The Psychometric Principles
Maximizing the quality of assessment

• Reliability (freedom from error)
• Validity (‘... what is says on the tin’)
• Standardisation (compared with what?)
• Equivalence (is it biased?)

If we tabulate the marks given by the different examiners they will tend to be disposed after the fashion of a gendarme’s hat. I think it is intelligible to speak of the mean judgment of competent critics as the true judgment; and deviations from that mean as errors. This central figure which is, or may be supposed to be, assigned by the greatest number of equally competent judges, is to be regarded as the true value, just as the true weight of a body is determined by taking the mean of several discrepant measurements.

Reliability

Theory of True Scores

Francis Edgeworth, 1888

Reliability

• Sometimes called or ‘Latent Trait Theory’
• $X = T + E$
  • Where $X$ = Observed score
  • $T$ = True Score
  • $E$ = Error
• Latent Variable Analysis
Measuring reliability

- Reliability is reported as a positive correlation coefficient.
- The reliability of a score is a value between 0 and 1.
  - If zero, all is error, one is perfect accuracy.
- Can use it to:
  - Report the expected accuracy of our question or questionnaire
  - Improve the accuracy of our measure
  - Compare the accuracy of different forms of assessment
  - Assign a degree of confidence to a test result.
Reliability

- Inter-rater reliability
- Test – retest reliability (stability)
- Parallel forms reliability
- Split-half reliability
  - The Spearman-Brown Formula
- Cronbach’s alpha

Expected reliabilities

- Individual ability tests 0.92
- Group ability tests 0.85
- Personality scales 0.75
- Essays 0.66
- Creativity tests 0.50
- Projective tests 0.30
- Graphology/Astrology ?
Reliability and Validity

- Reliability is the extent to which a measurement is free from error

- Validity is the extent to which a measurement is measuring what it is purported to measure
Validity

- Face validity
- Content validity
- Predictive validity
- Construct validity
- Differential validity
- Consequential validity

Standardization

- Calculate means and standard deviation of norm group
- Provide norm table or conversion
  - Standard scores $z = (x - \text{mean})/\text{s.d.}$
  - Standardised scores
    - T-scores $= z*10 + 50$
    - Stanine $= z*2 + 5$ (min = 1, max = 9)
    - Sten $= z*2 + 5.5$ (min = 1, max = 10)
    - IQ format $= Z*15 + 100$
Equivalence

• Item bias
• Intrinsic test bias
• Extrinsic test bias
• Adverse impact
• Equivalence
• Differential Item Functioning

Constructing a psychometric test

• Defining the purpose
• Designing the blueprint
• The pilot study
• Item analysis
• Obtaining reliability and validity
• Writing the handbook
Developing the blueprint

- Curriculum based
  - Bloom’s taxonomy of educational objectives
- Job description
  - The job analysis
  - The person specification
- Theoretical
  - Ability
  - Personality

Knowledge Test Specification

<table>
<thead>
<tr>
<th>Manifestations</th>
<th>Arithmetic</th>
<th>Geometry</th>
<th>Algebra</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge of Terms (25%)</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Understanding (25%)</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Application (25%)</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Generalisation (25%)</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
Personality Test Specification

<table>
<thead>
<tr>
<th>Manifestations</th>
<th>High/Positive</th>
<th>High/Negative</th>
<th>Low/Positive</th>
<th>Low/Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Extraversion</td>
<td>Neuroticism</td>
<td>Detail</td>
<td>Tough-mindedness</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Classical item reduction

- Record form analysis
  - Non-responses
  - Altered items
  - Comments
- Delete extreme items
- Delete items with poor discrimination
- Retain the balance of the test
  - test specification
  - Positive and negative items
- Aim to reduce items by 50%
Software

- R
- Excel (with Real Statistic or other Add On)
- Stata
- Mplus
- SPSS
  - Analysis
    - Scale
      - Reliability analysis
        » Statistics
        » Item
        » Scale if item deleted

Item analysis

- p should be between 20% and 80%
- Adjusted Item-Total Correlation (point or rank biserial) > 0.2
- Remember the test specification!
Carmen

1. When on public transportation and in public places, I often start a conversation with people I do not know.
13. In a group or a meeting, I love it when everyone notices me.
25. I am an attractive person.
37. I am often told that I am charming.
49. I am often the centre of attention at business meetings.
61. I like to draw attention to the way I am dressed and my style.
73. I am not afraid to be the centre of attention.
85. I enjoy having a lot of people around me.
97. I seek out and love strong emotions.
109. I express my feelings and emotions very easily.

SPSS Reliability: Item-Total Statistics (Cronbach = 0.767)
Scale Mean if Item Deleted
Scale Variance if Item Deleted
Corrected Item-Total Correlation
Squared Multiple Correlation
Cronbach's Alpha if Item Deleted

<table>
<thead>
<tr>
<th></th>
<th>Scale Mean</th>
<th>Scale Variance</th>
<th>Corrected Item-Total Correlation</th>
<th>Squared Multiple Correlation</th>
<th>Cronbach's Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>21.63</td>
<td>15.946</td>
<td>.264</td>
<td>.108</td>
<td>.770</td>
</tr>
<tr>
<td>13.</td>
<td>21.47</td>
<td>14.668</td>
<td>.511</td>
<td>.359</td>
<td>.737</td>
</tr>
<tr>
<td>25.</td>
<td>21.06</td>
<td>15.536</td>
<td>.429</td>
<td>.247</td>
<td>.749</td>
</tr>
<tr>
<td>37.</td>
<td>21.17</td>
<td>14.986</td>
<td>.467</td>
<td>.262</td>
<td>.743</td>
</tr>
<tr>
<td>49.</td>
<td>21.74</td>
<td>15.052</td>
<td>.515</td>
<td>.343</td>
<td>.738</td>
</tr>
<tr>
<td>61.</td>
<td>21.76</td>
<td>14.813</td>
<td>.431</td>
<td>.243</td>
<td>.748</td>
</tr>
<tr>
<td>73.</td>
<td>21.19</td>
<td>14.152</td>
<td>.561</td>
<td>.394</td>
<td>.729</td>
</tr>
<tr>
<td>85.</td>
<td>21.13</td>
<td>15.072</td>
<td>.445</td>
<td>.223</td>
<td>.746</td>
</tr>
<tr>
<td>97.</td>
<td>21.35</td>
<td>15.226</td>
<td>.400</td>
<td>.217</td>
<td>.752</td>
</tr>
<tr>
<td>109.</td>
<td>21.25</td>
<td>15.519</td>
<td>.320</td>
<td>.160</td>
<td>.764</td>
</tr>
</tbody>
</table>
Writing the handbook

• Include copyright notice
• Include the scoring key and instructions
• Give evidence of reliability and validity
• Provide norms