Interpretation and Follow Up of Assessment Results

Undergraduate Medical Education Course Directors’ Retreat

Don A. Klinger
Associate Professor
Assessment and Evaluation
Faculty of Education
Overview

- Norm Referenced and Criterion Referenced Assessment
- Reliability and Validity
- Test and Item Analyses
The Purpose of Testing

- Measure student achievement of learning expectations.

The determination of achievement is relative.
Norm-referenced Assessment

- Achievement is measured relative to the performance of others.
- Historically, this is the most common form of assessment
  - Bell curve
  - Tests are designed to maximise the spread between students
Criterion-referenced Assessment

- Achievement is measured relative to a identified standard.
- Theoretically, this is now the most common form of assessment
  - Tests are designed to maximise our confidence that the student has mastered the intended learning.
Our Previous/Present Models

- The use of easy and difficult test items to better identify the stronger students.
- Testing a broad domain of skills.
- Pass scores based on pre-determined percentages.
Reliability

Reliability is all about consistency.

- If a student wrote parallel forms of the same test, how consistent would their scores be?
Reliability

Reliability is a technical procedure

- There are different techniques to conduct reliability analyses
  - Parallel forms (Gold Standard)
  - Cronbach’s Alpha (the most commonly used estimate of reliability)
- The result is a value between 0 and 1
  - For high-stakes decisions, >0.85
Random measurement error lowers reliability

- Test item quality
- Testing conditions
- Preceptors (markers)
- The test taker
Random Error
Validity

Validity is about Truthfulness.

- Are the interpretations we make about a student’s achievement correct?
Validity

Validity is an ongoing judgmental procedure.

- Does the test measure the intended construct?
  - Relevance and representativeness
- Does the test result in correct decisions?
Validity

Systematic measurement error lowers validity

- Poorly defined constructs
- Test/construct misfit
- Biases
- Construct irrelevant variance
Systematic Error
Our Reality

- Errors are built into the **instruments** we construct to measure what students know and are able to do.

- Errors are built into our **interpretation** of the information students provide us on these instruments.
The Fallibility of Assessment

Assessment is a human activity

<table>
<thead>
<tr>
<th>1. What a student really knows and is able to do</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. What assessment instrument you choose</td>
</tr>
<tr>
<td>3. How difficult you make the tasks</td>
</tr>
<tr>
<td>4. How you interpret what a student demonstrates</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>= A Student’s Score (mark)</td>
</tr>
</tbody>
</table>

Theories, policies, principles, standards and practice help us to reduce error in assessment.
## Contingency Tables
(Confusion Matrices)

<table>
<thead>
<tr>
<th></th>
<th>Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competent</td>
<td>60</td>
<td>5</td>
</tr>
<tr>
<td>Not Competent</td>
<td>15</td>
<td>20</td>
</tr>
</tbody>
</table>
Test and Item Analyses

- Item Analyses
  - Difficulty
    - $p$-values
    - mean scores
  - Discrimination (Validity)
    - Point biserial
    - Other indices
<table>
<thead>
<tr>
<th>Quest</th>
<th># Cor.</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>% Corr.</th>
<th>Diff.</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21</td>
<td>5</td>
<td>4</td>
<td>21</td>
<td>20</td>
<td>42</td>
<td>.42</td>
<td>0.18</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>1.0</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>2</td>
<td>0</td>
<td>40</td>
<td>8</td>
<td>80</td>
<td>.80</td>
<td>0.30</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>5</td>
<td>30</td>
<td>4</td>
<td>11</td>
<td>60</td>
<td>.60</td>
<td>0.06</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>1</td>
<td>25</td>
<td>18</td>
<td>6</td>
<td>50</td>
<td>.50</td>
<td>-0.10</td>
</tr>
</tbody>
</table>
Formulas used by the Medicine program

Difficulty = \frac{R1 + R2}{N1 + N2}

Validity = \frac{R1 - R2}{N1}