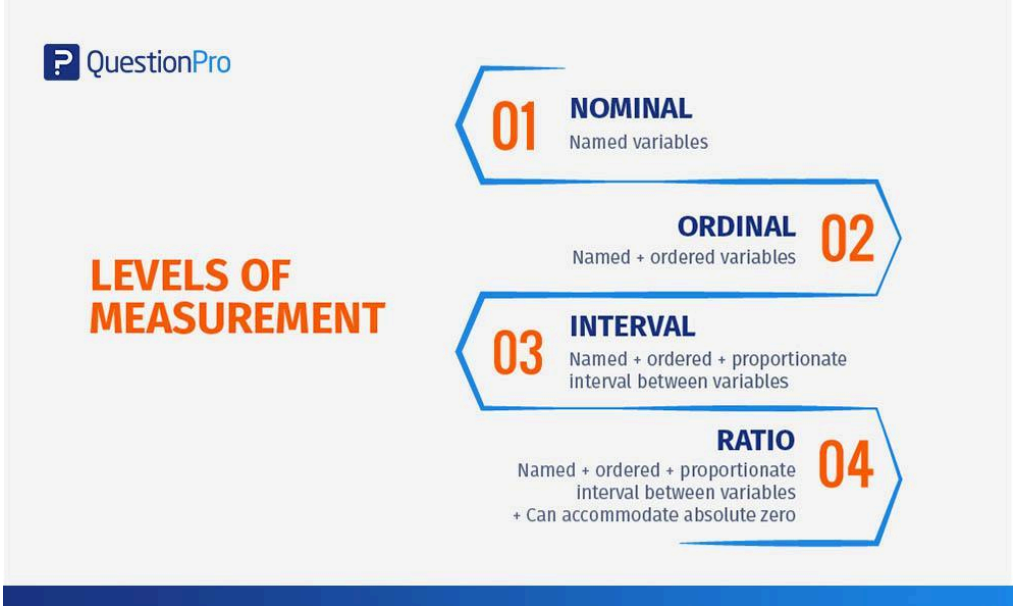
Assignment 3
[Dr. Krishankant Tiwari]

Q1. Levels of Measurement

A variable has one of four different levels of measurement: Nominal, Ordinal, Interval, or Ratio. (Interval and Ratio levels of measurement are sometimes called Continuous or Scale). It is important for the researcher to understand the different levels of measurement, as these levels of measurement, together with how the research question is phrased, dictate what statistical analysis is appropriate.



Four Different Levels of Measurement

In descending order of precision, the four different levels of measurement are:

- Nominal–Latin for name only (Republican, Democrat, Green, Libertarian)
- Ordinal–Think ordered levels or ranks (small–8oz, medium–12oz, large–32oz)
- Interval–Equal intervals among levels (1 dollar to 2 dollars is the same interval as 88 dollars to 89 dollars)
- Ratio–Let the "o" in ratio remind you of a zero in the scale (Day 0, day 1, day 2, day 3, ...)

Now in details

1. The first level of measurement is **nominal level** of measurement. In this level of measurement, the numbers in the variable are used only to classify the data. In this level of measurement, words, letters, and alpha-numeric symbols can be used. Suppose there are data about people belonging to three different gender categories. In this case, the person belonging to the female gender could be classified as F, the person belonging to the male gender could be classified as M, and transgendered classified as T. This type of assigning classification is nominal level of measurement.
2. The second level of measurement is the **ordinal level** of measurement. This level of measurement depicts some ordered relationship among the variable's observations. Suppose a student scores the highest grade of 100 in the class. In this case, he would be assigned the first rank. Then, another classmate scores the second highest grade of a 92; she would be assigned the second rank. A third student scores a 81 and he would be assigned the third rank, and so on. The ordinal level of measurement indicates an ordering of the measurements.
3. The third level of measurement is the **interval level** of measurement. The interval level of measurement not only classifies and orders the measurements, but it also specifies that the distances between each interval on the scale are equivalent along the scale from low interval to high interval. For example, an interval level of measurement could be the measurement of anxiety in a student between the score of 10 and 11, this interval is the same as that of a student who scores between 40 and 41. A popular example of this level of measurement is temperature in centigrade, where, for example, the distance between 940C and 960C is the same as the distance between 1000C and 1020C.
4. The fourth level of measurement is the **ratio level** of measurement. In this level of measurement, the observations, in addition to having equal intervals, can have a value of zero as well. The zero in the scale makes this type of measurement unlike the other types of measurement, although the properties are similar to that of the interval level of measurement. In the ratio level of measurement, the

divisions between the points on the scale have an equivalent distance between them.

Q2 Reliability and Validity of Measurement

Reliability

Reliability means that a measurement procedure yields consistent or equivalent scores when the phenomenon being measured is not changing (or that the measured scores change in direct correspondence to actual changes in the phenomenon). If a measure is reliable, it is affected less by random error or chance variation than if it is unreliable. Reliability is a prerequisite for measurement validity: We cannot really measure a phenomenon if the measure we are using gives inconsistent results. In fact, because it usually is easier to assess reliability than validity, you are more likely to see an evaluation of measurement reliability in a research report than an evaluation of measurement validity.

Test–Retest Reliability

When researchers measure a phenomenon that does not change between two points separated by an interval of time, the degree to which the two measurements are related to each other is the **test–retest reliability** of the measure. If you take a test of your research methodology knowledge and retake the test 2 months later, the test is performing reliably if you receive a similar score both times—presuming that nothing happened during the 2 months to change your research methodology knowledge. We hope to find a correlation between the two tests of about .7 and prefer even a higher correlation, such as .8.

Internal Consistency

When researchers use multiple items to measure a single concept, they are concerned with **internal consistency**. For example, if the items composing the CES–D (like those in Exhibit 4.2) reliably measure depression, the answers to the questions should be highly associated with one another. The stronger the association among the individual items and the more items that are included, the higher the reliability of the scale

One method to assess internal consistency is to divide the scale into two parts, or **split-half reliability**. We might take a 20-item scale, such as the CES–D, and sum the scores of the first 10 items, sum the scores of the second 10 items (items 11–20), and then correlate the scores for each of the participants. If we have internal consistency, we should have a fairly high correlation, such as .8 or .9. This correlation typically gets higher the more items there are in the scale. So what may be considered a fairly high split-half reliability score for a 6-item scale might not be considered a high score for a 20-item scale.

As you can imagine, there are countless ways in which you might split the scale, and in practical terms, it is nearly impossible to split the scale by hand into every possible combination. Fortunately, the speed of computers allows us to calculate a score that indeed splits the scale in every combination. A summary score, such as **Cronbach's alpha coefficient**, is the average score of all the possible split-half combinations. In Radloff's (1977) study, the alpha coefficients of different samples were quite high, ranging from .85 to .90.

Alternate-Forms Reliability

Researchers are testing **alternate-forms reliability** (or parallel-forms reliability) when they compare subjects' answers to slightly different versions of survey questions (Litwin, 1995). A researcher may reverse the order of the response choices in a scale, modify the question wording in minor ways, or create a set of different questions. The two forms are then administered to the subjects. If the two sets of responses are not too different, alternate-forms reliability is established.

Validity

Validity refers to the extent to which measures indicate what they are intended to measure. More technically, a valid measure of a concept is one that is (a) closely related to other apparently valid measures of the concept, (b) closely related to the known or supposed correlates of that concept, and (c) not related to measures of unrelated

concepts (adapted from Brewer & Hunter, 2005). Measurement validity is assessed with four different approaches: face validation, content validation, criterion validation, and construct validation.

Face Validity

Researchers apply the term **face validity** to the confidence gained from careful inspection of a concept to see whether it is appropriate "on its face." A measure is face valid if it obviously pertains to the meaning of the concept being measured more than to other concepts (Brewer & Hunter, 2005). For example, a count of how many drinks people consumed in the past week would be a face-valid measure of their alcohol consumption.

Although every measure should be inspected in this way, face validation does not provide any evidence of measurement validity. The question "How much beer or wine did you have to drink last week?" looks valid on its face as a measure of frequency of drinking, but people who drink heavily tend to underreport the amount they drink. So the question would be an invalid measure in a study that includes heavy drinkers.

Content Validity

Content validity establishes that the measure covers the full range of the concept's meaning. To determine that range of meaning, the researcher may solicit the opinions of experts and review literature that identifies the different aspects or dimensions of the concept.

In contrast, experts may disagree with the range of content provided in a scale. The CES–D depression scale includes various dimensions of somatic symptoms and negative feelings. Some experts (e.g., Liang, Tran, Krause, & Markides, 1989) have questioned the presence of some items such as "feeling fearful" or "people dislike me," suggesting that these items are not reflective of the dimensions of depression.

This example illustrates one of the difficulties in relying solely on face or content validity. In the end, they are subjective assessments

of validity and, therefore, are weaker forms of validity than the next two types of validity, which are based on empirical assessments.

Criterion Validity

Criterion validity is established when the scores obtained on one measure are similar to scores obtained with a more direct or already validated measure of the same phenomenon (the criterion). A measure of blood-alcohol concentration or a urine test could serve as the criterion for validating a self-report measure of drinking as long as the questions we ask about drinking refer to the same period. A measure of depression could be compared to another accepted self-administered depression scale. SAT or ACT scores could be compared to academic success in college. In each of these cases, the measure is being compared to some criterion believed to measure the same construct.

The criterion that researchers select can be measured either at the same time as the variable to be validated or after that time. **Concurrent validity** exists when a measure yields scores that are closely related to scores on a criterion measured at the same time. A store might validate its test of sales ability by administering the test to sales personnel who are already employed and then comparing their test scores to their sales performance. A measure of walking speed based on mental counting might be validated concurrently with a stop watch. **Predictive validity** is the ability of a measure to predict scores on a criterion measured in the future. For example, a store might administer a test of sales ability to new sales personnel and then validate the measure by comparing these test scores with the criterion—the subsequent sales performance of the new personnel.

An attempt at criterion validation is well worth the effort because it greatly increases confidence that the measure is measuring what was intended. However, for many concepts of interest to social work researchers, no other variable might reasonably be considered a criterion. If we are measuring feelings, beliefs, or other subjective

states, such as feelings of loneliness, what direct indicator could serve as a criterion?

Construct Validity

Measurement validity can also be established by showing that a measure is related to a variety of other measures as specified in a theory. This validation approach, known as **construct validity**, is commonly used in social research when no clear criterion exists for validation purposes. This theoretical construct validation process relies on using a deductive theory with hypothesized relationships among the constructs (Koeske, 1994). The measure has construct validity (or theoretical construct validity) if it "behaves" as it should relative to the other constructs in the theory. For example, Danette Hann, Kristin Winter, and Paul Jacobsen (1999) compared subject scores on the CES–D to a number of indicators that they felt from previous research and theory should be related to depression: fatigue, anxiety, and global mental health. The researchers found that individuals with higher CES–D scores tended to have more problems in each of these areas, giving us more confidence in the CES–D's validity as a measure.

A somewhat different approach to construct validation is **discriminant validity**. In this approach, scores on the measure to be validated are compared to scores on another measure of the same variable and to scores on variables that measure different but related concepts. Discriminant validity is achieved if the measure to be validated is related most strongly to its comparison measure and less so to the measures of other concepts. The CES–D would demonstrate discriminant validity if the scale scores correlated strongest with the Beck Depression Inventory (a validated scale to measure depression) and correlate lower with the Beck Anxiety Inventory (a validated scale to measure anxiety).

Convergent validity is achieved when you can show a relationship between two measures of the same construct that are assessed using different methods (Koeske, 1994). For example, we might compare the CES–D scale scores to clinical judgments made by

practitioners who have used a clinical protocol. The CES–D scores should correlate with the scores obtained from the clinical protocol.