ITEM ANALYSIS FOR TEACHERS

MAKING SENSE OF ASSESSMENT RESULTS
DEFINITION

• Item analysis= The examination of individual items on a test, rather than the test as a whole, for its difficulty, appropriateness, relationship to the rest of the test, etc.

• Item analysis is useful in helping test designers determine which items to keep, modify, or discard on a given test; and how to finalize the score for a student

• If you improve the quality of the items on a test, you will improve the overall quality of the test – hence improving both reliability and validity
ITEM DIFFICULTY INDEX

• Item difficulty index = The proportion of test takers who answer an item correctly
• For maximizing validity and reliability, the optimal item difficulty level is 0.50
• However, this does not mean every item should have a difficulty level of 0.50, simply that the average of all items should be 0.50
The act of guessing on a test can cause some complications when a multiple choice test is used as the form of assessment; therefore, having different numbers of choices can influence the optimal mean p value.

<table>
<thead>
<tr>
<th>Number of Choices</th>
<th>Optimal Mean p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (i.e. true-false)</td>
<td>0.85</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
</tr>
<tr>
<td>4</td>
<td>0.74</td>
</tr>
<tr>
<td>5</td>
<td>0.69</td>
</tr>
<tr>
<td>Constructed response (essay)</td>
<td>0.50</td>
</tr>
</tbody>
</table>
SPECIAL ASSESSMENT SITUATIONS AND ITEM DIFFICULTY

- However, much of this is only true for norm-referenced tests
- For criterion-referenced tests, such as mastery tests, item difficulty is evaluated differently
- On mastery tests, $p$ values may be as high as 0.90, since student performance is a function of repeated attempts with feedback.
- Generally, students are expected to do much better than “chance” because they have been taught the material that is being assessed.
ITEM DISCRIMINATION

• Item discrimination = How well an item can discriminate among test takers who differ on the construct being measured by the test.

• Discrimination index = A calculation that determines the difference between those test takers who score well on a test and those who score poorly.

\[ D = \frac{T(%) - B(\%)}{(T+B) \times 0.5} \]

\[ T(27) = \text{proportion of examinees getting the item correct from the top group (e.g., top 27%) } \]

\[ B(27) = \text{proportion of examinees getting the item correct in the bottom groups (e.g., bottom 27%)} \]
GUIDELINES FOR EVALUATING D VALUES

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.40 and larger</td>
<td>Excellent</td>
</tr>
<tr>
<td>0.30-0.39</td>
<td>Good</td>
</tr>
<tr>
<td>0.11-0.29</td>
<td>Fair</td>
</tr>
<tr>
<td>0.00-0.10</td>
<td>Poor</td>
</tr>
<tr>
<td>Negative values</td>
<td>Miskeyed or other major flaw</td>
</tr>
</tbody>
</table>
ITEM-TOTAL CORRELATION COEFFICIENTS

- Item-total correlation = The correlation of performance on the test items (scored as either a 0 or 1) with the total test score
- This is usually calculated using a point-biserial correlation
- A large item-total correlation indicates that an item is measuring the same construct as the overall test measures
ITEM ANALYSIS OF SPEED TESTS

- Item analyses of speed tests are sometimes complicated by the fact that a majority of items are designed to be easy to be answered in a short amount of time.
- However, items later on the test should be somewhat more difficult as it is intended that less students will get them correct.
- There is no perfect calculation that takes these factors into consideration, however, they should be kept in mind when designing speed tests.
DISTRACTOR ANALYSIS

- Distracter= An incorrect alternative on a multiple choice item
- Because of time constraints, teachers designing tests usually only perform **distracter analysis** for items that need further scrutiny based on their $p$ or $D$ values
- A distracter analysis allows you to examine how many students in the top and bottom groups selected each option on a multiple choice item
- When examining each distracter, consider:
  - Did the distracter distract some students?
  - Did the distracter attract more examinees in the bottom group than the top group?
  - If “yes” for both questions, the option is a good distracter
PRACTICAL STRATEGIES FOR TEACHERS

• Most computers and scanning devices/programs provide item analyses
• However, if your school does not have access to such devices and programs, Website Reactions is a site that allows you to compute common item analysis statistics (www.surveyreaction.com/itemanalysis.asp)
PRACTICAL STRATEGIES FOR TEACHERS

• If you prefer to do item analysis by hand, you can separate students’ tests by grouping the top 10 (or 5 or 15) and the bottom 10 (or 5 or 15), and setting the middle group aside.

• Determine how many students in the top group answered an item correctly, and how many in the bottom group did.

• You can use this information to calculate the overall item difficult index \( (p) \), and separate item difficulty indexes, which can then be used to calculate the discrimination index \( (D) \).

• Item difficulty can also be calculated for the entire group of test takers (not just the top and bottom group).
USING ITEM ANALYSIS TO IMPROVE ITEMS

Consider the following questions based on your analyses:

- Is the item difficulty level appropriate for the testing application?
- Does the item discriminate adequately?
- Are the distractors performing adequately?
- What is your overall evaluation of the item?

Use your answers to these questions to decide which items to include, revise, or omit from a test.
ITEM ANALYSIS - EXAMPLES

• So, a test item may have an item difficulty of .70, and item discrimination of .40. This means that 70% of the test takers passed the item, and more students in the top group than the bottom group got the item correct.

• Another example – item difficulty of .40 and item discrimination of -.35. This means that only 50% of the test takers answered the item correctly, and more individuals in the bottom group (compared to the top group) answered in correctly. This might be a item that needs revision (may be ambiguously worded) or may have a miskeyed answer. Other explanations may also be possible.
ITEM ANALYSIS OF PERFORMANCE ASSESSMENTS

• For performance assessments that include one task, item analysis can typically not be done.

• However, if a performance assessment includes multiple tasks that receive scores on a numeric scale (i.e., from a rubric), item analysis may be done.

• You can compare how the top 10 students performed versus the bottom 10; if the average discrimination score is relatively high, it means that the item (or task) is discriminating between the two groups of students.
Along with performing equations for item analysis, you can set a test aside for a few days following its creation, and then proof it for errors.

- This “cool-off period” can help with the proofing process and identifying problematic test items.

Consider having a colleague review the test and give you feedback on it.

Ideally, you should incorporate both quantitative and qualitative item analysis into test development.
USING ITEM ANALYSIS TO IMPROVE CLASSROOM INSTRUCTION

• When a large number of students miss an item, this indicates that the content or construct it is assessing probably needs to be reviewed in class
• Distracter analysis can also help the teacher identify which misconceptions are shared by the majority of her students and correct them
• Discarding bad items from the student score may make the score more reliable and valid. It might also serve to improve student motivation.