

Chapter 5 Answers to Study Questions

5.1. What is measurement?

Measurement is the act of measuring by assigning symbols or numbers to something according to a specific set of rules. It involves identifying the dimensions, quantity, capacity, type, or degree of something.

5.2. What are the four different levels or scales of measurement and what are the essential characteristics of each one?

The four levels of measurement are nominal, ordinal, interval, and ratio scales. Note that the first letters spell NOIR (which means black in French).

- The most basic level of measurement is the nominal level which simply involves assigning symbols or names to identify the groups or categories of something (e.g., gender and college major are nominal variables).
- The next level of measurement is the ordinal level in which the levels take on the new property of rank order (e.g., students' ranks on an exam, 1st, 2nd, 3rd, etc. is an ordinal variable).
- The next level of measurement is the interval level which takes on the new property that the distances between adjacent points is the same (in addition to having the property of rank ordering). An example is the Fahrenheit temperature scale, where the difference between 70 and 75 degrees is the same as the difference between 75 and 80 degrees. Note however that you cannot say that 80 degrees is twice as hot as 40 degrees because the zero point on an interval scale is arbitrary.
- The highest level of measurement is the ratio scale which has the properties of rank order and equal distances and it has the new property of having an absolute or true zero point. You have a true zero point when zero means none of the property being measured. Annual income and height are examples. Note now that a person who is six feet tall is twice as tall as a person who is three feet tall. Unlike with interval scales, we can make these types of ratio statements with ratio scales (e.g., $50/25=2$).

5.3. What are the twelve assumptions underlying testing and measurement?

Note that it takes a lot of hard work to make the 12 assumptions happen in practice. I will list the 12 assumptions here:

1. Psychological traits and states exist (they are social constructions that name phenomena of interest to researchers as they attempt to understand the world)
2. Psychological traits and states can be quantified and measured
3. Various approaches to measuring aspects of the same thing can be useful
4. Assessment can provide answers to some of life's most momentous questions
5. Assessment can pinpoint phenomena that require further attention or study
6. Various sources of data enrich and are part of the assessment process
7. Various sources of error are always part of the assessment process
8. Tests and other measurement techniques have strengths and weaknesses
9. Test-related behavior predicts non-test-related behavior
10. Present-day behavior sampling predicts future behavior
11. Testing and assessment can be conducted in a fair and unbiased manner

12. Testing and assessment benefit society.

Also be sure to know the three definitions included in this section *traits* (distinguishable, relatively enduring ways in which one individual differs from another), *states*, (less enduring ways in which individuals vary), and *error* (the difference between a person's true score and the person's observed score).

5.4. What is the difference between reliability and validity? Which is more important?

Reliability refers to the consistency or stability of the test scores; validity refers to the accuracy of the inferences or interpretations you make from the test scores. Both of these characteristics are important. Note also that reliability is a necessary but not sufficient condition for validity (i.e., you can have reliability without validity, but in order to obtain validity you must have reliability).

5.5. What are the definitions of reliability and reliability coefficient?

Reliability refers to the consistency or stability of a set of test scores. The reliability coefficient is a correlation coefficient that is used as an index of reliability.

Note that there are several different forms of reliability. First is test-retest reliability (the consistency of a group of individuals' scores over time). The second type is equivalent-forms reliability (consistency of a group of individuals' scores on two equivalent forms of a test). The third type or reliability is internal consistency reliability (consistency of items in measuring a single construct). The two subtypes of internal consistency are split-half reliability and coefficient alpha. The fourth major type of reliability is inter-scorer reliability (consistency or degree of agreement between two or more scorers, judges, or raters).

5.6. What are the different ways of assessing reliability?

Most of the types of reliability are assessed with simple correlation coefficients (called reliability coefficients). Test-retest reliability is the correlation between a group's scores on the same test given at two different times (i.e., give a set of people a test twice and see if the two sets of scores are correlated). Equivalent-forms reliability is the correlation between a group's scores on two forms of the same test (i.e., give everyone in a group two forms of the same test and correlate those two sets of scores). Split-half reliability is the correlation between a group's scores on two halves of the same test (everyone in the group takes the test once and you give everyone a score on both of the two halves of the test; then you correlate those two sets of scores). Coefficient alpha can be viewed as the average of the correlations of all of the items on a test with each other (e.g., if a test only had 3 items it would be the average of the correlation between items 1 and 2, 1 and 3, and 2 and 3). It tells you if the items tend to be related. The basic inter-scorer reliability is the correlation between two raters' ratings of a set of objects (e.g., a set of essay questions).

5.7. Under what conditions should each of the different ways of assessing reliability be used?

Test-retest is used to determine consistency of the scores on a test over time.

Equivalent forms reliability is used to see if different forms of a test give consistent results.

Internal consistency reliability is used to see if the different items on a test give consistent results.

Inter-scorer reliability is used to see if two raters of a set of items give consistent results.

5.8. What are the definitions of validity and validation?

Validity is the accuracy of the inferences, interpretations, or actions made on the basis of test scores. Validation is the process of gathering evidence that supports the inferences made on the basis of test scores.

5.9. What is meant by the unified view of validity?

It means that all validity can be viewed as part of construct validity. That's because to be discussing measurement validity, there has to be something that we intend to measure. The term "construct" simply refers to what we want to measure whether it be age, gender, IQ, knowledge.

5.10. What are the characteristics of the different ways of obtaining validity evidence?

The three major types of evidence include:

- (1) Evidence based on content.
- (2) Evidence based on internal structure of the test.
- (3) Evidence based on relations to other variables.

This is summarized in Table 5.6.

■ **TABLE 5.6** Summary of Methods for Obtaining Validity Evidence

Type of Evidence	Procedures
Evidence based on content	Study the construct to measure, examine the test content, and make a decision whether the test content adequately represents the construct. This is usually done by experts.
Evidence based on internal structure	First, determine how many dimensions or constructs the test measures using the technique called <i>factor analysis</i> . Second, examine the <i>homogeneity</i> of the items (for the whole test and for each of the subscales if the test measures more than one dimension). You can do this by calculating the <i>item-to-total correlation</i> (for a test measuring only one dimension) and by calculating <i>coefficient alpha</i> for the test and for each subscale for a test measuring more than one dimension.
Evidence based on relations to other variables	Relate the test scores to a known criterion by collecting <i>concurrent</i> and/or <i>predictive</i> evidence. Correlate the test scores with measures of the same construct and measures of different constructs to obtain <i>convergent</i> and <i>discriminant</i> evidence. Determine whether groups differ on the test in the way that would be expected (e.g., for a liberalism scale, determine whether Republican Party members differ from Democratic Party members).

5.11. What are the purposes and key characteristics of the major types of tests discussed in your textbook?

The major types of tests discussed are:

- Intelligence tests (goal is to measure one or more types of intelligence)
- Personality tests (goal is to measure one or more dimensions of personality)
- Educational assessment tests (including preschool assessment tests for identifying “at risk” children, achievement tests for measuring learning from formal learning experiences, aptitude tests for measuring informal learning that goes on in life, and diagnostic tests for identifying academic difficulties in students).

5.12. What is a good example of each of the major types of tests that are discussed in this chapter?

- Some examples intelligence tests are the Stanford-Binet Intelligence Test, the Wechsler Adult Intelligence Scale, the Slosson Intelligence Test.
- Some examples of personality tests are the Minnesota Multiphasic Personality Inventory, the California Psychological Inventory, Work Values Inventory, Minnesota School Attitude Survey, and the Thematic Apperception Test.
- Some examples of educational assessment tests are Peabody Individual Achievement Test, Nelson Reading Skills Tests, and the Basic English Skills Test.

For more information on these tests and for more tests go to the companion website under bonus materials for Chapter 5.