MEASUREMENT ISSUES

There are many different ways to assess learning outcomes, but regardless of the type of procedure selected, all assessment should possess certain characteristics. The most important of these are reliability and validity.

Reliability

Next to validity, reliability is the most important characteristic of assessment results. Reliability provides the consistency that makes valid interpretations possible. It looks at issues related to stability and consistency of test scores over time, test administrations, test forms, and raters as well as homogeneity of items within an instrument. For example, if different faculty members obtain similar ratings on the same assessment task we can conclude that our results are reliable from rater to rater and if similar scores are obtained when the same assessment instrument or equivalent forms are used in a pre/post design we can conclude that our results are reliable across administrations and test forms. However, we cannot expect assessment results to be perfectly stable since there are many factors that may contribute to fluctuations. These factors contribute to measurement error, and methods for determining reliability essentially are a means for determining how much measurement error there is in our results. We want to minimize measurement error as much a possible. When making criterion-referenced interpretations (i.e., comparison to a fixed standard as opposed to relative standing) our desire for consistency of measurement is similar to norm-referenced interpretations (i.e., consistency across raters, task, time, forms); however, the focus is more often on whether the performance meets the standard than on the actual scores. For a more detailed discussion on how to estimate reliability, consult Linn and Miller (2005).

Tips to Increase Reliability

Test Length	Tests with more items have higher reliability assuming items are homogeneous
Time Limits	Increasing test time increases reliability; decreasing time between two test administrations of the same test or similar form increases reliability
Training Raters	Training raters increases consistency across raters

Crocker & Algina, 1986

Validity:The Most Important Consideration

Validity refers to the meaningfulness and appropriateness of the uses and interpretations to be made of assessment results and is considered the most important criteria when selecting an assessment procedure (Miller & Linn, 2005). There are many factors that may affect validity of interpretations and uses of assessment. These may include factors within the assessment itself, in the relationship between teaching and testing, in the administration and scoring of instruments, and in the nature of the group being assessed. A major goal in the construction, selection and use of assessment instruments is to control for those factors that will have the potential effect on validity and to interpret the results in accordance to what validity evidence is available. Presented below are some questions for evaluating assessment methods in light of validity considerations.

1) Does the content represent the construct that you are interested in assessing? Does the method of assessment align with your student outcomes and prompt students to represent the dimensions of learning desired? Are you measuring the content too narrowly leading to a narrow interpretation or are you measuring the content too broadly (e.g., measuring something more than the learning outcomes that you are looking for) ?

2) Will the assessment method elicit responses from the students that are consistent with the learning outcomes desired?

3) How do your assessment results compare to other measures like it? You would expect students scoring high on one criterion to score high on another criterion like it. You might use grades as a proxy but remember to interpret results carefully. Grades are not a flawless criterion as we have already have mentioned as they are lacking in the comprehensiveness and are contaminated by other factors.

Tips on Selecting Instruments

Look at the instrument's measurement properties
Has it been validated? Does it have good measurement properties?
Identify the kind of inferences that can be drawn
Determine its limitations and restrictions (i.e., will this work for your sample of students at this university?)