PLS 506
Mark T. Imperial, Ph.D.

Lecture Notes: Reliability & Validity

- Measurement & Variables
  - Initial step is to *conceptualize* and clarify the concepts embedded in a hypothesis or research question with precisely stated definitions
  - Concepts may be a single category (i.e., male) or several categories (i.e., gender).
  - Measurement assumes you can assign *different* values or categories to units of analysis. Because the concepts vary they are called *variables*.
  - Specific instances of a variable are called *indicators*
  - So we go from the abstract to the specific Concepts – variables – indicators
    - Education – level of education – to years of school
  - Indicators provide imperfect representations of variables for two reason
    - They often contain errors of classification
    - They rarely capture the meaning of a concept
  - The final step, *operationalization*, delineates the procedures for sorting units into categories
    - Ask people a question on a survey about the number of years in school

- Sources of measurement error
  - Potential sources in variation among cases when an operational definition is applied
    - True differences in the concept that you intend to measure. If you have a valid operational measure, this variation should account for most of the variation observed.
  - *Systematic measurement error*: results when factors systematically influence either the process of measurement or the concept being measured. Typically, it results in ratings or scores being consistently biased in one direction.
    - *Reactive measurement effect*: When the respondent's sensitivity or responsiveness to a measure is affected by the process of observation or measurement
    - *Social desirability effect*: people are less willing to admit to holding undesirable positions and attitudes when they are aware they are being "tested"
    - Respondents are more likely to agree than to disagree with statements irrespective of their content
    - Respondents will tend to give stereotyped responses such as answering the right-hand or left-hand response when a sequence of questions is asked in a similar format

  - *Random measurement error* is unrelated to the concept being measured. It is the result of temporary or chance factors.
    - Examples: upswings or downturns in a respondent's health, temporary variation in the administration of coding, momentary investigator fatigue. A bored, tired, or distracted respondent may provide erroneous responses.
    - Error is random because its presence, extent, and direction are unpredictable from one person to the next.
    - While they produce imprecise and inaccurate measurements affecting reliability, because they are unsystematic these random errors tend to cancel themselves out.
- Reliability and Validity
  - Reliability
    - *Reliability* is concerned with questions of stability and consistency
    - Is the operational definition measuring something consistently and dependably? Do repeated applications of the operational definition under similar conditions yield consistent results
  - Validity
    - *Validity* refers to the extent of matching or congruence or "goodness of fit" between an operational definition and the concept it is purported to measure
  - A highly unreliable measure cannot be valid. How can you get at what you want to measure if your results fluctuate wildly
  - A reliable measure may still be invalid. You can measure something very consistently but it is something other than what you intended to measure.

- How can you assess reliability?
  - *Reliability assessment* is a matter of checking for such consistency – either over time (when the same measurements are repeated) or over slightly different but equivalent measures (more than one indicator or more than one interviewer/observer/recorder is used)
  - *Test-retest reliability*: involves testing or measuring the same person or unit on two separate occasions and then looks at the correlation among the two measurements. Anything less than 0.8 for a correlation coefficient would be dangerously low. Ideally you will be near 1.00 (a perfectly reliable measure)
  - You can also estimate the equivalence among measurements made on the same occasion
    - *Parallel- or alternate forms procedure*: two alternate forms of a measure designed to be as similar as possible are administered successively to the same group of persons. The correlation between the scores on the two forms indicates the degree of reliability of either form taken separately. Advantages include:
      - Because the forms are different, the recall from the first form is unlikely to affect the second form
      - Without the danger of recall, the period of time between application of the forms can be reduced
      - When the forms are administered close together the chance of real change in the concept being measured is slight.

- *Slit-half method*: a scale or index measuring several items is applied to a sample of cases. After which, the items are divided into two halves randomly and each half is treated as a subtest with the result of the two subsets correlated to obtain an estimate of reliability.
  - Not having to worry about creating two equivalent forms is an advantage
  - Still assumes the existence of equivalent subsets of items
- *Internal consistency*: Research examines the relationships among all items simultaneously rather than arbitrarily splitting the items or comparing the results of parallel forms. The basic question is to what extent the measures are homogenous and measures the same concept. This can be measure using different statistical techniques.
- *Intercoder reliability*: Measures the extent to which different interviewers, observers, coders using the same instrument or measure get equivalent results. Various approaches to comparing how different interviewers, raters, coders, etc. are using the instrument and recording data.

- What can be done to improve reliability of your operational definitions
  - *Exploratory studies, pretests, preliminary interviews* of a small sample of persons similar in characteristics to the target group
  - *Adding items to the same type of scale* will usually increase reliability
  - *Item-by-item analysis* will reveal which items discriminate well between units with different values on a particular variable. Eliminate those items that do not discriminate well.
  - *Clear instructions*: the respondents may misinterpret the question or directions

- Four types of validity of interest to evaluation researchers
  - *Statistical conclusion validity*: the validity of inferences about the correlation (covariation) between treatment and outcome. (appropriate use of statistics to infer whether independent and dependent variables covary.
  - *Construct validity*: the validity of inferences about the higher order constructs that represent sampling particulars.
  - *Internal validity*: the validity of inferences about whether observed covariation between A (presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables are manipulated or measured. Does the covariation result from a causal relationship?
  - *External validity*: the validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables. To what populations can the effect from the samples be generalized?

- How can you assess validity?
  - Validity cannot be assessed directly. If it could, if we knew the case's true value, there would be no need for the measure.
  - Many of the threats to validity discussed below cannot be ruled out by experimental or quasi-experimental designs.
  - Researchers always need to be mindful and explore the potential role and influence of each threat given the particulars of a study and take steps to minimize these threats
    - How might the threat apply in this case?
    - Is there evidence that the threat is plausible rather than just possible?
    - Does the threat operate in the same direction as the observed effect so that it might explain the observations? Conversely, if it works in the opposite direction are my findings conservative?
- *Statistical Conclusion Validity* (related to causal inferences – statistical inferences)
  - Two related statistical inferences can affect the covariation component of causal inferences
    - Whether the presumed cause and effect covary
      - *Type I error*: You can incorrectly conclude that cause and effect covary when they do not
      - *Type II error*: incorrectly conclude they do not covary when in fact they do.
    - How strongly they covary
      - You can over- or under-estimate the size of the effect
      - You can over- or under-estimate the degree of confidence that we have in the magnitude of the effect
    - Conventional approach of many statistical tests is to report the probability that the result would have occurred by chance
      - If $p < .05$ the result is statistically significant
      - However, if $p > .05$ it could still be the case that A causes B (Type II)
      - Conversely, $p < .05$ and A many not cause B (Type I)
      - Statistical significance tells us little about the substantive significance (i.e., size of the effect)
      - Conversely, nonsignificance does not imply zero effect
      - When possible should report 95% confidence that the result is X ± Y
  - *Threats to statistical conclusion validity*: reasons why inferences about covariation between two variables may be incorrect
    - *Low statistical power*: power refers to the ability to test direct relationships that exist in the population. It is conventionally defined as the probability that a statistical test will reject the null when it is falls. Usually set at $\alpha = .05$. It can usually be improved through such techniques as:
      - Increasing sample size
      - Using equal cell sample sizes
      - Improving measurement
      - Increasing the strength or variability of the treatment
      - Use powerful statistical tests and ensure their assumptions are met

    - *Violated assumptions test statistics*: can lead to over- or underestimating the size and significance of the effect
    - *Fishing and error rate problem*: repeated tests for significant relationships, if uncorrected for the number of tests, can inflate statistical significance of any one test.
    - *Unreliability of measures*: measurement errors weaken the relationship between two variables and can increase or decrease the relationships between three or more variables
    - *Restriction of range*: reduced range of a variable will weaken the relationship between it and another variable
    - *Unreliability of treatment implementation*: if a treatment is intended to be implemented in a standardized manner is implemented only partially for some respondents it may underestimate the effects compared to full implementation
    - *Extraneous variance in the experimental setting*: some features of an experimental setting may inflate error, making the effect more difficult to detect.
    - *Heterogeneity of units*: increased variability on the outcome variable increases error variance making detection of a relationship more difficult
    - *Inaccurate effect size estimation*: some statistics systematically overestimate or underestimate the size of an effect
- ▪ *Internal validity* (related to causal inferences – does A cause B)
  - Refers to inferences about whether observed covariation between A and B reflects a causal relationship from A to B in the form in which the variables were manipulated and measured
    - Internal validity is necessary for experimental and quasi-experimental designs probing causal hypotheses
    - To support this inference the research must be able to show that A preceded B in time
    - A must covary with B (statistical conclusion validity)
    - No other explanations for the covariation are plausible
    - A preceding B is easy to control for in experiments but less so in non-experimental cross-sectional research
  - *Threats to internal validity*: reasons why inferences that the relationship between two variables is causal may be incorrect
    - *Ambiguous temporal precedence*: lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect
    - *Selection*: systematic differences in respondent characteristics could also cause the observed effect
    - *History*: Events occurring concurrently with the treatment could cause the observed effect
    - *Maturation*: Naturally occurring changes over time could be confused with a treatment effect
    - *Regression*: when units are selected for their extreme scores, they will often have less extreme scores on other variables, which might be confused with a treatment effect.
      - Need to explore anytime respondents are selected (or select themselves) because they have scores that are higher or lower than average

- *Attrition (experimental mortality)*: loss of respondents to treatment or to measurement that can produce artificial effects if the loss is correlated with important conditions
- *Testing*: exposure to a test can affect scores on subsequent exposures to that test, an occurrence that could be confused with a treatment effect
- *Instrumentation*: The nature of the measure may change over time or conditions in a way that could be confused with a treatment effect
- *Additive and interactive effects of threats to internal validity*: the impact of a threat can often exacerbate or depend on the levels of other threats.
  - *Selection-maturation*: additive effect when nonequivalent experimental groups formed at the start are maturing at different rates over time.
  - *Selection-history*: additive effect if nonequivalent groups from different settings with a unique history
  - *Selection-instrumentation*: additive effect if nonequivalent groups have different means on a test with unequal intervals along its distribution (a ceiling or floor for one group but not another)

- *Random assignment* of targets to treatment and control groups helps to minimize many of these threats to internal validity

- *Construct validity* (related to generalizations)
  - Construct validity involves making inferences from the sampling particulars of a study to the higher order constructs they represent.
    - Economist may be interested in the construct of unemployed, disadvantaged workers but the sample studied may actually be people who had income below the poverty level in the last 6 months or who participate in government welfare or food stamp programs.
    - Sometimes, important discrepancies exist between the sample and the construct
  - Construct validity is important because
    - Research cannot be done without constructs
    - Constructs connect the experiment to the theory to the communities that use the information
    - Construct labels carry social, political, and economic implications
    - Creation and defense of constructs is a fundamental task of science (e.g., periodic table, species taxonomies, etc.)
  - *Threats to construct validity*: reasons why inferences about the constructs that characterize study operations may be incorrect
    - *Inadequate explication of constructs*: failure to adequately explicate a construct can lead to incorrect inferences about the relationship between the operation and construct
      - Construct may be identified at too general a level
      - Construct may be too specific
      - Wrong construct may be identified – think you are measuring one concept when it really is another
      - Two or more constructs are described with only one construct

- *Construct confounding*: failure to measure all of the constructs may confound the results
  - For example, you may be interceded in the unemployed but this group may be predominantly uneducated and male – which construct is being measured?
- *Mono-operation bias*: Any one operationalization of a construct will underrepresent the construct of interest and measure irrelevant constructs complicating inferences
  - Often possible to use different measures of an outcome but many experiments only have one or two manipulations of an intervention and one setting
- *Mono-method bias*: when all operationalizations use the same method (e.g., self-reported data) that method is part of the construct being studied
- *Confounding constructs with levels of constructs*: inferences about constructs that best represent study operations may fail to describe the limited levels of the construct actually studied.
- *Treatment sensitive factorial structure*: the structure of a measure may change as a result of treatment, change that may be hidden if the same scoring is always used
  - Those exposed to a treatment test might begin to see the test in a different way than those not exposed
- *Reactive self-report changes*: self-reports can be affected by participant motivation to be in a treatment condition, motivation that can change after assignment is made
- *Reactivity to the experimental situation*: participant responses reflect not just treatments and measures but also participants' perceptions of the experimental situation. These perceptions are part of the treatment construct that is tested
- *Experimenter expectancies*: Experimenter can influence participant responses by conveying expectations about desirable responses and those expectations are actually part of the treatment construct being tested
- *Novelty and disruption effects*: participants may respond unusually well to a novel innovation or unusually poorly to one that disrupts their routine
- *Compensatory equalization*: when treatment provides desirable goods or services, compensatory goods or services may be provided to those not receiving the treatment, and this action must be included as part of the treatment construct description
- *Compensatory rivalry*: participants not receiving treatment may be motivated to show they can do as well as those receiving the treatment, and this should be included as part of the treatment construct description
- *Resentful demoralization*: participants not receiving the treatment may be so resentful or demoralized that they may respond more negatively than otherwise and this should be included as part of the treatment construct description
- *Treatment diffusion*: participants may receive services from a condition to which they were not assigned making construct descriptions of both conditions more difficult.

- *External validity* (related to generalizations – From sample to and across populations)
  - External validity concerns inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes that were or were not in the experiment. Targets for generalization can be
    - Narrow to broad
    - Broad to narrow
    - At a similar level
    - To a similar or different kind
    - Random sample to population members
  - *Threats to external validity*: reasons why inferences about how study results hold over variations in persons, settings, treatments, and outcomes may be incorrect
    - Interaction of the causal relationships with units: an effect found with certain kinds of units might not hold if other kinds of units had been studied
      - If most research is done on white males will it hold for females or minorities
      - Those successfully recruited to an experiment may be different than those who are not
    - *Interaction of the causal relationship over treatment variations*: an effect found with one treatment variation might not hold with other variations of that treatment or when that treatment is combined with other treatments or when only part of that treatment is used
      - Size or direction of a causal relationship varies over different treatment variations
      - Reduced class size may work when it is accompanied by funding to higher skilled teachers and build new schools but fail when students are taught in trailers by poorly trained teachers
    - *Interaction of the causal relationship with outcomes*: an effect found on one kind of observation may not hold if other outcome observations were used
      - In cancer research treatments may vary in effectiveness whether the outcome is quality of life, 5-year metastasis free survival, or overall survival
      - Sometimes treatments have positive effects on one outcome but different effects on other outcomes
      - Offset by examining different outcome measures
    - *Interactions of the causal relationship with settings*: an effect found in one kind of setting may not hold if other kinds of settings were used
      - Program may work in urban areas but not in rural areas or vice versa
      - Offset by examining sub-settings or using larger multisite studies
    - *Context-dependent mediation*: an explanatory mediator of a causal relationships in one context may not mediate in another context
      - One part of explanation is identifying mediating processes.
      - However, mediating processes that work in one setting may have different effects in other settings
  - Random sampling before randomly assigning to treatment and control groups simplifies external validity inferences.