CHAPTER 5

MEASUREMENT CONCEPTS

LEARNING OBJECTIVES

- ✓ Define reliability of a measure of behavior:
  - ✓ describe the difference between test-retest, internal consistency, and interrater reliability
- ✓ Discuss ways to establish construct validity:
  - ✓ face validity, content validity, predictive validity, concurrent validity, convergent validity, and discriminant validity
- ✓ Describe the problem of reactivity of a measure of behavior
  - ✓ discuss ways to minimize reactivity
- ✓ Describe the properties of the four scales of measurement:
  - ✓ nominal, ordinal, interval, and ratio

RELIABILITY OF MEASURES

- **Reliability**: *deals with the consistency or stability of a measure of behavior*
  - It is the extent to which an experiment, test, or any measuring procedure yields the same result on repeated trials.

RELIABILITY OF MEASURES

Reliability

- Without the agreement of independent observers to replicate research procedures, or the ability to use research tools and procedures that yield consistent measurements, researchers would be unable to satisfactorily draw conclusions, formulate theories, or make claims about the generalizability of their research.

RELIABILITY OF MEASURES

- A more formal way of understanding reliability is to use the concepts of true score and measurement error.
- Any measure that you make can be thought of as comprising two components:
  - a **true score**, which is the real score on the variable, (e.g., if an intelligence test score reflects the actual intelligence of the individual being tested, it is the true intelligence score of that person)
  - **measurement error**. An unreliable measure of intelligence contains considerable measurement error and so does not provide an accurate indication of an individual's true intelligence.
- **True score**: is the real score on a variable
- **Measurement error**: is a measurement that produces unreliable scores

RELIABILITY OF MEASURES

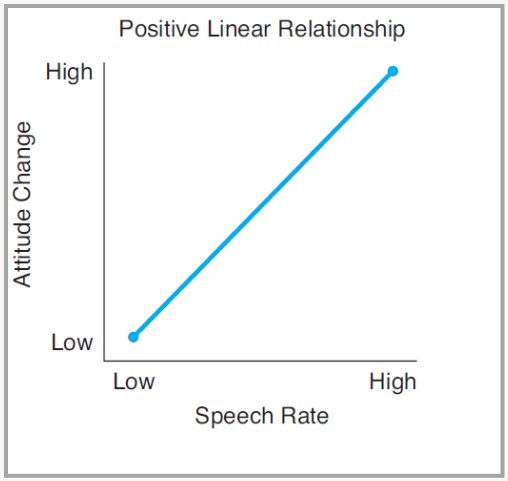
- *The most common correlation coefficient when discussing reliability is the* **Pearson product-moment correlation coefficient**.
- The Pearson correlation coefficient (symbolized as *r*) can range from 0.00 to +1.00 and 0.00 to -1.00.

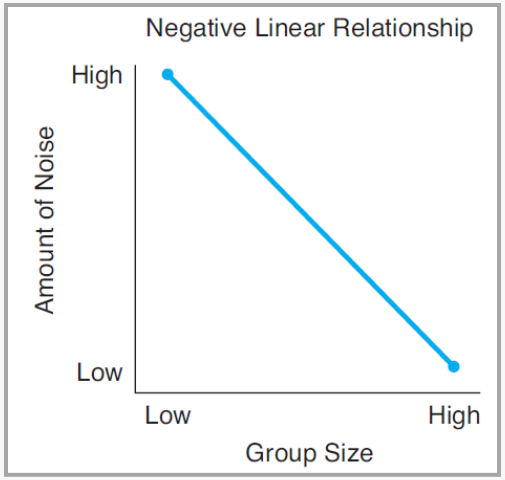
RELIABILITY OF MEASURES

- A correlation of 0.00 tells us that the two variables are not related at all.
- The closer a correlation is to 1.00, either +1.00 or -1.00, the stronger is the relationship.
- *The positive and negative signs provide information about the direction of the relationship.*

RELIABILITY OF MEASURES

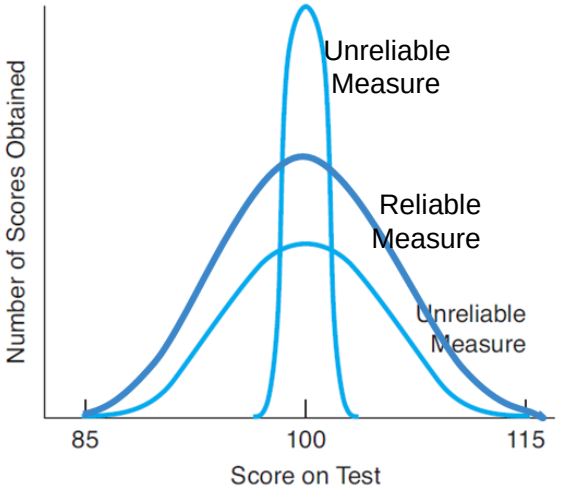
- When the correlation coefficient is positive (a plus sign), there is a positive linear relationship—high scores on one variable are associated with high scores on the second variable.
- A negative linear relationship is indicated by a minus sign—high scores on one variable are associated with low scores on the second variable.

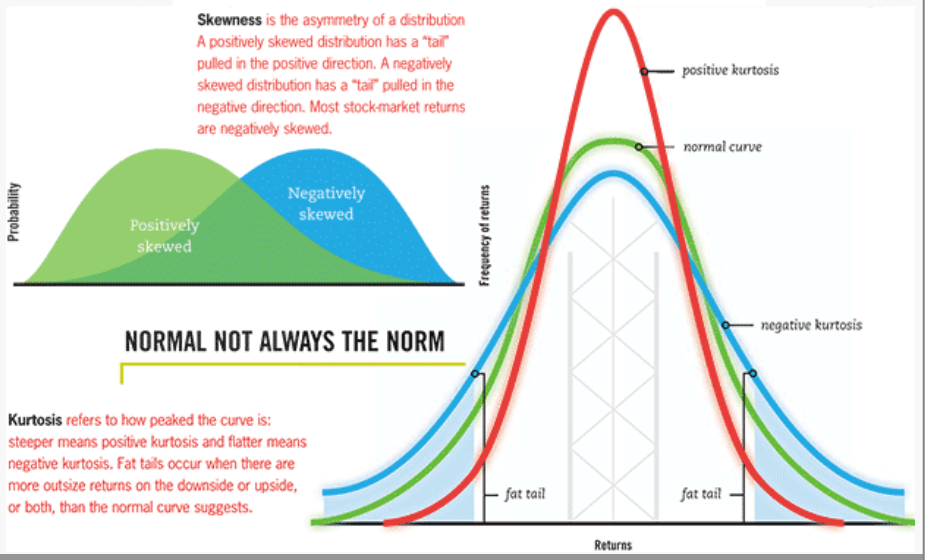




COMPARING DATA OF A RELIABLE AND UNRELIABLE MEASURE

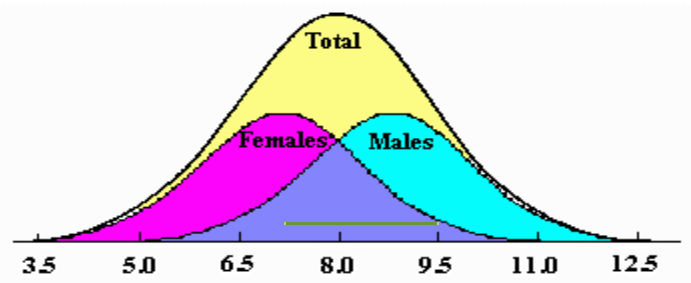
- Reliable vs. Reliable
  - The top figure demonstrates how reliable measures will yield scores where students tend to earn similar scores. Here, in a reliable measure, students will earn about 100 on their test scores. In an unreliable measure, scores will spread out where students aren't really obtaining similar scores one way or the other.
- A fundamental task in many statistical analyses is to characterize the *location* and *variability* of a data set.
- A further characterization of the data includes skewness and kurtosis.
- Measures of Dispersion tells us about the variation of the data set.
- **Skewness** tells us about the *direction of variation* of the data set.
- **Kurtosis** is a parameter that describes the shape of a random variable's probability distribution.



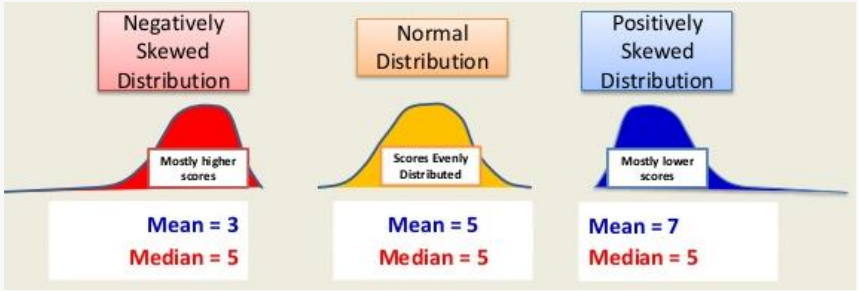


COMPARING DATA OF A RELIABLE AND UNRELIABLE MEASURE: Skewness

- **Types of Skewness:**
- Teacher expects most of the students get good marks. If it happens, then the curve looks like the yellow, normal curve bellow:



  - But for some reasons (e. g., lazy students, not understanding the lectures, not attentive etc.) it is not happening. So we get skewed curves.



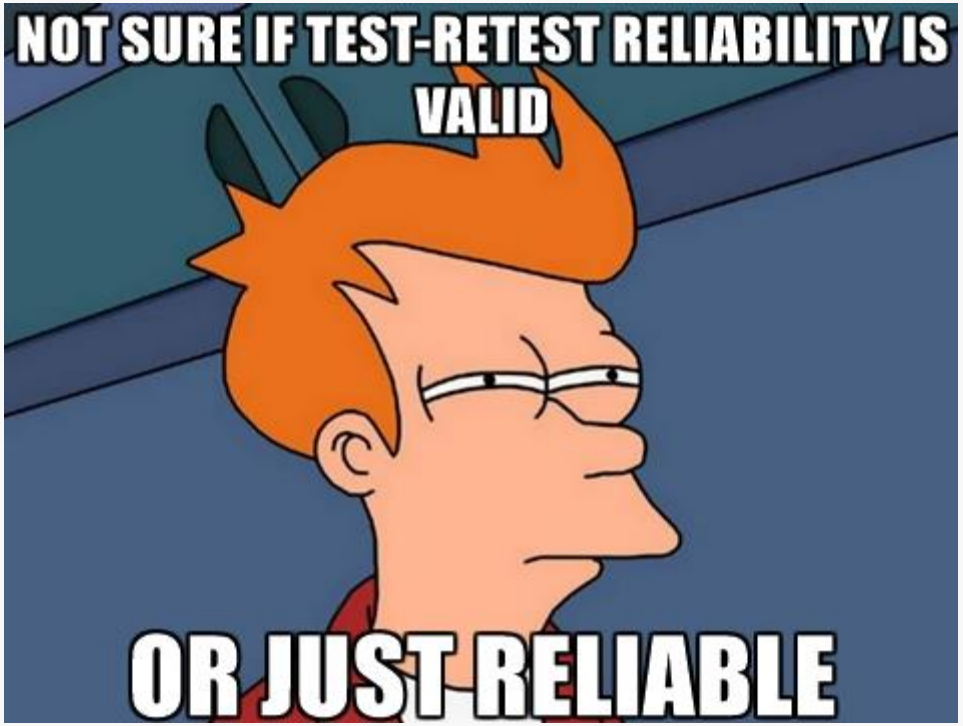
- The first one is known as positively skewed and the last one is known as negatively skewed curve.

COMPARING DATA OF A RELIABLE AND UNRELIABLE MEASURE: Kurtosis

- **Kurtosis** is a parameter that describes the shape of a random variable's probability distribution.
- These graphs illustrate the notion of kurtosis. The graph on the right has higher kurtosis than the graph on the left. It is more peaked at the center, and it has fatter tails. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution.
- Higher kurtosis means more of the variance is the result of infrequent extreme deviations (outliers). If a distribution of test scores is very tall, it may indicate a problem with the validity of your decision making processes or a problem with sampling.
- Flat topped distributions indicate not enough variability among scores, which could indicate a problem with your test, measure, or survey.

TESTS OF RELIABILITY

- **Test-retest reliability**
- **Internal consistency reliability**
- **Split-half reliability**
- **Item-total**
- **Interrater reliability**



RELIABILITY OF MEASURES

- **Test-retest reliability**: Assessed by measuring the same individuals at two points in time
  - For example, if students take a test, you would expect them to show very similar results if they take the same test a few months later.
  - One would then have two scores for each person, and a correlation coefficient could be calculated to determine the relationship between the first test score and the retest score.
    - High reliability is indicated by a high correlation coefficient showing that the two scores are very similar.

RELIABILITY OF MEASURES

- **Test-retest reliability**:
  - **Vulnerable to artificiality**
    - Given that test-retest reliability requires administering the same test twice, the correlation might be artificially high because the individuals remember how they responded the first time, also known as **practice effects**.
      - **Alternate forms of reliability** are sometimes used to avoid this problem; it requires administering two different forms of the same test to the same individuals at two points in time.
  - **Vulnerable to maturation**
    - If the test and retest are taken at the beginning and at the end of the semester, it can be assumed that the intervening lessons will have improved the ability of the students.
    - Even if a test-retest reliability process is applied with no sign of intervening factors, there will always be some degree of error.

RELIABILITY OF MEASURES

- **Internal consistency reliability**: Is assessed using responses at only one point in time. Because all items measure the same variable, they should yield similar or consistent results.
  - It measures the correlations between different items on the same test (or the same subscale on a larger test), because items that propose to measure the same general construct produce similar scores.
  - For example, if a respondent expressed agreement with the statements "I like to ride bicycles" and "I've enjoyed riding bicycles in the past", and disagreement with the statement "I hate bicycles", this would be indicative of good internal consistency of the test.
  - Internal consistency is usually measured with Cronbach's alpha, a statistic calculated from the pairwise correlations between items.

RELIABILITY OF MEASURES

- **Split-half reliability**: *Correlation of the total score on one half of the test with the total score on the other half*. Split-half reliability is another subtype of internal consistency reliability.
  - The two halves are created by randomly dividing the items into two parts.
  - Another way to test split-half reliability is by "splitting in half" all items of a test that are intended to probe the same area of knowledge in order to form two "sets" of items.
  - The *entire* test is administered to a group of individuals, the total score for each "set" is computed, and finally the split-half reliability is obtained by determining the correlation between the two total "set" scores.
  - The indicator of reliability based on internal consistency that is commonly used is called **Cronbach's alpha**,
    - Cronbach's alpha provides one with the average of all possible split-half reliability coefficients

RELIABILITY OF MEASURES

- It is also possible to examine the correlation of each item score with the total score based on all items. This is called an **Item-total correlation**.
- They are very informative because they provide information about each individual item.
  - An item-total correlation test is performed to check if any item in the set of tests is inconsistent with the averaged behavior of the others, and thus can be discarded.
  - The analysis is performed to purify the measure by eliminating 'garbage' items prior to determining the factors that represent the construct; that is, the meaning of the averaged measure.

RELIABILITY OF MEASURES

- **Interrater reliability** *is the extent to which raters agree in their observations*
- Reliability and accuracy of measures
  - **Reliability** is clearly important when researchers develop measures of behavior, however, it is not the only characteristic of a measure or the only thing that researchers worry about.
  - Reliability tells one about measurement error but it does not tell one about whether there is a good measure of the variable of interest.

VALIDITY vs RELIABILITY

- A measure can be highly reliable but not accurate
- For example, an instrument can consistently measure a certain **construct**, or theory, of personality, such as Neuroticism.
  - However, it may not capture all of the factors, or elements, that make up that construct.
  - Everyone possesses some characteristics of neuroticism in their personalities, but that does not say that everyone is strictly neurotic.
- **Construct validity** concerns whether one's methods of studying variables are accurate. That is, it refers to the adequacy of the operational definition of variables.
  - To what extent does the operational definition of a variable actually reflect the true theoretical meaning of the variable?

INDICATORS OF CONSTRUCT VALIDITY OF A MEASURE

Validity	Definition
Face validity	As the name suggests, it is a measure of how a measure, 'at face value,' *appears to measure the construct of interest*. A direct measurement of face validity is obtained by asking people to rate the validity of a test as it appears to them.
Content validity	Refers to the extent to which a measure represents all facets of a given social construct. *Is it really measuring the construct it attempts to measure?*
Predictive validity	*is the extent to which a score on a scale or test predicts scores on some criterion measure*. For example, the **validity** of a cognitive test for job performance is the correlation between test scores and, for example, supervisor performance ratings..

INDICATORS OF CONSTRUCT VALIDITY OF A MEASURE

Validity	Definition
Concurrent validity	Is demonstrated by research that examines the relationship between the measure and a criterion behaviour at the same time (concurrently). Such as asking questions about how much exercise an individual gets and then monitoring their actual activity with a step counter. If outcomes from both tests have close agreement they have concurrent validity.
Convergent validity	Is the extent to which scores on the measure in question are related to scores on other measures of the same construct or similar constructs.
Discriminant validity	*When the measure is **not** related to variables with which it should not theoretically be related.*

REACTIVITY OF MEASURES

- **Reactivity**: A measure is reactive if awareness of being measured changes an individual's behavior
- Measures of behavior vary in terms of their potential reactivity
  - Doctors often experience this when having their patience blood pressure taken. The test itself tends to make some people nervous, thus raising their blood pressure.

SCALES OF MEASUREMENT

Scale	Description
Nominal	*Categories with no numeric scales*. Instead, categories or groups simply differ from one another
Ordinal	Rank ordering. The numeric values are limited. Instead of having categories that are simply different, as in a nominal scale, the categories can be ordered from first to last. Letter grades are a good example of an ordinal scale.
Interval	Difference between the numbers is meaningful. The intervals between the numbers are equal in size. The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees.
Ratio	*Zero indicates absence of variable measured*. Variables like height, weight, enzyme activity are ratio variables. Temperature, expressed in F or C, is not a ratio variable.

IMPORTANCE OF MEASUREMENT SCALES

- When one reads about the operational definitions of variables, one will recognize the levels of the variable in terms of these types of scales.
- The conclusions one draws about the meaning of a particular score on a variable depend on which type of scale was used.
- Rating scale provides quantitative information about the variables
- Scale determines the types of statistics that are appropriate when the results are analyzed

OK to compute…	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Frequency Distribution	Yes	Yes	Yes	Yes
Median and Percentiles	No	Yes	Yes	Yes
Add or Subtract.	No	No	Yes	Yes
Mean, Standard Deviation, Standard Error	No	No	Yes	Yes
Ratio or Coefficient of Variation (i.e., standardized measure of dispersion)	No	No	No	Yes

LAB

- **Complete:**
  - "Hypothesis and Operational Definition Exercise"
  - "Reliability, Validity, & Reactivity"
    - Due Before Class Next Tuesday