Technical Adequacy of Assessments: Validity and Reliability

Dr. K. A. Korb

University of Jos

Importance of Good Measurement

- The conclusions of assessment (in a study) are only as good as the information (data) that is collected.
- The information (data) that is collected is only as good as the assessment (instrument) that collects the data.
- A poorly designed assessment (instrument) will lead to bad information (data), which will lead to bad conclusions.
- **Therefore, developing a good assessment (instrument) is a key part of good educational interventions (conducting a high quality research study).**

Considerations in Assessment Development

- **Error:** Anything that causes a student's score on the assessment to be an inaccurate representation of their true score
  - The assessment must be carefully developed so as to avoid error
- For formal assessments, allow plenty of time to revise, pilot test, re-revise, and re-re-revise the assessment

Considerations in Assessment Development

- The characteristics of the pupils taking the assessment should guide instrument development and language used
- Keep the assessment as short as possible while still including sufficient items to measure the key variables
- DIRECTLY measure key variables

Developing a Good Assessment

- Step 1: Identify other studies assessing the same variable or other assessments
  - Two reasons for identifying other studies:
    - Develop a construct definition of the variable.
    - Provide ideas on how each variable should be measured.
- Step 2: Develop a construct definition for each variable

Developing a Good Assessment

- Step 3: Operationalize the construct definition
  - *It is vital that the construct and operational definitions are clearly related*
  - Consider practical limitations in the variable definition when operationalizing the variable

Types of Assessments

- **Self-Report:** Participants report their own demographic characteristics, attitudes, beliefs, knowledge, feelings, and behavior
  - Can take the form of Questionnaire or Interview
- **Performance Assessment:** Directly assess performance on a contrived task
- **Observation:** Researchers observe participants' behavior
- **Checklist:** Identify the frequency or presence of behaviors or characteristics
- **Examination/Test:** Test participants' knowledge of a topic
- **Archival Data:** Collect information from existing records

Developing a Good Assessment

- Step 4: Choose an instrument to measure each variable (except an IV that is manipulated in an experimental design)
  - "Development of new tests is a complex and difficult process that requires considerable training in…psychological measurement. Therefore, we recommend that you make certain no suitable test is available before developing your own" (Gall, Gall, & Borg, 2003, p. 216).
- Advantages of using an already-developed instrument
  - Saves time and energy
  - Likely has already been well-validated
  - Connects your study to the entire body of research that uses that instrument

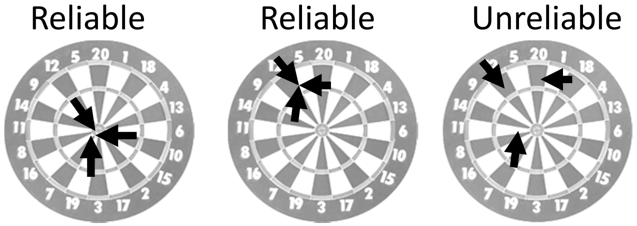
Adopting or Adapting an Assessment

- **Adopting:** Use the assessment nearly verbatim
- **Adapting:** Significantly alter the assessment
- Generally, adopting is preferable to adapting because reliability and validity studies still apply
- However, the instrument may not be applicable to your population, requiring adaptation

Developing a Good Assessment

- Step 5: Write a draft of the assessment
- Step 6: Revise the draft
  - Give the draft to other colleagues/experts to vet
- Step 7: Pilot the draft on a sample with similar characteristics to your population
  - Ask them to note places where they are unclear on the instrument
- Step 8: Revise, revise, revise, Re-Pilot, revise, revise, revise, Re-Pilot, revise, revise, Repeat

Reliability: Consistency of results



Reliability Theory

- Actual score on test = True score + Error
  - True Score: Hypothetical score on test
- The reliability coefficient indicates the ratio between the **true score** variance on the test and the **total** variance
  - As the error in testing decreases, reliability increases

Reliability: Sources of Error

- Error in test construction
- Error due to test construction with two or more forms of the instrument
- Error in test administration
- Error in test scoring

Error in Test Construction

- Results from items measuring more than one variable
- **Internal Consistency:** Measured by statistics looking at item consistency, e.g., Cronbach's alpha, split-half reliability
  - Correlate the items with each other
- Low internal consistency indicates poor test construction
  - Items are measuring more variables than they were designed to measure.
  - Solution: Revise the items to focus more directly on the key variable based on its definition

Error in Multiple Forms of the Instrument

- **Parallel Forms Reliability:** Determines the similarity of two different versions of the same instrument
- To calculate:
  - Administer the two tests to the same participants within a short period of time.
  - Correlate the test scores

Error in Test Administration

- **Test environment:** Room temperature, amount of light, noise, etc.
- **Test-taker variables:** Illness, amount of sleep, test anxiety, etc.
- **Examiner-related variables:** Absence of examiner, examiner's demeanor, etc.

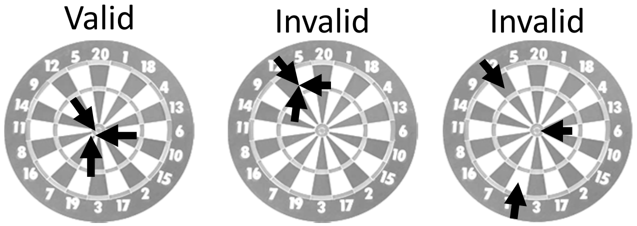
Error in Test Administration

- **Test-Retest Reliability**: Determines how much error in a test score is due to error in test administration.
- To calculate:
  - Administer the same test to the same participants on two different occasions
  - Correlate the two test scores

Error in Test Scoring

- When educators give subjective assessments different educators may give different scores to the same responses
- **Inter-Rater Reliability:** Determines how closely two different raters mark the assessment
- To calculate
  - Give the exact same test results from one test administration to two different raters.
  - Calculate the inter-rater reliability, generally with Cohen’s Kappa

Validity: Measuring what is supposed to be measured



Validity

- Three types of validity:
  - **Construct validity:** Measure the appropriate psychological construct
  - **Criterion validity:** Predict appropriate outcomes
  - **Content validity:** Adequate sampling of content

Construct Validity

- **Construct Validity**: Appropriateness of inferences drawn from test scores regarding an individual's status of the psychological construct of interest
- Two considerations:
  - Construct underrepresentation
  - Construct irrelevant variance

Construct Validity

- **Construct underrepresentation:** A test does not measure all of the important aspects of the construct.
  - Academic self efficacy may measure self efficacy only in math and science, ignoring other important academic subjects
- **Construct-irrelevant variance:** Test scores are affected by other unrelated processes
  - A mathematics test requires students to understand a language they are not familiar with

Criterion Validity

- **Criterion Validity**: Correlation between the measure and a criterion
- A criterion can be any standard with which the test should be related
  - Behavior
  - Other test scores
  - Ratings
  - Psychiatric diagnosis

Criterion Validity Example

Criterion Validity Evidence for New Science Reasoning Test: Correlations between Science Reasoning and Other Measures

	New Science Reasoning Test
WAEC Science Scores	.83
School Science Marks	.75
WAEC Writing scores	.34
WAEC Reading Scores	.24
Future marks in university science courses	.65

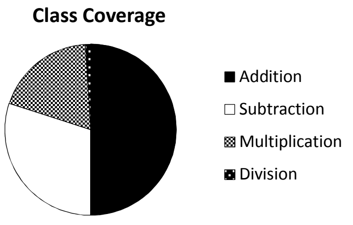
High correlations indicate good convergent validity.

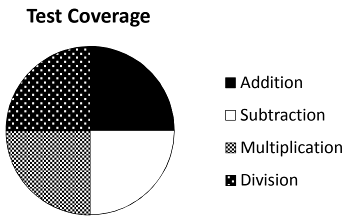
Low correlations indicate good divergent validity.

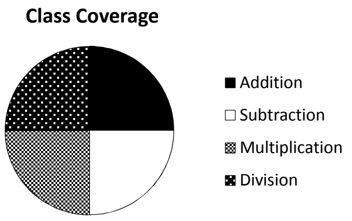
High correlation indicates good predictive validity.

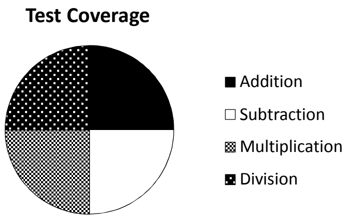
Content Validity

- **Content Validity:** Sampling the entire domain of the variable it was designed to measure
- To assess:
  - Gather a panel of judges
  - Give the judges a table of specifications of the content in the domain
  - Give the judges the instrument
  - Judges draw a conclusion as to whether the proportion of content covered on the instrument matches the proportion of content in the domain



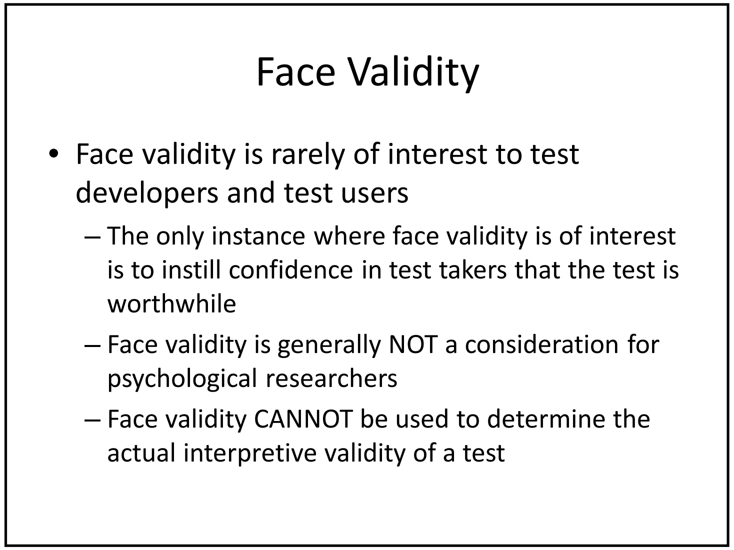






Face Validity

- **Face validity:** Whether the instrument appears to measure what it purports to measure
- To assess: Ask test users and test takers to evaluate whether the test appears to measure the construct of interest



Face Validity

- Face validity is rarely of interest to test developers and test users
  - The only instance where face validity is of interest is to instill confidence in test takers that the test is worthwhile
  - Face validity is generally NOT a consideration for psychological researchers
  - Face validity CANNOT be used to determine the actual interpretive validity of a test