INTRODUCTION

Sociologist James A. Quinn states that the tasks of scientific method are related directly or indirectly to the study of similarities of various kinds of objects or events. One of the tasks of scientific method is that of classifying objects or events into categories and of describing the similar characteristics of members of each type. A second task is that of comparing variations in two or more characteristics of the members of a category. Indeed, it is the discovery, formulation, and testing of generalizations about the relations among selected variables that constitute the central task of scientific method.

Fundamental to the performance of these tasks is a system of measurement. S.S. Stevens defines measurement as "the assignment of numerals to objects or events according to rules." This definition incorporates a number of important distinctions. It implies that if rules can be set up, it is theoretically possible to measure anything. Further, measurement is only as good as the rules that direct its application. The "goodness" of the rules reflects on the ***reliability*** and ***validity*** of the measurement--two concepts which we will discuss further later in this lab. Another aspect of definition given by Stevens is the use of the term numeral rather than number. A numeral is a symbol and has no quantitative meaning unless the researcher supplies it through the use of rules. The researcher sets up the criteria by which objects or events are distinguished from one another and also the weights, if any, which are to be assigned to these distinctions. This results in a scale. We will save the discussion of the various scales and levels of measurement till next week. In this lab, our discussion will be focusing on the two fundamental criteria of measurement, i.e., reliability and validity.

The basic difference between these two criteria is that they deal with different aspects of measurement. This difference can be summarized by two different sets of questions asked when applying the two criteria:

Reliability:

a. Will the measure employed repeatedly on the same individuals yield similar results? (stability)

b. Will the measure employed by different investigators yield similar results? (equivalence)

c. Will a set of different operational definitions of the same concept employed on the same individuals, using the same data-collecting technique, yield a highly correlated result? Or, will all items of the measure be internally consistent? (homogeneity)

Validity:

a. Does the measure employed really measure the theoretical concept (variable)?

EXAMPLE: GENERAL APPROACHES TO RELIABILITY/VALIDITY OF MEASURES

1. **Concept**: "Exposure to Televised News"

2. **Definition:** the amount of time spent watching televised news programs

3. **Indicators:**
   a. frequency of watching morning news
   b. frequency of watching national news at 5:30 p.m.
   c. frequency of watching local news
   d. frequency of watching television news magazine & interview programs

4. **Index:**
   Design an eleven-point scale, where zero means "never watch at all," one means "rarely watch" and ten "watch all the time." Apply the eleven-point scale to each of the four indicators by asking people to indicate how often they watch each of the above TV news programs.

Combining responses to the four indicators/or survey questions according to certain rules, we obtain an index of "exposure to televised news program," because we think it measures TV news exposure as we defined it above. A sum score of the index or scale is calculated for each subject, which ranges from 0 (never watch any TV news programs) to 40 (watch all types of TV news program all the time). Now, based on the empirical data, we can assess the reliability and validity of our scale.

DETERMINING RELIABILITY

1. **Stability** (Test-Retest Correlation)

Synonyms for reliability include: dependability, stability, consistency (Kerlinger, 1986). Test-retest correlation provides an indication of stability over time. For example, if we asked the respondents in our sample the four questions once in this September and again in November, we can examine whether the two waves of the same measures yield similar results.

2. **Equivalence**

We want to know the extent to which different investigators using the same instrument to measure the same individuals at the same time yield consistent results. Equivalence may also be estimated by measuring the same concepts with different instruments, for example, survey questionnaire and official records, on the same sample, which is known as multiple-forms reliability.

3. **Homogeneity** (Internal Consistency)

We have three ways to check the internal consistency of the index.

a) **Split-half correlation.** We could split the index of "exposure to televised news" in half so that there are two groups of two questions, and see if the two sub-scales are highly correlated. That is, do people who score high on the first half also score high on the second half?

b) **Average inter-item correlation.** We also can determine internal consistency for each question on the index. If the index is homogeneous, each question should be highly correlated with the other three questions.

c) **Average item-total correlation.** We could correlate each question with the total score of the TV news exposure index to examine the internal consistency of items. This gives us an idea of the contribution of each item to the reliability of the index.

Another approach to the evaluation of reliability is to examine the relative absence of random measurement error in a measuring instrument. Random measurement errors can be indexed by a measure of variability of individual item scores around the mean index score. Thus, an instrument which has a large measure of variability should be less reliable than the one having smaller variability measure.

<u>DETERMINING VALIDITY</u>

1. **Criterion (Pragmatic) Validity**

Based on different <u>time frames</u> used, two kinds of criterion-related validity can be differentiated.

a) **Concurrent validity.** The measures should distinguish individuals -- whether one would be good for a job, or whether someone wouldn't. For example, say a political candidate needs more campaign workers; she could use a test to determine who would be effective campaign workers. She develops a test and administers it to people who are working for her right now. She then checks to see whether people who score high on her test are the same people who have been shown to be the best campaign workers <u>now</u>. If this is the case, she has established the concurrent validity of the test.

b) **Predictive validity.** In this case our political candidate could use the index to predict who would become good campaign workers in the future. Say, she runs an ad in the paper for part-time campaign workers. She asks them all to come in for an interview and to take the test. She hires them all, and later checks to see if those who are the best campaign workers are also the ones who did best on the test. If this is true, she has established the predictive validity of the test and only needs to hire those who score high on her test. (Incidentally, criticisms of standardized tests such as GRE, SAT, etc. are often based on the lack of predictive validity of these tests).

2. **Construct Validity**

Three types of evidence can be obtained for the purpose of construct validity, depending on the research problem.

a) **Convergent validity**. Evidence that the same concept measured in different ways yields similar results. In this case, you could include two different tests. For example:

1. You could place people on meters on respondent's television sets to record the time that people spend with news programs. Then, this record can be compared with survey results of "exposure to televised news"; or

2. You could send someone to observe respondent's television use at their home, and compare the observation results with your survey results.

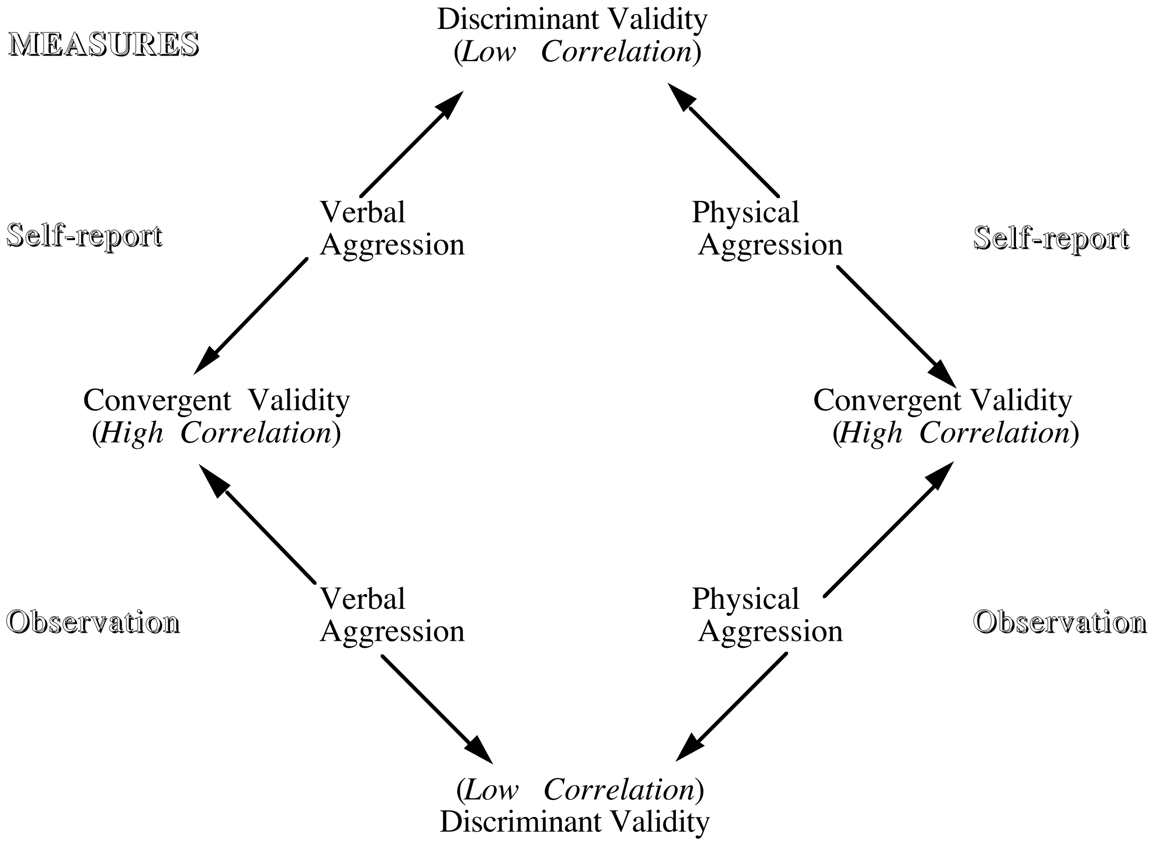
b) **Discriminant validity**. Evidence that one concept is different from other closely related concepts. So, in the example of TV news exposure, you could include measures of exposure to TV entertainment programs and determine if they differ from TV news exposure measures. In this case, the measures of exposure to TV news should not related highly to measures of exposure to TV entertainment programs.

Convergent Validity: Where different measures of the same concept yield similar results. Here we used self-report versus observation (different measures). Yet, these two measures should yield similar results since they were to measure verbal (or physical) aggression. The results of verbal aggression from the two measures should be highly correlated.

Discriminant Validity: Evidence that the concept as measured can be differentiated from other concepts. Our theory says that physical aggression and verbal aggression are different behaviors. In this case, the correlations should be low between questions asked that dealt with verbal aggression and questions asked that dealt with physical aggression in the self-report measure.

Example: Convergent/Discriminant Validity

Theoretical Statement: Physical violence in television leads to physical aggression.



c) **Hypothesis-testing**. Evidence that a research hypothesis about the relationship between the measured concept (variable) and other concept (variable), derived from a theory, is supported. In the case of physical aggression and television viewing, for example, there is a social learning theory stating how violent behavior can be learned from observing and modeling televised physical violence.

From this theory we derive a hypothesis stating a positive correlation between physical aggression and the amount of televised physical violence viewing, then, can be derived. If the evidence collected supports the hypothesis, we can conclude a high degree of construct validity in the measurements of physical aggression <u>and</u> viewing of televised physical violence since the two theoretical concepts are measured and examined in the hypothesis-testing process.

3. **Face Validity**

The researchers will look at the items and agree that the test is a valid measure of the concept being measured just on the face of it. That is, we evaluate whether each of the measuring items matches any given conceptual domain of the concept.

4. **Content Validity**

Content validity regards the representativeness or sampling adequacy of the content of a measuring instrument. Content validity is always guided by a judgment: Is the content of the measure representative of the universe of content of the concept being measured (Kerlinger, 1986)?

Although both face validation and content validation of a measurement is judgmental, the criterion for judgment is different. While the belonging of each item to the concept being measured is to be determined in the evaluation of face validity, content validation determines whether any left-out item should be included in the measurement for its representativeness of the concept.

An example may clarify the distinction. Now, the task here is to determine content validity of a survey measure of "political participation." First, we may specify all the aspects/or dimensions of this concept. Then, we may take the measurement apart to see if all of these dimensions are represented on the test (e.g., the questionnaire). For example:

POLITICAL PARTICIPATION

Dimensions	Behavior: Expressing own viewpoint	Behavior: Learning other's viewpoint	Cognitions
Indicators	Political activity Voting registration Voted in past Membership in organizations	Viewing broadcasts Discuss with family/friends Reading campaign materials	Interest in politics Party affiliation Political knowledge

Have we left out any dimensions? If we are not representing all the major dimensions of the concept, we've got low validity. We won't be measuring some aspects of the concept. Some people will probably get different "scores" on the political participation test than they should, since we haven't measured some of the things we need to. You can think of the domain of the concept "political participation" as a universe consisting of different aspects (dimensions). The measures of the concept are a sample from the universe. The question dealt with in content validity is whether the sample (measurement) is representative enough to cover the whole universe of the concept domain.

Presented in the following are two tables outlining the different ways of establishing reliability and validity. TABLE 4-1 shows that, to establish any form of reliability, one needs two or more independent observations on the same people. As we may realize later, the more independent observations we have on a measurement of a concept taken with different points of time or forms, the more freedom we gain to establish reliability.

TABLE 4-1

TYPES OF RELIABILITY

		Time dimension	
		Multiple-Time-Point Study	Single-Time-Point Study
Forms	Multiple	Equivalence Stability	Equivalence
	Single	Stability	
Items	Multiple	Homogeneity Stability	Homogeneity
	Single	Stability	

TABLE 4-2 shows different types of validity and three criteria which distinguish them. The three criteria are where to start the validation, the evidence and criteria for establishing validity. As you may see, construct validity is the most demanding in that both theory and empirical data are required in the process of validation. Nonetheless, it is the most valuable in theory construction.

TABLE 4-2

TYPES OF VALIDITY

Validity types	Where to Start	Evidence	Criteria
Judgmental (Pre-Data)			
Face Validity	Indicator	Judgmental	What's there
Content Validity	Concept	Judgmental	What's not there
Data-Based (Post-Data)			
Criterion-Related Validity 1. Concurrent 2. Predictive	Criterion Group 1. criterion manifesting currently 2. criterion occurring in the future	Empirical	*Empirical Criterion* Prediction
Construct Validity	Theory	Empirical	*Theoretical Criterion* Convergent Discriminant Hypothesis-testing