Social Science Research Methods
Module 12: Psychometrics

PROFESSOR JOHN RUST
HTTP://WWW.PSYCHOMETRICS.CAM.AC.UK
THE PSYCHOMETRICS CENTRE
JUDGE BUSINESS SCHOOL, 17 MILL LANE
The Psychometric Centre

Education and diagnostic tests e.g. BAS-III, WISC, CELF
Organisational e.g. Watson-Glaser, Orpheus, Giotto
Statistical, IRT and AI techniques
Software products e.g. Concerto, Apply Magic Sauce
Web based assessment, ‘Testing Website’
BPS Certificate in Test Use courses
Short courses, seminars and events (SEM in R, Mplus)
PhDs in psychometrics or related areas
Tutorial materials on website
  www.psychometrics.cam.ac.uk
What is psychometrics?

- The science of psychological assessment
- Brief history
- Theory of True Scores
- The Psychometric Principles
- Types of psychometric test
- Test construction exercise
The first examinations
Measuring candidates’ ability in order to determine their position in court. Xun Zi, 310BC - 289BC
‘Mental Age’ Binet, Paris, 1904
‘Mental Tests’ US Army Alpha, WW1
Scholastic Aptitude (11+, SAT)
IQ testing, Ellis Island 1912-1954
What is psychometrics today?

- In research
- Much assessment is “high stakes”
  - Recruitment and staff development
  - Licensing and chartering (e.g. Accountants, Surgeons)
  - School and University examinations
  - Psychiatric and ‘special needs’ diagnosis
  - Credit ratings, risk analysis
- Online digital footprint analysis
  - Marketing
  - Forensics
  - Robotics
The theory of true scores

- Whatever precautions have been taken to secure unity of standard, there will occur a certain divergence between the verdicts of competent examiners.

- 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.

What can be measured?

- length, blood pressure, knowledge, desire, intelligence
- “Temperature is what thermometers measure”
- Measurements, decisions, the umpire, judgements, competitions …. 
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
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?)

Types of assessment

- First impressions
- Application forms and references
- Objective tests (on or off line)
- Projective tests
- Interviews
- Essays and examinations
- Research questionnaires
Constructing a psychometric test

- Defining the purpose
- Designing the blueprint
- The pilot study
- Item analysis
- Obtaining reliability and validity
- Writing the handbook
Appropriate latent traits

- Personality rather than ability – e.g. “Tendency to feel ....”
- Single traits with about 32 pilot items
- Reduce to 20 through item analysis
- Examples: Apathy, Bewilderment, Collywobbles, Disgruntlement, Embarrassment, feeling like a Fraud, Guilt, Hatred, Humility, Indignation, Jealousy, Loneliness, Love, Morbid Curiosity, Nostalgia, Overwhelmed, Perversity, Road Rage, Self-Pity, Triumph, uncertainty, Vengefulness, Wanderlust
Designing the blueprint

- **Curriculum based**
  - Bloom’s taxonomy of educational objectives

- **Job description**
  - The job analysis
  - The person specification

- **Theoretical**
  - Ability
  - Personality
## Knowledge Test Specification

### Content areas

<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>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>
What can be wrong with multiple choice items?

- Language
  - appropriateness
  - understandability
  - discontinuity
- Order
- Distractors
  - distractor analysis
- Bias
The pilot study

- Pre-piloting
  - Are the correct items correct?
  - Are the distractors incorrect?
  - Are any items offensive or likely to be biased?
- The sample and sample size
- Data collection
- Data entry
Point biserial correlation

- $P_{bi} = ((x - \text{mean}) \sqrt{p(1-p)})/s.d.$

- Contribution of intercorrelations to the total variance
  - From the item variances
  - From the Item covariances
Using difficulty (p) and discrimination (r) indices:

- p should be between .2 and .8
- r should be above approx .2
- Remember the test specification!
Software

- R
- Mplus
- SPSS
  - Analysis
    - Scale
      - Reliability analysis
        - Statistics
        - Item
        - Scale if item deleted
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%
Reliability

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

- WAIS 0.92
- Ability tests 0.82
- Personality test 0.75
- Essay marks 0.60
- Supervisors ratings 0.40
- Projective tests 0.30
- Creativity tests 0.25
- Astrology/graphology ?
Validity

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

- Calculate means and standard deviation of norm group
- Provide norm table or conversion
  - Standard scores $z = (x - \text{mean})/s.d.$
  - 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 = $z \times 15 + 100$
Freedom from bias

- Item bias
- Intrinsic test bias
- Extrinsic test bias
- Adverse impact
- Equivalence
- Differential Item Functioning
Writing the handbook

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