SSRMC Psychometrics Module
Classical Test Theory and Practice

Professor John Rust

http://www.psychometrics.cam.ac.uk
The Psychometrics Centre
Judge Business School
Trumpington Street

The Psychometrics Centre

- Educational and diagnostic tests e.g. BAS-III, WISC, CELF
- Organisational e.g. Watson-Glaser, Orpheus/OBPI, Giotto
- Statistical, IRT, machine learning and AI techniques
- Software products e.g. Concerto, Apply Magic Sauce
- Web based assessment, ‘Discover My Profile’
- BPS Professional training courses
- Short courses, seminars and events (SEM in R, Mplus, ML)
- PhDs in psychometrics
- Tutorial materials on website - www.psychometrics.cam.ac.uk
The programme

1: Introduction to psychometrics (John Rust)
   • Psychometrics yesterday, today and tomorrow
   • How to design and build your own psychometric test

2: Testing in the online environment (Dr David Stillwell)
   • Testing via the internet. How to, plus do's and don'ts
   • Putting your test online

3. Modern Psychometrics (Dr Luning Sun)
   • How to build your own CAT using Concerto
   • Practical

4: Online Computer Adaptive Tests (Dr Aiden Loe)
   • Item Response Theory (IRT) models and their assumptions
   • Assessment and development of item parameters

Introduction to psychometrics

• Definition: The science of psychological assessment
• From IQ to AI: Past, present and future
• Intended and unintended consequences
• Behavioural Economics
• Online digital footprints
• ‘Psychology’ in Cyberspace
Psychometric Prediction

The 20th Century. Testing IQ with the best of intentions

MERITOCRACY
And sometimes the worst. Ellis Island

Eugenics and Dysgenics

- Virginia (1924) “Sterilization Act” for the “feebleminded”
- Germany (1933) “Law for the Prevention of Genetically Diseased Offspring” (sterilization for feeblemindedness, mental illness, blindness, deafness, physical deformity)
- Germany (1937) “Commission Number 3” compulsory sterilization of children of mixed-race origins.
- Germany (1939) Euthanasia introduced in psychiatric hospitals for those with disabilities (broadly defined, this included homosexuality and “social deviancy”)  
- Soviet Union (1949), Incarceration on psychiatric diagnosis of “philosophical intoxication”, “sluggish schizophrenia” (poor social adaptation), etc.
Herrnstein and Murray “The Bell Curve” (USA, 1994)

The Flynn Effect

<table>
<thead>
<tr>
<th>Wechsler Intelligence Scale for Children</th>
<th>Average at 1949</th>
<th>Rate of change p.a.</th>
</tr>
</thead>
<tbody>
<tr>
<td>WISC</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>WISC-R</td>
<td>107.63</td>
<td>0.311</td>
</tr>
<tr>
<td>WISC-III</td>
<td>113.00</td>
<td>0.322</td>
</tr>
<tr>
<td>WISC-IV</td>
<td>117.63</td>
<td>0.363</td>
</tr>
</tbody>
</table>
Openness

Openness-to-Experience
I am always on the lookout for new ideas to explore

Conventionalism

Conventionality
I stick with tried and tested ways of doing things

Conscientious

Detail conscious
If you get the details right everything will come out fine.

Disorganised

Seeing the bigger picture
I prefer to leave the boring details to my staff
Extraversion

I can be the life and soul of any party

Introversion

I have my best ideas when I am left alone

Agreeableness

I will always try and help my friends when they need me

Toughmindedness

I have a special ability to make tough decisions when needed
Neurotic

Emotional
I get anxious when I have to make difficult decisions

Unemotional
I am more able than most to cope with disaster

Stable

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.

The Theory of True Scores

• Sometimes called or ‘Latent Trait Theory’
• $X = T + E$
  • Where $X =$ Observed score
  • $T =$ True Score
  • $E =$ Error
• Latent Variable Analysis
• 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
• Concurrent validity
• Predictive validity
• Construct 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 = \( z \times 10 + 50 \)
    • Stanine = \( z \times 2 + 5 \) (min = 1, max = 9)
    • Sten = \( z \times 2 + 5.5 \) (min = 1, max = 10)
    • IQ format = \( Z \times 15 + 100 \)
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>Content areas</th>
<th>Arithmetic</th>
<th>Geometry</th>
<th>Algebra</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge of Terms (25%)</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Understanding (25%)</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Application (25%)</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Generalisation (25%)</td>
<td></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>Extraversion</th>
<th>Neuroticism</th>
<th>Detail</th>
<th>Tough-mindedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>High/Positive</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>High/Negative</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Low/Positive</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Low/Negative</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!
TD12 Subscale 6

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)

<table>
<thead>
<tr>
<th>Item</th>
<th>Scale Mean if Item Deleted</th>
<th>Scale Variance if Item Deleted</th>
<th>Corrected Item-Total Correlation</th>
<th>Squared Multiple Correlation</th>
<th>Cronbach's Alpha if Item Deleted</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