TEST ITEM ANALYSIS: I HAVE IT.... NOW WHAT DO I DO WITH IT???
Objectives

By the end of this presentation, participants will be able to:

- Interpret test analysis results to determine overall test performance.
- Interpret test analysis results to determine test item level of difficulty ($p$) and discrimination ($D$).
- Use test analysis results to determine the need for test item revision.
Item Analysis: What is it?

- Numerical method for analyzing test items.
  - Identifies distractors not doing what they are supposed to do: distract
  - Ideally suited for multiple-choice tests (not so well for essay or completion tests).
  - Best for norm-referenced tests (comparing students within a group, not against a criterion).
Item Analysis: Purpose

- Test student comprehension/knowledge of material as accurately as possible
- Improve the overall test
- Identify deficient test items to:
  - Improve them
  - Eliminate them
Test Reliability: Kuder-Richardson

- Estimate of test **internal consistency**
  - Would the same items test the same on different forms of the test?
  - Would the test yield the same or similar scores consistently (all things being equal).

- Reliability differs from **validity**. **Validity** refers to: the test measuring what it is supposed to measure (Does it address instructional objectives?)
Kuder-Richardson

- KR-20 formula considers:
  - Number of items on the test
  - Student performance on each test item
  - Variance between/among test items
Test Reliability:

- Chronbach’s alpha
  - Also known as coefficient alpha
  - Best measures surveys or attitude data (no right/wrong answer)
Step One: the test

- Analyze the test as a whole
- Review the KR-20
  - How reliable was the test?
- Look at the mean score
  - How high did the class score as a whole?
- Look to see how many questions everyone got right (100%)!
  - Was the test too easy? (More on that later).
KR-20: (need to know)

- **KR-20 range**: 0.00 – 1.00

- Values near 0.00 – measure many unknown factors, but not what you wanted to measure: score consistency.

- Values near 1.00 – closely measure a single factor – score consistency.

- **Take-Away**: A high KR-20 indicates **reliable (consistent)** student scores. A KR-20 score of **0.60 or higher** is desirable.
Let’s Practice: the test

100 point tests N = 70 students

<table>
<thead>
<tr>
<th></th>
<th>KR-20</th>
<th>Mean</th>
<th>Median</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>0.35</td>
<td>85.50</td>
<td>86.10</td>
<td>73-93</td>
</tr>
<tr>
<td>Test 2</td>
<td>0.64</td>
<td>82.68</td>
<td>83.78</td>
<td>65-93</td>
</tr>
<tr>
<td>Test 3</td>
<td>0.73</td>
<td>79.48</td>
<td>79.42</td>
<td>63-90</td>
</tr>
</tbody>
</table>

Which test is most reliable?
Step 2: Analyze test items

- Now that you’ve analyzed the test as a whole, you need to review individual test items.

- Two very important item analysis components:
  - Difficulty Index (p-value)
  - Discrimination Index (D) also known as point biserial
Item Analysis: Difficulty Index (p-value)

- $p = \frac{\text{total number of students selecting correct answer (24)}}{\text{total number of students who answered the item (30)}}$

Item #1          A  B  C*  D  
2  0  24  4

$p = \frac{24}{30} = 0.80$  (rule: the higher the $p$ value, the easier the test item)

What do you observe about this test item?

- Moderately easy (80% of students got it right)
- “B” needs to be modified or replaced
- “A” is suspect.
Let’s Practice

N = 25

<table>
<thead>
<tr>
<th>Item #</th>
<th># answering correctly</th>
<th>p value</th>
<th>Level of difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0.40</td>
<td></td>
</tr>
</tbody>
</table>
Item Difficulty: p-value (need to know)

- Range: 0.00 – 1.00

- Values near 0.00 (0.10 or 10%): difficult test item; more students answered the item incorrectly than correctly.

- Values near 1.00 (0.90 or 90%): easy test item; more students answered the item correctly than incorrectly.

Take-Away: Ideally, items should have p-values that range between 0.4 (40%) -0.6 (60%) (relatively difficult). Note: Oermann & Gaberson: 0.3 (30%) -0.7 (0.70%) for norm-referenced exams

Note: Very difficult and very easy items affect discrimination values ($D$) because they do not discriminate well (e.g. everyone gets the item correct or incorrect).
BUT......

- Does the p-value (item level of difficulty) tell you:
  - Whether or not distractors actually distracted those who did not know content?
  - Whether or not those who answered correctly were those who did know content?

- No, so now we must discriminate between the groups.
Item Discrimination: Point Biserial (D)

- The point biserial (D) measures the relationship between the student’s score (right/wrong) and the aggregate assessment score.

- Correlates with item difficulty (p-value):
  - Very easy/difficult items do not discriminate well.
  - Moderately difficulty items (p-values 0.4-0.6/0.3-0.7) generally discriminate very well.
Discrimination Index (D) (point biserial)

1. Arrange scores from highest to lowest
2. Divide into upper and lower groups (quartiles work) N=80; Q=20
3. Calculate the number in each group (upper, lower)

<table>
<thead>
<tr>
<th>Options</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>N=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper</td>
<td>15</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Lower</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

4. Compute D value:

\[
D = \frac{\text{Number correct (upper group)} - \text{number correct (lower group)}}{\text{Number in ether group (larger group number if unequal)}}
\]

\[
D = \frac{15 - 6}{20} = 0.45
\]

What does this mean? The D value is **positive**, which means more students who did well on the test overall also answered the item correctly. The higher the D value, the better the discrimination.
Point biserial (D-Value need to know)

- **Range**: -1.00 to +1.00

- Values **below 1.00** (negative values, e.g. – 0.10) mean higher scorers (e.g. upper quartile students) answered the item incorrectly more often than low scorers (e.g. lower quartile students).

- Values **above 1.00** (positive values) mean higher scorers answered the item correctly more often than low scorers (ideal)

- **Take-Away**: ideally, biserial should be +0.20 or greater. The higher the better!

*Note: Negative biserials likely indicate: miskey, ambiguity, guessing, information not covered/inadequately covered, etc. but definitely need review and most likely need revision.*
Let’s practice!

<table>
<thead>
<tr>
<th>Item</th>
<th>Total N= 70</th>
<th>Upper 27% N=19</th>
<th>Lower 27% N=19</th>
<th>Point Biseri al (D)</th>
<th>P-value</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>71.43%</td>
<td>94.74% (18)</td>
<td>52.63% (10)</td>
<td>18-10/19= 0.42</td>
<td>0.71</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>50*</td>
</tr>
<tr>
<td>2</td>
<td>41.43%</td>
<td>68.42% (13)</td>
<td>21.05% (4)</td>
<td>13-4/19= 0.47</td>
<td>0.41</td>
<td>15</td>
<td>25</td>
<td>1</td>
<td>29*</td>
</tr>
<tr>
<td>3</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>19-19/19= 0</td>
<td>1.0</td>
<td>0</td>
<td>70*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>44.29%</td>
<td>57.89% (11)</td>
<td>52.63% (10)</td>
<td>11-10/19= 0.05</td>
<td>0.44</td>
<td>39</td>
<td>0</td>
<td>31*</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>91.43%</td>
<td>89.74% (17)</td>
<td>94.74% (18)</td>
<td>17-18/19= -0.05</td>
<td>0.91</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>64*</td>
</tr>
</tbody>
</table>

\[ D = \frac{\text{#correct (upper)} - \text{#correct (lower)}}{\text{number in either group (N=19)}} \]
How do you correct tests & test items?

- **Increase test reliability:**
  - Increase the length of the test, e.g. 75 or 100 items vs. 25 or 50 items
  - **Manage item difficulty** (increase/decrease item difficulty if your items are consistently above p-value >.70 (too easy) or < 0.30 (too difficult)
  - **Manage item discrimination** (eliminate or revise p-value 1.0 items = 100% correct items); eliminate or revise negative point biserial items.
How do you correct tests & test items?

- Improve your test-writing skills! (That’s another workshop!)

- Not so easy but definitely possible
  - Revise non-distractors
  - Remove distracting content from the stem (e.g. long cases not relevant to the question)
  - Proof for spelling and grammar
How do you correct tests & test items?

- **Remove ambiguity from distractors** (run additional stats – upper quartile)
  - This item suggests ambiguity:
    - **Upper quartile**: A B C D*
      - 12 1 0 12

- **Remove guessing** (run additional stats – upper quartile)
  - This item suggests guessing:
    - **Upper quartile**: A B* C D
      - 6 6 7 5
Now that you know all this, let’s analyze a test!

- **Sample Test #1**  
  Test Items = 100  
  N = 67 students  
  KR-20 = 0.70  
  Mean = 79.48  
  p-Value Range = 0.16-0.70

<table>
<thead>
<tr>
<th>Item</th>
<th>Upper 25% N=17</th>
<th>Lower 25% N=17</th>
<th>P-Value</th>
<th>Point Biserial (D)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.66%</td>
<td>55.56%</td>
<td>0.68</td>
<td>0.35</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>46*</td>
</tr>
<tr>
<td>2</td>
<td>16.42%</td>
<td>22.22%</td>
<td>0.16</td>
<td>0.05</td>
<td>11*</td>
<td>32</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>68.66%</td>
<td>50.00%</td>
<td>0.68</td>
<td>0.29</td>
<td>13</td>
<td>46*</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>70.15%</td>
<td>83.33%</td>
<td>0.70</td>
<td>-0.17</td>
<td>47*</td>
<td>8</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>
Good, now let’s analyze another test!

Sample Test #2  # Test Items 100
N = 70 students  KR-20 = 0.30  Mean = 65.70
pValue Range = 0.30 – 0.96

<table>
<thead>
<tr>
<th>Item</th>
<th>Upper 25% (N=25)</th>
<th>Lower 25% (N=25)</th>
<th>P-Value</th>
<th>Point Biserial (D)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41.43%</td>
<td>21.05%</td>
<td>0.41</td>
<td>0.48</td>
<td>15</td>
<td>24</td>
<td>1</td>
<td>29*</td>
</tr>
<tr>
<td>2</td>
<td>31.48%</td>
<td>20.10%</td>
<td>0.30</td>
<td>0.12</td>
<td>21*</td>
<td>31</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>45.71%</td>
<td>31.58%</td>
<td>0.46</td>
<td>0.32</td>
<td>32*</td>
<td>15</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>94.74%</td>
<td>94.74%</td>
<td>0.96</td>
<td>0.02</td>
<td>67*</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
## Cheat sheet!

| P-value (item difficulty) | Point biserial (Item Discrimination, D-value) |  
|---------------------------|---------------------------------------------|---
| ≥ 0.70 (easy)             | D = neg value (e.g. - 0.20)                  | D = 0.00 to < +0.20               | D = > +0.20  
|                           | Requires review & revision (easy item; does not discriminate) | Consider Review (easy item; BUT it does it discriminate at 0.01 or 0.19???) | Keep & Use (easy item; high discrimination)  
| 0.30-0.70 (moderate)     | Requires review & revision (moderate item; does not discriminate) | Consider Review (moderate item; BUT it does it discriminate at 0.01 or 0.19???) | Keep & Use (moderate item; high discrimination)  
| ≤ 0.30 (difficult)       | Requires review & revision (difficulty item; does not discriminate) | Consider Review (difficult item; BUT it does it discriminate at 0.01 or 0.19???) | Keep & Use (difficult item; high discrimination)  

Objectives:

Hopefully, you can now:

- Interpret test analysis results to determine overall test performance.
- Interpret test analysis results to determine test item level of difficulty ($p$) and discrimination ($D$).
- Use test analysis results to determine the need for test item revision.
QUESTIONS???
greenig@graceland.edu
References