permanent patterns that characterize and can be used to classify individuals.
      2. Most personality tests are **self-report** (a test-taking method in which participants check or rate the degree to which various characteristics are descriptive of themselves) measures.
      3. **Performance** (a test-taking method in which the participants perform some real-life behavior that is observed by the researcher) measures of personality are also used.
      4. Personality has also been measured with **projective measures** (a test-taking method in which the participants provide responses to ambiguous stimuli). The test administrator searches for patterns on participants’ responses. Projective tests tend to be quite difficult to interpret and are not commonly used in quantitative research
      5. Discussion Question: Compare and contrast the different types of personality tests. What are their strengths and weaknesses?
   3. Educational Assessment Tests
      1. Preschool Assessment Tests: These are typically screening tests because the predictive validity of many of these tests is weak.
      2. **Achievement Tests** are tests that are designed to measure the degree of learning that has taken place after a person has been exposed to a specific learning experience. They can be teacher-constructed or standardized tests. These two types of achievement tests differ in terms of psychometric properties.
      3. **Aptitude Tests** focus on information acquired through the informal learning that goes on in life. They are often used to predict future performance whereas achievement tests are used to measure current performance.
      4. **Diagnostic Tests** are tests are used to identify where a student is having difficulty with an academic skill. Do not give information about why the difficulty exists.
      5. Discussion Question: compare and contrast the different types of educational assessment tests.
6. **Sources of Information about Tests**
   1. There are resources available for students to use in evaluating tests in terms of their suitability as measurement tools.
   2. The two most important main sources of information about tests are the **Mental Measurements Yearbook** (MMY, a primary source of information about published tests) and Tests in Print (TIP, a comprehensive primary source of information about published tests).
   3. Some additional sources are provided in Table 7.7.
   4. Discussion Question: Why is it important for researchers to access test information and reviews?