Reliability and Validity of Methods Used to Gather Evidence

Reliability and Validity narrow down the pool of possible summative and formative methods:

- **Reliability** – refers to the extent to which trial tests of a method with representative student populations fairly and consistently assess the expected traits or dimensions of student learning within the construct of that method.
- **Validity** – refers to the extent to which a method prompts students to represent the dimensions of learning desired. A valid method enables direct and accurate assessment of the learning described in outcome statements.

(Assessing for Learning: Building a sustainable commitment across the institution by Maki 2004)

### Reliability

Reliable measures can be counted on to produce consistent responses over time:

- Reliable data – variance in scores is attributable to actual differences in what is being measured, such as knowledge, performance or attitudes
- Unreliable data – score variance is due to measurement error; which can include such things as the individuals responding to the instrument, the administration and scoring of the instrument, and the instrument itself

(Assessment Essentials: planning, implementing, and improving assessment in higher education by Palomba and Banta 1999)

Sources of error in regard to reliability:

(i) from the student

- lack of motivation to take assessment seriously
- prior experience with being evaluated
- test anxiety, coaching, and physiological variables

(ii) from the assessment instrument

- test items ambiguously worded
- rating scale is confusing or vague
• representativeness of test items; e.g., if two test forms differ in their emphasis on program content, inconsistent scores may result from the conditions of assessment administration

(iii) • varying the style of test administration procedures produces varying scores

The greater the error in any assessment information, the less reliable it is, and the less likely it is to be useful.


Types of reliability:

(i) Stability – usually described in a test manual as test-retest reliability; if the same test is readministered to the same students within a short period of time, their scores should be highly similar, or stable

(ii) Equivalence – the degree of similarity of results among alternate forms of the same test; tests should have high levels of equivalence if different forms are offered

(iii) Homogeneity or internal consistency – the interrelatedness of the test items used to measure a given dimension of learning and development

(iv) Interrater reliability – the consistency with which raters evaluate a single performance of a given group of students


Barriers to establishing reliability include rater bias – the tendency to rate individuals or objects in an idiosyncratic way:

• central tendency – error in which an individual rates people or objects by using the middle of the scale
• leniency – error in which an individual rates people or objects by using the positive end of the scale
• severity – error in which an individual rates people or objects by using the negative end of the scale
• halo error – when a rater’s evaluation on one dimension of a scale (such as work quality) is influenced by his or her perceptions from another dimension (such as punctuality)
### Major Types of Reliability

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-retest reliability</td>
<td>A reliability estimate based on assessing a group of people twice and correlating the two scores. This coefficient measures score stability.</td>
</tr>
<tr>
<td>Parallel forms reliability (or alternate forms reliability)</td>
<td>A reliability estimate based on correlating scores collected using two versions of the procedure. This coefficient indicates score consistency across the alternative versions.</td>
</tr>
<tr>
<td>Inter-rater reliability</td>
<td>How well two or more raters agree when decisions are based on subjective judgments.</td>
</tr>
<tr>
<td>Internal consistency reliability</td>
<td>A reliability estimate based on how highly parts of a test correlate with each other.</td>
</tr>
<tr>
<td>Coefficient alpha</td>
<td>An internal consistency reliability estimate based on correlations among all items on a test.</td>
</tr>
<tr>
<td>Split-half reliability</td>
<td>An internal consistency reliability estimate based on correlating two scores, each calculated on half of a test.</td>
</tr>
</tbody>
</table>

### Validity

*Valid* measures on ones in which the instrument measures what we want it to measure:

- Construct-related validity – refers to the congruence between the meaning of the underlying construct and the items on the test or survey; i.e., do results correlate with other instruments examining the same construct?
- Criterion-related validity – includes predictive validity: how dependable is the relationship between the scores or answers on an instrument and a particular future outcome?
- Content-related validity – refers to the match between the content of the instrument and the content of the curriculum or other domain of interest

(Assessment Essentials: planning, implementing, and improving assessment in higher education by Palomba and Banta 1999)

*Validity* must be judged according to the application of each use of the method. The validity of an assessment method is never proved absolutely; it can only be supported by an accumulation of evidence from several categories. For any assessment methods to be used in decision making, the following categories should be considered:

- Content relevance and representativeness
  - The selected test should be a representative sample from those educational objectives which the test is supposed to measure
  - The test should cover what the program covered and should place emphasis in proportion to the program’s emphases
- Tests may be reliable but not valid for a particular program
- Internal test structure
  - Typically demonstrated through intercorrelations among items covering the same content domain
- External test structure
  - Necessary when the educator wishes to compare test scores or ratings with other measures or related variables
- Process of probing responses
  - Typically sought at two points during any test or scale construction: initially in the test construction to determine whether the students’ interpretations are consistent with the intent of the test designer; and at the point of probing the process to see if a pattern might be discovered on those students who scored very high or very low
- Test’s similarities and differences over time and across groups and settings
  - In studying validity evidence over time, some outcome measures should increase over time
- Value implications and social consequences
  - If a test or rating scale discriminates against certain groups of people, that test or scale should be considered suspect.


Validity:

- involves establishing that an assessment measures what it is supposed to measure
- can be thought of as the extent of the relationship between an assessment and the construct the assessment is supposed to predict

<table>
<thead>
<tr>
<th>Major Types of Validity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct validity</td>
<td>Construct validity is examined by testing predictions based on the theory (or construct) underlying the procedure. For example, faculty might predict that scores on a test that assesses knowledge of anthropological terms will increase as anthropology students progress in their major. We have more confidence in the test’s construct validity if predictions are empirically supported.</td>
</tr>
<tr>
<td>Criterion-related validity</td>
<td>Criterion-related validity indicates how well results predict a phenomenon of interest, and it is based on correlating assessment results with this criterion. For example, scores on an admissions test can be correlated with college GPA to demonstrate criterion-related validity.</td>
</tr>
<tr>
<td>Validity Type</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Face validity</td>
<td>Face validity is assessed by subjective evaluation of the measurement procedure. This evaluation may be made by test takers or by experts in what is being assessed.</td>
</tr>
<tr>
<td>Formative validity</td>
<td>Formative validity is how well an assessment procedure provides information that is useful for improving what is being assessed.</td>
</tr>
<tr>
<td>Sampling validity</td>
<td>Sampling validity is how well the procedure’s components, such as test items, reflect the full range of what is being assessed. For example, a valid test of content mastery should assess information across the entire content area, not just isolated segments.</td>
</tr>
</tbody>
</table>