

8

Collecting Dependable Data

*Validity
Reliability
Objectivity*

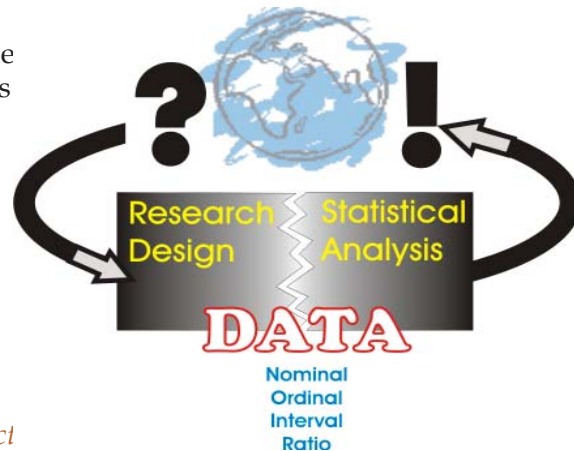
We have discussed variables and problems, hypotheses and purposes, populations and samples. The theoretical foundation of your study must sooner or later yield to concrete action: *the collection of real pieces of data*. The tools used to collect data are called *instruments*. An instrument may be an observation checklist, a questionnaire, an interview guide, a test or attitude scale. It may be a video camera or cassette recorder. An instrument is any device used to observe and record the characteristics of a variable.

Before you can accurately measure the stated variables of your study, you must translate those variables into measurable forms. This is done by *operationally defining the variables of your study* (Chapter 3). Data collection is meaningless without a clearly operationalized set of variables.

The second step is to insure that the *select instrument accurately measures the variables* you've selected.

The naive researcher rushes past the instrument selection or development phase in order to collect data. The result is faulty, error-filled data -- which yields faulty conclusions. The accuracy of the instrument used in your study is an important factor in the usefulness of your results. If the data is incomplete or inadequate, the study is destined for failure. A wonderful design and precise analysis yields useless results if the data quality is poor. So carefully design or select the instrument you will use to collect data. Three characteristics -- "the Great Triad" -- determine the precision with which an instrument collects data.

The Great Triad consists of (1) *validity*, "Does the instrument measure what it says it measures?"; (2) *reliability*, "Does the instrument measure accurately and consistently?"; and (3) *objectivity*, "Is the instrument immune to the personal attitudes and opinions of the researcher?"



Validity

The term validity refers to the ability of research instruments to measure what they say they measure. *A valid instrument measures what it purports to measure.* A

Content
Predictive
Concurrent
Construct

12-inch ruler is a valid instrument for measuring length. It is not a valid instrument for measuring I.Q., or a quantity of a liquid, or an amount of steam pressure. These require an I.Q. test, a measuring cup, and a pressure gauge.

Let's say a student wants to measure the variable "spiritual maturity," and operationally defines it as "the number of times a subject attended Sunday School out of the past 52 Sundays." The question we should ask is whether "attendance count" in Sunday School is a **valid measure** of spiritual maturity – does "count" really measure "spiritual maturity"?

Can one attend Sunday School and be spiritually immature? (Yes, for coffee, fellowship and business contacts).

Can one be spiritually mature and not attend Sunday School? (Yes, pastors usually use this time for pastoral work). If either of these questions can be answered yes, (and they are), then the measure is not a valid one.

There are four kinds of instrument validity: **content, concurrent, predictive, and construct**. Each of these have specific meaning, and helps establish the nature of valid instruments.

Content Validity

The content validity of a research instrument represents the **extent to which the items in the instrument match the behavior, skill, or effect the researcher intends them to measure**.¹ In other words, a test has content validity if the items actually measure mastery of the content for which the test was developed. Tests which ask questions over material not covered by objectives or study guidelines, or draw from other fields besides the one being tested, violate this kind of validity. Content validity is different from *face validity*, which is a subjective judgement that a test *appears to be valid*.

Researchers establish content validity for their instruments by submitting a long list of items (such as statements or questions) to a "validation panel." Such a validation panel consists of six to ten persons who are considered experts in the field of study for which the instrument is being developed. The panel judges the clarity and meaningfulness of each of the items by means of a 4- or 6-point rating scale. Compute the means and standard deviations (see Chapter 16) for each of the items. Select the items with the highest mean and lowest standard deviation on "meaningfulness" and "clarity" to be included in your instrument. ***In summary, content validity asks the question, "How closely does the instrument reflect the material over which it gathers data?"*** Content validity is especially important in **achievement testing**.

Predictive Validity

The predictive validity of a research instrument represents the **extent to which the test's results predict such things as later achievement or job success**. It is the degree to which the predictions made by a test are confirmed by the later success of the subjects.

Suppose I developed a "Research and Statistics Aptitude Test" to be given students at the beginning of the semester. If I correlated these test scores of incoming students with their final grade in the course, I could use the test as a predictor of success in the course. In this example, the Research and Statistics Test provides the predictor measures and the final course grade is the criterion by which the aptitude test is analyzed for validity. In predictive validity, the criterion scores are gathered some time after the predictor scores. The Graduate Record Examination (GRE) is taken

¹Merriam, p. 140

by college students and supposedly predicts which of its users will succeed in (future) doctoral level studies. *Predictive validity asks the question, "How closely does the instrument reflect the later performance it seeks to predict?"*

Concurrent Validity

Concurrent validity represents the extent to which a (usually smaller, easier, newer) test reflects the same results of a (usually larger, more difficult, established) test. The established test is the criterion, the benchmark, for the newer, more efficient test. Strong concurrent validity means that the smaller, easier test provides data as well as the larger, more difficult one. A popular personality test, called the Minnesota Multi-Phasic Inventory (MMPI), once had only one form consisting of about 550 questions. The test required several hours to administer. In order to reduce client frustration, a newer short-form version was developed which contained about 350 questions. Analysis revealed that the shorter form had high concurrent validity with the longer form. That is, psychologists found the same results with the shorter form as with the long form, but also reduced patient frustration and administration time.

A researcher wanted to determine whether anxious college students showed more preference for female role behaviors than less anxious students. To identify contrasting groups of anxious and non-anxious students, she could have had a large number of students evaluated for clinical signs of anxiety by experienced clinical psychologists. However, she was able to locate a quick, objective test, the Taylor Manifest Anxiety Scale, which has been demonstrated to have high concurrent validity with clinical ratings of anxiety in a college population. She saved considerable time conducting the research project by substituting this quick, objective measure for a procedure that is time-consuming and subject to personal error.¹ *Concurrent validity asks the question, "How closely does this instrument reflect the criterion established by another (usually more complex or costly) validated instrument?"*

Construct Validity

Construct validity reflects the extent to which a research instrument measures some abstract or hypothetical construct.² Psychological concepts, such as intelligence, anxiety, and creativity are considered hypothetical constructs because they are not directly observable -- they are *inferred* on the basis of their observable effects on behavior.³ In order to gather evidence on construct validity, the test developer often starts by setting up hypotheses about the differentiating characteristics of persons who obtain high and low scores on the measure.

Suppose, for example, that a test developer publishes a test that he claims is a measure of anxiety. How can one determine whether the test does in fact measure the construct of anxiety? One approach might be to determine whether the test differentiates between psychiatric and normal groups, since theorists have hypothesized that anxiety plays a substantial role in psychopathology. If the test does in fact differentiate the two groups, then we have some evidence that it measures the construct of anxiety.⁴ *Construct validity asks the question, "How closely does this instrument reflect*

¹Borg and Gall, 279

²A construct is a theoretical explanation of an attribute or characteristic created by scholars for purposes of study. Merriam, p. 141

³Borg and Gall, 280

⁴*Ibid.*

⁵Sax, 206

⁶Ary et. al., 200

⁷David Payne, *The Assessment of Learning: Cognitive and Affective* (Lexington, Mass.: D.C. Heath and Company, 1974), 259

the abstract hypothetical construct it seeks to measure?"

Reliability

Stability
Consistency
Equivallance

Reliability is the “extent to which measurements reflect true individual differences among examinees.”⁵ It is the “degree of consistency with which [an instrument] measures what it is measuring.”⁶ The higher the reliability of an instrument, the less influenced it is by random, unsystematic factors.⁷

In other words, is an instrument confounded by the “smoke” and “noise” of human characteristics, or can it measure the true substance of those variables? Does the instrument **measure accurately**, or is there extraneous error in the measurements? Do the scores produced by a test **remain stable over time**, or do we get a different score every time we administer the test to the same sample?

There are three important measures of reliability. These are the coefficients of **stability**, **internal consistency**, and **equivalence**. All three use a correlation coefficient to express the strength of the measure. We will study the correlation coefficient in detail when we get to chapter 22. For the time being, we will merely state that a reliability coefficient can vary from 0.00 (no reliability) to +1.00 (perfect reliability, which is never attained). A coefficient of **0.80 or higher is considered very good**.

Coefficient of Stability

The coefficient of stability, also called **test-retest reliability**,⁹ measures how **consistent scores remain over time**. The test is given once, and then given to the same group at a later time, usually several weeks. A correlation coefficient is computed between the two sets of scores to produce the stability coefficient.

The greatest problem with this measure of reliability is determining how much delay to use between the tests. If the delay is too short, then subjects will remember their previous answers and the reliability coefficient will be higher than it should be. If the delay is too long, then subjects may actually change in the interval. They will answer differently, but the difference is due to a change in the subject, not in the test. This will yield a coefficient lower than it should be.¹⁰ Still, **science does best with consistent, stable, repeatable phenomena, and the stability of responses to a test is a good indicator of the stability of the variable being measured**.

Coefficient of Internal Consistency

The purpose of a test is to measure, honestly and precisely, variables resident in subjects. The structure of the test itself can sometimes reduce the reliability of scores it produces. The coefficient of internal consistency¹¹ measures consistency within a given test.

The coefficient of internal consistency has two major forms. The first is the **split-half test**. After a test has been administered to a single group of subjects, the items are divided into two parts. Odd items (1, 3, 5...) are placed in one part and even items (2, 4, 6...) in the other. Total scores of the two parts are correlated to produce a measure of item consistency. Since reliability is related to the length of the test, and the split-half coefficient reduces test length by half, a correction factor is required in order to obtain the reliability of the entire test.

The **Spearman-Brown prophecy formula** (formula at right) is used to make this correction.¹² Here the r' is the corrected reliability coefficient and r equals the com-

$$r' = \frac{2r}{1+r}$$

⁹Borg and Gall, 283

¹⁰Ibid., 284

¹¹Ibid., 284-5

¹²Ibid., 285

puted correlation between the two halves. If $r=0.60$, then the formula yields $r' = 0.75$.

Another measure of internal consistency can be obtained by the use of the **Kuder-Richardson formulas**. The most popular of the formulas are known as K-R 20 and K-R 21. The K-R 20 formula is considered by many specialists in education and psychology to be the most satisfactory method for determining test reliability. The K-R 21 formula is a simplified approximation of the K-R 20, and provides an easy method for determining a reliability coefficient. It requires much less time to apply than K-R 20 and is appropriate for the analysis of teacher-made tests and experimental tests written by a researcher which are scored dichotomously.¹³ (A dichotomous variable is one which has two and only two responses: yes-no, true-false, on-off).

Cronbach's Coefficient Alpha is a general form of the K-R 20 and can be applied to multiple choice and essay exams. Coefficient Alpha compares the sum of the variances for each item with the total variance for all items taken together. If there is high internal consistency, coefficient alpha produces a strong positive correlation coefficient.

Coefficient of Equivalence

A third type of reliability is the coefficient of equivalence, sometimes called parallel forms, or alternate-form reliability. It can be applied any time one has two or more parallel forms (different versions) of the same test.¹⁴ One can administer both forms to the same group at one sitting, or with a short delay between sittings. A correlation coefficient is then computed on the two sets of parallel scores.

A common use of this type of reliability is in a pretest-posttest research setting. By using the same test for both testing occasions, the researcher cannot know how much of the gain in scores is due to the treatment and how much is due to subjects remembering their answers from the first test. **If one has two parallel forms of the same exam, and the coefficient of equivalence is high, one can use one form as the pretest and the other as the posttest.**

Reliability and Validity

A test can be reliable and not valid for a given testing situation. But can a test be unreliable and still be valid for a given testing situation?

Answer 1: A Test Must be Reliable in Order to be Valid

You will read in some texts that an unreliable instrument is not valid. For example, Bell states "If an item is unreliable, then it must also lack validity, but a reliable item is not necessarily also valid."¹⁵ Sax writes, "a perfectly valid examination must measure some trait without random error...in a reliable and consistent manner."¹⁶ Both of these statements subsumes the concept of reliability under validity rather than depicting them as interdependent concepts.

Nunnally agrees in the sense that "reliability does place a limit on the extent to which a test is valid for any purpose." Further, "high reliability is a necessary, but not sufficient, condition for high validity. If the reliability is zero, or not much above zero, then the test is invalid."¹⁷

Dr. Earl McCallon of North Texas University put it more directly in class. The

¹³Borg and Gall, 285-6

¹⁴Ibid., 283

¹⁵Judith Bell, *Doing Your Research Project* (Philadelphia: Open University Press, 1987), 51

¹⁶Sax, 220

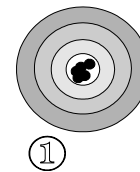
maximum validity of a test is equal to the square root of its reliability.¹⁸ Therefore, test validity is dependent upon test reliability.

Answer 2: A Test Can be Valid Even If It Isn't Reliable

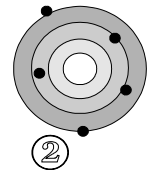
Both Bell and Sax reflect what Payne calls the “cliche of measurement that a test must be reliable before it can be valid.”¹⁹ Payne explains validity in terms of a theoretical inference. Validity is not strictly a characteristic of the instrument but of the “inference that is to be made from the test scores derived from the instrument.” Payne differentiates between validity and reliability as interdependent concepts: validity deals with systematic errors (clarity of instructions, time limits, room comfort) and reliability with unsystematic errors (various levels of subject motivation, guessing, forgetting, fatigue, and growth) in measurement.²⁰

Babbie uses marksmanship to demonstrate the inter-relationship of validity and reliability.²¹ **High reliability** is pictured as a *tight shot pattern* and **low reliability** as a *loose shot pattern*. This is a measure of the shooter’s ability to consistently aim and fire the weapon. **High validity** is pictured as a shot-cluster *on target* and **low validity** as a cluster off to the side. This is a measure of the trueness or accuracy of the sights. Using this analogy we can define four separate conditions:

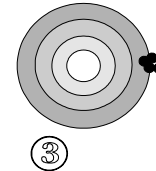
1. High reliability and high validity is a tight cluster in the bull’s eye.



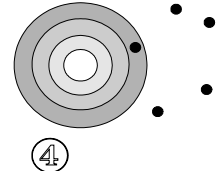
2. Low reliability with high validity is a loose cluster centered around the bull’s eye. (One could certainly question the “validity” of such data)



3. High reliability with low validity is a tight cluster off the target.



4. Low reliability with low validity is a loose cluster off the target.



Payne and Babbie would hold that an instrument can be unreliable and still be valid. A yardstick made out of rubber or a measuring tape made out of yarn are valid instruments for measuring length, even though their measurements would not be accurate. Bell, Sax and Nunnally would say a tape measure made of yarn is not valid if it cannot produce reliable measurements. McCallon demonstrates the boundary condition of $V_{max} = \sqrt{R}$.

In the final analysis, whether we are aiming a rifle or designing a research instrument, our goal should be to get a “tight cluster in the bull’s-eye.” Use instruments which demonstrate the ability to collect data with high validity and high reliability.

Objectivity

The third characteristic of good instruments is objectivity. Objectivity is the extent to which equally competent scorers get the same results. If interviewers A and B interview the same subject and produce different data sets for him, then it is clear that

¹⁷Jum Nunnally, *Educational Measurement and Evaluation*, 2nd ed. (New York: McGraw-Hill Book Company, 1972), 98-99

¹⁸Class notes, Research Seminar, Spring 1983

¹⁹Payne, 259

²⁰*Ibid.*, 254

the measurement is subjective.²² Something about the subject is “hooking” the interviewers differently. The difference is not in the subject, but in the interviewers. A pilot study which uses the researcher’s instrument with subjects similar to those targeted for the study will demonstrate whether it is objective or not. This is particularly important in interview or observation type studies in which human subjectivity can distort the data being gathered.

The validation panel described under “validity” also helps the researcher create an objective test. All items in an item bank should be as clear and meaningful as the researcher can make them. But after the validation panel has evaluated and rated them, the best of the items can be selected for the instrument. This will filter out much of the researcher’s own biases.

An illustration of the objective-subjective tension in instruments is the difference between essay and objective tests. The difference in grades produced on essay tests can be more related to the mood of the grader than the knowledge of the student. A well-written objective test avoids this problem because the answer to every question is definitively right or wrong.

Whether you are planning to use an interview guide, an observation checklist, an attitude scale, or a test, you must work carefully to insure that the data you gather reflects the real world as it is, and not as you want it to be.

Summary

The first element of the Great Triad is **validity**. The four types of validity – **content, predictive, concurrent, and construct** – focus on how well an instrument measures what it purports to measure. The fifth type of validity, “face” validity, is nothing more than a subjective judgement on the part of the researcher and should not be used as a basis for validating instruments.

The second element of the Great Triad is **reliability**. These three approaches to reliability – **stability, internal consistency, equivalence** – focus on how accurate the gathered data is.

The third element of the Great Triad is **objectivity**, which concerns the extent that data is free from the subjective characteristics of the researchers.

Authentic scientific knowing is based on data that is...

VALID

it says what it purports to say

RELIABLE

it says what it says accurately and consistently, and

OBJECTIVE

it says what it says without subjective distortion or personal bias

Vocabulary

coefficient of stability
concurrent validity
construct validity

measure of steadiness, or sameness, of scores over time
degree a new (easier?) test produces **same results** as older (harder?) test
degree to which test **actually measures specified variable** (e.g. 'intelligence')

²¹Babbie, 118

²²Sax, 238

content validity	degree to which test measures course content
Cronbach's coefficient α	measure of internal consistency of a test
coefficient of equivalence	measure of sameness of two forms of a test
face validity	degree a test looks as if it measures stated content
coefficient of internal consistency	degree each item in a test contributes to the total score
Kuder-Richardson formulas	measures of internal consistency
objectivity	the degree that data is not influenced by subjective factors in researchers
parallel forms	tests used to establish equivalence
predictive validity	degree test measures some future behavior
reliability	degree a test measures variables accurately and consistently
Spearman-Brown prophecy formula	used to adjust the r value computed in split-half test
split half test	procedure used to establish internal consistency
test-retest	test given twice over time to establish stability of measures
validity	degree a test measures what it purports to measure

Study Questions

1. Define the terms "instrument," "validity," "reliability," and "objectivity."
2. Discuss the relationship between an operational definition and the procedures for collecting data.
3. Of these three essentials of research, which is most important? Clear research design, accurate measurement, precise statistical analysis. Why?

Sample Test Questions

1. Which of the following is not part of the Great Triad?
 - a. predictive validity
 - b. internal consistency
 - c. instrument objectivity
 - d. empirical measurement
2. Content validity is most concerned with how well the instrument
 - a. predicts some future behavior
 - b. defines an hypothetical concept
 - c. matches the results of another instrument
 - d. measures a specific universe of knowledge
3. The coefficient of stability is more commonly known as the
 - a. split-half test
 - b. test-retest
 - c. the K-R 20
 - d. parallel forms
4. Using Babbie's analogy of shots on a target, a tight cluster off to the side of the target would represent
 - a. high reliability with high validity
 - b. low reliability with high validity
 - c. high reliability with low validity
 - d. low reliability with low validity