The Psychometric Principles
Maximizing the quality of assessment

- Reliability
- Validity
- Standardisation
- Equivalence
What can be measured?

• Length, blood pressure, knowledge, desire, intelligence
• “Temperature is what thermometers measure”
• Measurements, decisions, the umpire, judgements, competitions, awards ….

Psychometrics as measurement

• Reliability is the extent to which a measurement is free from error.
• “If anything exists it must exist in some quantity and can therefore be measured”. (Lord Kelvin 1824, 1907)
• In 1900, Lord Kelvin claimed “There is nothing new to be discovered in physics now. All that remains is more and more precise measurement.”
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.

The Theory of True Scores

• Charles Spearman (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-293. If we have two measures of the same characteristic we can estimate true values. The accuracy of this estimation is called its reliability.
• Melvin Novik, Frederick Lord And Allan Birnbaum, used Classical Test theory to derive Latent Trait Theory, the fundamental building block of Item Response Theory and Rasch. Ref: Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. 
Measuring reliability

• The reliability of a score is a value between 0 and 1. If zero, all is error, 1 is perfect accuracy. Once we have an estimate of reliability we can use it to:
  • 1. Compare different forms of assessment
  • 2. Assign confidence to a test result.

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 0.32
• Graphology/astrology ?
Using reliability

• Reliability gives us the standard error of measurement:

• Standard Error of Measurement = \( S \times \sqrt{1-r} \)
  • where \( S \) = standard deviation of test scores
  • and \( r \) = reliability

Example

• Emma obtains a mark of 67 on her final year essay. Assuming the reliability of essays is 0.66 and a standard deviation of 10, the standard error of measurement is \( 10 \times \sqrt{1-0.66} \), which is approximately 6.

• The 95% confidence interval is this value ± 1.96, = approx 12

• The 95% confidence interval of her mark is 67 ± 12. That is, her ‘true score’ could be anything between 55 and 79
More uses for reliability

• Spearman Brown Prophesy Formula
• New reliability = \( n \times r / 1 + (n-1)r \)
• Where \( n \) = ratio by which test length has changed
  • \( r \) = old reliability

Example

• If Emma completed 3 essays as part of her examination paper in a single subject, then the new reliability = \( 3 \times 0.66 / (1 + (3-1) \times 0.66) = 0.85 \).
• This gives a confidence interval of 67 ± 8 i.e. from 59 to 75
Forms of validity

- Face validity
- Content validity
- Predictive validity
- Concurrent validity
- Criterion related validity
- Construct validity

Face Validity

- Appropriateness
- Relevance
- Fairness
- Face validity for the candidate AND client
- Face reliability
Content validity

- The extent to which the content of the test matches the content of the:
  - Job description
  - Person specification
  - Curriculum

Test specification

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>Bloom Taxonomy</td>
</tr>
<tr>
<td>25%</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>Knowledge</td>
</tr>
<tr>
<td>25%</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>Understanding</td>
</tr>
<tr>
<td>25%</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>Application</td>
</tr>
<tr>
<td>25%</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>Generalisation</td>
</tr>
</tbody>
</table>
Concurrent Validity

• Does the test measure the same thing as other tests that also purport to measure it?
• Concurrent validity as differential validity
• Multitrait-multimethod approach (Campbell & Fiske): 3 or more traits assessed by 3 or more methods
• Convergent validity (concurrent)
• Discriminant validity

Differential Validity

• Does the test measure the trait it purports to measure?
• ‘Anxiety’ but not ‘Depression’
• Potential’ but not ‘Ability’
• ‘Critical Thinking’ but not ‘Intelligence’
• ‘Conscientiousness’ but not ‘Impression Management’
The Multitrait-Multimethod Technique

<table>
<thead>
<tr>
<th></th>
<th>Self-report</th>
<th>360°</th>
<th>Projective</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E  N  C</td>
<td>E  N  C</td>
<td>E  N  C</td>
</tr>
<tr>
<td>SR-E</td>
<td>1 -0.15 0.32</td>
<td>0.65 -0.04 0.23</td>
<td>0.46 -0.56 0.13</td>
</tr>
<tr>
<td>SR-N</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SR-C</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>360-E</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>360-N</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>360-C</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>P-E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-N</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>P-C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Criterion-related validity

- Does the test predict success on a criterion
  - E.G Are students with three straight A’s at A level more likely to become successful doctors? I.e. Do they ‘do better’:
    - (a) In their medical school exams?
    - (b) as doctors?
Predictive validity

- Validates the test against its ability to predict
- Behaviour
- Motivation
- Success
- Potential

Accuracy of Predictors
Construct validity

- Constructs (e.g. Intelligence, Justice)
- Definitions
- Networks of associated ideas e.g. Biological Basis of personality
  - Arousal
  - Brain structure
  - Mental Illness
  - Conditioning
  - Sensory deprivation

Three types of standard

1. Criterion referenced – what can a person with this score be expected to do or know how to do?
2. Norm referenced – compare with others
3. Ipsative – strengths and weaknesses (or training needs)
The normal distribution

The standard score (z score)

- \( z = \frac{\text{raw score} - \text{mean}}{\text{standard deviation}} \)
- Scores range between -3 and +3 with a mean of zero
- Eg for a set of scores with a mean of 60 and a standard deviation of 6, what is the z score of persons with raw scores of:
  - 60?, 66?, 54?, 69?
- Percentiles are obtainable from z tables
Standardised scores

- T scores = z*10 + 50
- Stanine scores = z*2 + 5
- Sten scores = z*2 + 5.5
- IQ format scores = z*15 + 100
- ‘A’ Level grades?

Bias and ‘offensiveness’
How are tests perceived?

• The predictive model
• The competition model
• The examinations model
• Popular conceptions of bias

The correction for guessing

• Corrected Score = R – W/(N-1)
• Where R = number correct
• W = number incorrect
• N = number of response options
• (in True/False, N=2)

<table>
<thead>
<tr>
<th>Raw</th>
<th>R</th>
<th>W</th>
<th>Corrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>75</td>
<td>75</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>47</td>
<td>47</td>
<td>53</td>
<td>0</td>
</tr>
</tbody>
</table>
Equivalence (bias)

- Differential Item Functioning (DIF)
  - Item bias
- Test Equivalence
  - Intrinsic test bias
- Adverse Impact
  - Extrinsic test bias
  - Cultural insensitivity

Item bias

- Second languages
- Dialects within a language
- Language subsets
- Pictorial forms
- Puzzles
## Testing for item bias

(using difficulty values)

<table>
<thead>
<tr>
<th>Item number</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>.86</td>
<td>.87</td>
</tr>
<tr>
<td>29*</td>
<td>.75</td>
<td>.35</td>
</tr>
<tr>
<td>3</td>
<td>.72</td>
<td>.74</td>
</tr>
<tr>
<td>48</td>
<td>.68</td>
<td>.59</td>
</tr>
<tr>
<td>15</td>
<td>.61</td>
<td>.55</td>
</tr>
<tr>
<td>9*</td>
<td>.45</td>
<td>.60</td>
</tr>
</tbody>
</table>

## Example of US case law

- All bilingual children must be tested in their primary language.
- Unfair verbal items should not be used.
- Currently enrolled bilingual children to be retested.
- State psychologists to develop tests for Mexican American children, with appropriate items and their own norms.
- Any school district with a disparity must submit an explanation.
Translinguistic and transcultural equivalence

- Obtained by:
  - Translation and back translation
  - Focus groups
  - Cognitive interviews

Intrinsic Test Bias 1

- The predictive validity model allows us to predict a candidate's success from their test score using regression.
- But suppose this regression equation is different between two groups.
Intrinsic Test Bias 2

- This is a statistical model for positive discrimination
- But do psychometricians agree on the procedures? No
  - Cleary \( y = \alpha + \beta x \)
  - Einhorn and Bass \( y = \alpha + \beta x + \varepsilon \)
  - And others .....
Example of US case law

- Bakke vs the Regents of the University of California Medical School at Davis.
- 1977 California Supreme Court ruled that positive discrimination on grounds of race violated the equal protection provision of the US constitution
  - 1978 US Supreme Court also ruled by 5 to 4
  - Court upheld affirmative action provided race was not involved.

UK Equal Opportunities Legislation

- Sex Discrimination Act (1975)
- Race Relations Act (1976)
- Data protection Act (1984)
- Disabilities Act (1998)
  - Sexual orientation
  - Religion and belief
Conclusions

• The four psychometric principles can be used
• To evaluate an assessment
• To improve an assessment
• To establish degrees of confidence
• To address issues of inequality
• To improve efficiency