Measurement

Systematically assigning categories (names) and/or magnitudes (numbers) to things.

Four topics will be addressed:

1. Measurement scales
2. Reliability
3. Construct Validity
4. Reactivity

I. Properties of Scales

A) *Difference* - Measurement scales require that objects have attributes that can be distinguished (assigned different labels).

Example - sex (male and female) or eye color (blue, brown, hazel).

B) *Magnitude* - Can the scale show that one object has a greater, equal or lesser value than another object. Examples above (sex, eye color) do not have this.

Example - An individual's ordering of preferences for some set of objects or events (e.g. color preferences, food preferences).

C) *Equal intervals* - Can the distances between adjacent items be compared? Examples above can not.

Example - The IQ scale has equal intervals. The difference between 120 and 135 is the same as the difference between 95 and 110.

D) *True zero* - Is there a point where the attribute or property measured by the scale does not exist? None of examples above meet this.

Example - Speed of response (time) has a true zero.

If a scale has one of these properties, it also has all of the "earlier" ones. That is, a scale with equal intervals also has magnitude and difference.

II. Four Types of Scales

These properties are the basis for four types of scales. The arithmetic operations that can be performed on a scale are determined by these properties. Since these arithmetic operations determine the statistics that can be used on the data, we need to understand the types of scales.

Since we are measuring behavior, some of our measurements will reflect physical attributes (e.g. how fast someone does a task) and some will reflect psychological attributes (subjective answers to questions). Keep this in mind as we describe the types of scales.

A) *Nominal* - Has Property of Difference.

No arithmetic operations can be done. For example, suppose we collected the zip-codes of all students. What does it mean to compute an average zip-code? Zip codes are simply labels. If we were to determine the number of times each zip code occurred, we could make claims about which one occurred more often.

Example - Are men and women represented in the faculty in proportion to their representation in the population? To answer this, we would make counts of the numbers in each of our categories. Now, how do we test the data. Remember, we can not do arithmetic operations.

The method is to use the counts for the population as our "baseline" and determine the odds that the counts for the faculty are "the same" (come from the same population or distribution).

B) *Ordinal* - Difference and Magnitudes.

If you were given a list of occupations and you were to rank order them from most preferred to least preferred, your preference ranking would be on an *ordinal* scale.

These scales allow you to make statements about more and less. You can run a correlation comparing orderings (to see how similar they are). However, since intervals are not equal, you can not compute mean rankings.

C) *Interval* - Difference, Magnitudes, Equal Intervals.

Since there is a single size unit, you can add and subtract values and compute means. This property opens up a wide range of statistics.

A common use of interval scales is the *Likert Scale*. In a typical use of this on a questionnaire, a subject is asked to rate, on a 1 to 5 scale, their agreement with a statement. The numbers 1 through 5 are treated as equally spaced and means, across (or within) subjects, can be tabulated.

D) *Ratio* - Difference, Magnitudes, Equal Intervals and True Zero.

Since there is a true zero point, we can determine ratios between values. If it takes one person 30 sec to complete a task and another person 20 sec, then the second person took only 2/3 as long. Scales of basic properties in physics are ratio scales. Psychological scales are not.

Functionally, the same statistical tests are used on both Interval and Ratio scale data.

Summary - Why is the scale important?

1. For data summary (descriptive statistics) and hypothesis testing (inferential statistics).

2. For the types of statements about the data which are valid.

For example, if you know the IQ of Betty is 140 and Tom's is 70, you can **not** say that Betty is twice as smart as Tom or even that Betty's IQ has a value that is twice Tom's. *These data are not on a ratio scale*.

III. Reliability

Reliability refers to the consistency of a measurement. If you repeat a measurement, do you get the same or a similar result?

A) Factors

Any time a measurement is made, there are these elements or components:

1. True value or score on measure
2. Measurement error
3. Measurement of other attribute(s) or qualities

In measuring human behavior, the measurement error can be substantial. For example, in a task that requires participants to sustain their attention to a conveyor belt to detect pills that are miss-colored or miss-shaped, performance is not constant for an individual. From moment to moment, their gaze may shift and they could miss a "reject" pill. Their performance may be different depending upon how tired they are before they start the task or on their mental state (in common language, they have something else on their mind).

Basically, the internal mental state of the participant is not constant and this shows up as variability in the measurement.

B) Measuring Reliability

If we assume that the psychological variable that we are measuring is stable over short intervals of time, we can establish the degree to which the measurement is reliable.

1. *Test-retest reliability*. We make the measurement at two different times (*test* and *retest*) with the same individuals. Are their scores similar (the same)?

2. *Internal consistency*. We examine each part of our measurement to see how it relates to other parts (*Cronbach's Alpha*), or we compare the scores on one half of the items with the other half (*split-half reliability*).

3. *Inter-rater reliability*. When a human must observe and score the participants' behavior, two (or more) observers are often used. We compare their scores and check that they are similar.

In these cases, reliability is assessed with the correlation coefficient. A high, positive coefficient means reliable.

Note – A measure can be reliable, but not valid (our next topic). The only requirement for reliability is that the measure be consistent (repeatable).

IV. Construct Validity

Does a scale really measure what it is supposed to measure?

The SAT is supposed to be a measure that predicts the ability to succeed in college. For this measure, the question of validity is whether it really does this. For the Beck Depression Inventory, the question of validity is whether it actually measures depression.

We will consider three aspects of construct validity:

Face Validity
Convergent and Discriminant Validity
Criterion Validity

A) Face validity

In a questionnaire or dependent measure, does the content of the question (measure) relate to the concept being investigated? For example, if one of the symptoms of depression is lethargy, a question that asked whether an individual feels like they have no energy would have face validity. Face validity is based on "logic" and some (minimal) theory of the behavior or psychological process being investigated.

In order to measure the speed of a mental process, we typically measure Reaction Time – how long a participant takes to make an observable response to an item.

However, a valid measure may have no face validity. Rapid eye movements while sleeping do not, "on the face of it", appear related to dreaming. However, they are a valid measure of when someone is dreaming. The validity was established in other ways.

B) Convergent and Discriminant Validity

Is a measure systematically related to other measures that it is supposed to be related to? Is it unrelated to measures that it is NOT supposed to be related to?

1. If it yields results similar to other measures that are related to it, it has *convergent validity*. For example, if a new measure of depression yields scores systematically related to the BDI (Beck Depression Inventory), this is convergent validity.

2. If it is unrelated to measures of different concepts that are supposed to be different, it has *discriminant validity*. If our new measure of depression does not correlate with scores on schizophrenia and other, unrelated disorders, it has discriminant validity.

C) Criterion Validity

If a measure is supposed to predict scores on some other measure, does it? To the extent that it does (accurately), it has criterion validity.

The SAT is supposed to predict success in college. Does it have criterion validity?

1. It has a test-retest *reliability* of 0.9, which is high.

2. It correlates about 0.35 with freshman year grades, which is **not** high. We would hope for a higher correlation for a measure to have good criterion validity. *The criterion validity is limited*.

Note here that this measure (SAT) has high reliability but moderate validity. Reliability and validity are not the same.

Why is the validity so apparently limited?

a) It misses the motivation/effort that a person puts into their work. Grades reflect this motivation/effort.

b) The questions cover only a part of intellectual ability that only partly overlaps with grades. For example, in striving for questions that can be objectively scored, aspects of intellectual ability such as the capability for organizing material are not tapped.

c) Some individuals get anxious and do not work well under time pressure. In college, some aspects of class work do not have the same time pressure.

V. Reactivity in the Measurement of Behavior

One factor that alters the behavior we measure is reactivity on the part of the participants. Does the act of measuring a participant's behavior alter what is being measured? When a person or animal "knows" that they are being observed, do they behave differently than if they were not aware of being observed?

A) Influence.

The influence of reactivity can be small or large. When large, it destroys the validity of the measurement. What we are measuring is the participants' behavior while being observed instead of the intended concept.

Reactivity takes different forms for different measures. For some questions, the alternative answers differ with respect to group norms. Is the behavior of a participant their "true" behavior, or one that conforms to group norms?

B) Approaches to Controlling Reactivity

1. Unobtrusive observation. If the participant does not know that they are being observed, then there should be no reactivity.

2. Participant observer. Participants know that they are being observed, but the observer is a part of their world. They are "used" to the observer.

3. Deception. A cover story is used so that the participant does not know what is really being measured.

4. Role playing. A situation is described and participants are asked to describe how they would act.

5. Simulation. Participants are actors in a simulated world (a more realistic version of role playing).

Each has advantages and problems:

For example, in a role play, participants might describe how they think they *should* act rather than what would actually happen. Or, they may not actually "know" how they would act. If this is the case, then a simulation may be more accurate at revealing "normal" behavior, but involves much more "work" to set up and run.

Using deception has ethical limits. It also requires a good "cover story" to limit reactivity.

Unobtrusive observation may invade privacy and violate the principle of autonomy (the right of participants to "run their own lives").

A participant observer may identify with the group that they are observing and their observations may not be as "objectively accurate". Conversely, because they are also a participant, they might make "more accurate" observations. That is, because they have been a part of the group, they know more and can make more detailed, reliable observations.

Answers for Chapter 4 Sample Questions

In a naturalistic observation version of assessing reading curricula, the teachers have chosen what curriculum to use. Thus, because the teachers (including their motivation for using the curriculum) differ between the two groups, these differences (ant not the curriculum) could lead to differences in reading scores. As an example, the teacher using the new curriculum may be more enthusiastic and this results in a more involved set of students who are more motivated to learn to read. Or, one teacher may routinely spend more time on reading than the other as a part of instruction.

Doing an experiment either requires that we use one teacher who does everything (except the curriculum for reading) the same for both classes or that we take two teachers and randomly assign them to classes and train them to do their teaching identically except for the reading curriculum. As later parts of the course will highlight, this is difficult to do.

Multiple choice:

1. – c; 2. – b; 3. – c; 4. – b

Sample Exam Questions for Chapter 5

1. It is legitimate to say that one person is twice as anxious as another if anxiety is measured on a: a) ratio scale b) interval scale c) ordinal scale d) all of the above

2. Test reliability as determined by correlating scores from the test taken by the same individuals at two different times is called: a) parallel forms reliability b) split half reliability c) inter-item reliability d) none of the above

3. If a scale actually represents the variable that it is supposed to represent, then we say that the scale has: a) consistency b) reliability c) construct validity d) a & b above

4. The mean of a set of values may be computed if the values were measured on a scale that is: a) ordinal b) interval c) ratio d) b & c above

Answers for Sample Questions from Chapter 5

1. – a; 2. – d (this is test-retest reliability);

3. – c; 4. - d