**PRINT**

**INSTANCEBEGINEDITABLE NAME="DOCTITLE" ASSESSMENT 3 CONTEXT INSTANCEENDEDITABLE**

InstanceBeginEditable name="Body"

**Measurement and Statistics in Research**

Levels of Measurement

Different variables are measured on different scales. Continuous variables have many possible values, such as age, weight, or even height. Discrete or categorical variables have only a few possible values, such as gender, marital status, or letter grade. Categorical variables that only have two possible values are called "dichotomous variables" because they are either one thing or the other, but nothing else.

In addition to these distinctions, there are also four levels of measurement that help identify the type of variable in use in a study and, therefore, have implications on the types of analysis that can be done.

* Nominal variables are categorical. These are variables that are, in essence, qualitative characteristics such as hair color or political affiliation, to which we assign numerical values for the purpose of analyzing them statistically.
* Ordinal variables are also categorical, but they can be ordered along a continuum, such as with ranks.
* Interval variables are continuous. Ordered, like ordinal variables, they differ in that the distance between the points is equal and thus measurable in a mathematically meaningful way.
* Ratio variables are also continuous, and in addition to being ordered and having equal distance between points, variables on this scale also have an absolute zero point, thus making it the most precise form of measurement.

Reliability

Reliability of a measure is the degree to which it performs consistently. This is measured by correlating the first score with the second score—if the correlation is high, the reliability is high. There are four types of reliability.

* Test-Retest Reliability measures whether or not the same individual will have the same score on a measure if given more than once.
* Inter-Rater Reliability measures whether or not one rater rates a behavior the same as another rater. Problems with Inter-Rater reliability indicate a problem with the definition of the behavior being rated. For example if a rater is measuring the number of pro-social behaviors exhibited by children in a classroom, but that rater only counts verbal behaviors and not nonverbal behaviors, their ratings would not be consistent with a rater counting both verbal and nonverbal behaviors.
* Internal Consistency Reliability measures whether or not one section of a test is consistent with another section of the test. For example, are the first 20 questions measuring the target behavior in the same way as the second set of 20 questions? Are the scores on each set highly correlated?
* Parallel Forms Reliability measures whether or not one form of a measure is consistent with another. Occasionally, a measure will have two forms, Form A and Form B, for example. The different forms are used to measure the same behavior, but they allow the test taker to avoid answering the same, exact questions on multiple test administrations. If Parallel Forms Reliability is high, we can be assured that both forms are measuring the same behavior.

Validity

Validity of a measure is the degree to which it measures what it says it is measuring. Does the test do what it is supposed to do? A valid test measures the behavior it intends to measure, rather than some other extraneous variable. For example, if we are studying healthy eating, and the way we decided to measure healthy eating is weight loss, is that really a valid measure of healthy eating? People who became anorexic would have a high score on healthy eating if their weight dropped, but would they really be eating healthy?

There are three types of validity:

* Content Validity is often called face validity. It is a simple global measure of validity that tells us whether or not the content actually reflects the construct. The best way to determine content validity is simply to ask an expert. Does a test with questions about the Revolutionary War, the Civil War, and the Civil Rights movement represent a valid test of knowledge of American history? An expert could tell you, yes, it does.
* Criterion Validity is the degree to which a test can predict performance, either currently (concurrent validity) or in the future (predictive validity). In order to assess criterion validity, test results must be compared to a criterion. For example, people who have been successful supervisors could be given a personality test. Their scores would be made the criterion, and other employees that took the test and met that criterion would be expected to be good supervisory candidates.
* Construct Validity is the degree to which the test is measuring the underlying construct or behavior you think it is. Construct validity is determined by comparing it to another test that has been shown to be a valid measure of that construct.

Methods of Measurement

There are many different types of tests that can measure many different aspects of the human experience—thoughts, behaviors, emotions, skills, knowledge, attitudes, and so on. Most tests are multiple choice, but some are forced response, and some may even be open ended. There are six major types of tests:

* Achievement Tests measure performance, knowledge, and ability. Standardized tests are the most common type of achievement test. Achievement tests measure what a person already knows and what they can already do.
* Aptitude Tests measure what a person is capable of doing. Intelligence tests are a good example of an aptitude test.
* Attitude Tests measure how someone feels about something. They can be aimed at anything from politics to personal health problems.
* Personality Tests measure personality traits, or behaviors that represent those personality traits. Personality tests can be projective (like the famous Rorschach Inkblot test) or objective (true or false test questions that indicate one's feelings or opinions on something).
* Observational Techniques can be used to test the presence and intensity of a particular behavior.
* Questionnaires are self-report tests that can incorporate many different types of test questions, including multiple choice, true or false, or open-ended responses.

Collection of data is a very detailed and precise process. Errors in data collection can result in inaccurate results of a study, and that makes the entire process invalid. There are many procedures that make data collection an organized logical process so that you can maintain the integrity of the information and draw valid conclusions from your research.

It is essential that data be collected in an organized and careful way. Researchers have a responsibility to make sure that the data is protected, both in terms of confidentiality issues, as well as in terms of accuracy of the recording of the data so that the participants are not misrepresented, and that study's results are valid.

* Forms: Researchers collect data on forms, either preexisting test forms or forms the researcher creates to compile information in one place, such as demographic information.
* Coding: Often, data must be coded in order to be analyzed. For example, age may be coded into child, adolescent, young adult, and adult for the purposes of the analysis. A coding procedure is established and decided upon before data is collected to make sure you have enough information to proceed with coding.
* Collection: Actual collection of the data is done under controlled, planned circumstances. Few distractions should be present, and the data collection environment should be as similar as possible for each participant to avoid any extraneous factors impacting the results.
* Data Entry: Once data is collected, it must be entered into a data analysis program such as SPSS. It is extremely important to make sure that you check and double-check your data entry. Human error is a big possibility at every stage but especially in transferring data from one form to another. If there is a way to electronically record data directly into the data-analysis software as you are collecting it, that can reduce one instance of possible data-recording error.

Data Analysis

Analyzing your data may seem daunting, but that is the point at which you can begin to see the fruits of your labor. Data analysis is what allows us to determine if our hypothesis is confirmed. Data can be analyzed in several ways, the two main ways being descriptive and inferential.

Descriptive Statistics

Descriptive statistics tell us the general characteristics of our data. Descriptive statistics can be presented in the form of measures of central tendency such as the mean (most common and useful), the median, or the mode. The mean is basically the average. What is the average score on an intelligence test? What is the average score on a test of anxiety or depression. The mean of your sample can then be compared either to another sample such as a control group that did not receive any treatment, or a normative group that the test creator has indicated to give researcher a source of comparison.

Descriptive statistics can also be presented in the form of measures of variability, such as range and standard deviation. Determining the variability of the data can give you an idea of how your data is dispersed. Are a lot of people scoring on the high end of the scale? Are most of the participants scoring very low? You can get the same mean from two vastly different ranges of data, so the range gives us another important picture of our sample.

The standard deviation simply tells us the average amount of variation each score has from the mean. Understanding how scores vary, can help us when comparing them to other scores, either those of another group in our sample, or those in the general population.

Inferential Statistics

Inferential statistics are the tools we use to compare our data to determine the results of our study. Inference is the leap we make from the results of our study with our sample to the population as a whole. Inference is what we use to generalize our results.

The central limit theorem tells us that as long as our sample is greater than thirty, and the distribution of the sample is normal, we can infer that our results can be applied to the larger population. The idea of statistical significance also plays into this picture because while ideally, we would like for our sample to be the perfect representation of the population, it may not be, and there may be other sources of error as well—in the way the data was collected, in the way the data was recorded, and so on.

There is some degree of risk in the inference that your results are in fact due only to the variables in your study and not chance or error, and this risk is dealt with by using statistical significance to interpret the data. By stating the level of statistical significance, you are stating your degree of confidence in your result. A statistical significance of .05 means that you are accepting that there is a 5 percent chance that your results are due to some form of error.

Statistical Tests

A number of tests of significance are available for use in analyzing your data. The *t* test is the most common test and can be used to test the difference between two groups. Often though, studies are more complex, involving multiple variables, so an Analysis of Variance (ANOVA) can be used to further analyze the variables in question. Figure 8.3 in the Salkind text provides an excellent primer for choosing the appropriate statistic for the types of data you are analyzing.

Further discussion of statistics is beyond the scope of this course but is covered in great depth in an actual statistics course.

Interpreting Results

While it may be the goal of researchers to find statistical significance, that may not have much meaning in the end. A study that shows a girl's scores are significantly lower than boy's score on a measure of assertiveness means nothing unless we make some interpretation of the meaningfulness of those results. What does that mean for girls? How does that play out clinically, in real life? How does it impact girls' lives, relationships, and self-esteem?

Researchers must be careful not to overextend the meaning of one result of one study, but a body of research in one area demonstrating similar results in different settings can help a researcher come to meaningful conclusions about their data and spur additional research aimed at rectifying a situation or helping a certain group overcome a certain characteristic or circumstance. This is the real reason for research: to better understand, so that we can help.

Reference

Salkind, N. J. (2012). *Exploring research* (8th ed.). Upper Saddle River, NJ: Pearson.