ITEM ANALYSIS

I. BASICS

Item analysis = evaluation of individual questions

A. Types of item analysis

1. Qualitative
   - evaluating item in terms of content & form
   - content validity

2. Quantitative
   - evaluating items in terms of statistical properties
     - item difficulty
     - item discrimination

B. Use of item analysis
   - to build reliability & validity into a test
   - to “fine tune” a test

C. Number of items and test quality
   - more items = more reliable test
   IF the quality of the additional items is as high as those already on the test
   - low-quality items -> error variance
II. QUALITATIVE ITEM ANALYSIS
   - subjective process
   - to determine whether the domain is sampled adequately
   - systematically review content of domain & compare to test items
   - also consider form of the items

III. QUANTITATIVE ITEM ANALYSIS

A. Item difficulty
   - % of people who pass it
   EX item passed by 50% of people has p = .50

   Easier items placed at beginning of test
   - people with lesser abilities/knowledge will get credit for items they CAN pass
   - motivation/rapport

   Purpose of analyzing item difficulty
   => discriminate among people

   Most information = items at intermediate difficulty level
   - but also retain some items at extremes

   Score distribution (for each item) addresses this issue
   - normal distribution indicates adequate representation of difficulty levels

   Skewed distributions indicate problems
   - floor effect = a positive skew
   - ceiling effect = a negative skew
To fix:
- add items with difficulty levels in the region of the pile-up
- increases discrimination among individuals at this level of ability

Match difficulty level of items to examinees’ characteristics

B. Item discrimination
   = degree to which an item differentiates correctly among examinees
   
   Compare performance on an item to internal or external criteria
   Personality & interest inventories
      - no right/wrong answers
   - items compared to external criteria
   Achievement/aptitude -> internal criteria
   - the total test score
   - divide sample into 3 subgroups based on overall score
   - compare performance of the 2 extreme groups on each item
Table 1: Number of Individuals Giving Correct Response in Each Criterion Group

<table>
<thead>
<tr>
<th>Item</th>
<th>Upper (20)</th>
<th>Middle (20)</th>
<th>Lower (20)</th>
<th>Difficulty (U+M+L)</th>
<th>p (#correct / total #)</th>
<th>Discrimination (U - L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>9</td>
<td>7</td>
<td>31</td>
<td>.52</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>20</td>
<td>16</td>
<td>56</td>
<td>.93</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>18</td>
<td>9</td>
<td>46</td>
<td>.77</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>11</td>
<td>16</td>
<td>37</td>
<td>.62</td>
<td>-6</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>13</td>
<td>11</td>
<td>35</td>
<td>.58</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>14</td>
<td>9</td>
<td>39</td>
<td>.65</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>.08</td>
<td>5</td>
</tr>
</tbody>
</table>
Items passed more frequently by low-ability group than high-ability group are probably flawed.

Items with very high/low difficulty levels should be replaced.

Index of Discrimination ($D$)
- % passing in upper minus lower group

Table 2: Computation of Index of Discrimination

<table>
<thead>
<tr>
<th>Item</th>
<th>Percent Passing</th>
<th>Difference (Index of Discrimination)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upper Group</td>
<td>Lower Group</td>
</tr>
<tr>
<td>1</td>
<td>75</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>95</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>0</td>
</tr>
</tbody>
</table>

$D = 0$ indicates no discrimination at all on the part of an item

$D = 100$ -> maximum discrimination

Negative discrimination -> problems

Examine response pattern of the 2 groups to look for error source.
Table 3: Response analysis of individual items  
(Correct responses are in bold)

<table>
<thead>
<tr>
<th>Item</th>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Upper</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Lower</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Upper</td>
<td>0</td>
<td>10</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Lower</td>
<td>2</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>Upper</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Lower</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Upper</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Lower</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Use all of this information to strengthen efficiency of the individual items

CAVEAT: These same properties may not hold for a different sample
- wise to “cross-validate” a test on a second, independent sample