Lecture 4: Psychometric Principles

Professor John Rust
University of Cambridge
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?)

- Rust, J. & Golombok, S. (2009) Modern Psychometrics
The reliability of measurements

• Carl Friedrich Gauss (1777-1855) derived the equation for the normal distribution to deal with error in the measurement of the exact location of stars in the firmament.

• Reliability is the extent to which a measurement is free from error.
The Average Man

• Adolphe Quetelet (1835) “On Man and the Development of His Faculties” invented the science of Social Physics and the idea of the “average man”

• “O sad condition of the human species! The toll of the prisons, of the chains and of the scaffold seems fixed with as much probability as the revenue of the state. We are able to enumerate in advance, how many individuals will stain their hands with the blood of their fellows, how many will be forgers, how many prisoners, nearly as one is able to enumerate beforehand the births and deaths which must take place.”
Departures from the average

- Francis Galton’s (1888) Co-relations and Their Measurements, Chiefly from Anthropometric Data, *Proceedings of the Royal Society* 45, 135-145. Represented a systematic shift from the study of the average man to the study of differences from the average.

- “The law of deviation from an average shows that the number per million whose heights range between any limits we please to name can be predicted from the previous datum of the average.”

- Galton proposed the ‘standard deviation’ as a unit or measurement for any quality, including intelligence, that had a normal distribution.
The evolution of the Latent Trait

• Charles Spearman (1904) General intelligence objectively determined and measured, *American Journal of Psychology*, 15, 201-293. With two measures of the same characteristic we can estimate true values. The accuracy of this estimation is called its reliability.

• Melvin Novick and Frederick Lord (1968) “Statistical theories of mental test scores” use Classical Test Theory to derive Latent Trait Theory. Allan Birnbaum, in his supplement, established Item Response Theory of which Rasch Scaling is a special case.

• Today Latent Variable Analysis (LVA) is an integral part of statistical modelling in Psychometrics, Econometrics and Statistics.
What can be measured?

- length, blood pressure, knowledge, desire, intelligence
- “Temperature is what thermometers measure”
- Measurements, decisions, the umpire, judgements, competitions ....
Forms of reliability

• Inter-rater Reliability
• Test-Retest Reliability
• Parallel Forms Reliability
• Split half Reliability
• Cronbach’s $\alpha$ (alpha)
Measuring reliability

- 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.
## Expected reliabilities

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual ability tests</td>
<td>0.92</td>
</tr>
<tr>
<td>Group ability tests</td>
<td>0.85</td>
</tr>
<tr>
<td>Personality scales</td>
<td>0.75</td>
</tr>
<tr>
<td>Essays</td>
<td>0.66</td>
</tr>
<tr>
<td>Creativity tests</td>
<td>0.50</td>
</tr>
<tr>
<td>Projective tests</td>
<td>0.30</td>
</tr>
<tr>
<td>Graphology/Astrology</td>
<td>?</td>
</tr>
</tbody>
</table>
Standard Error of Measurement (SEM)

• Given the reliability of a test and its standard deviation, we can use the Theory of True Scores to obtain an estimate of the size of the error.

• The Standard Error of Measurement (SEM)

• \[ \text{SEM} = \text{SD} \times \sqrt{1 - r} \]

  – where SD = the standard deviation
  – and \( r \) = reliability
The Confidence Interval

• Once we have the standard error of measurement, we can obtain the confidence interval associated with a particular score. This procedure turns a point score into a range of scores in which we can have a particular degree of confidence that the true score lies.

• I.e. Instead of saying “Peter’s score on the test was 77” we might say “We have 95% confidence that Peter’s score lies between 72 and 82”

• A 95% confidence interval means there is still a 1 in 20 chance that we are wrong.
Example 1

- Peter’s IQ score on the WISC-IV was 105, while his sister Margaret’s score was 111. How much confidence should we have that Margaret has more academic potential than Peter? The reliability of the WISC is 0.95 and the standard deviation is 15, the Standard Error of Measurement (SEM) is therefore $15 \times \sqrt{1 - 0.95}$, which is $15 \times \sqrt{0.05} = 15 \times 0.22 = 3.35$.

- The SEM has a normal distribution. The 95% confidence interval is $\pm$ SEM*1.96, in this case $3.35 \times 1.96 = 6.57$, i.e. approximately 7.

- Hence the 95% confidence interval for the IQs are:
  - Peter: 98 to 112
  - Margaret: 104 to 118

- Clearly some caution required.
Example 2

• Emma obtains a mark of 67 on her final year essay. Assuming the reliability of essays is 0.66 and the standard deviation is 10, the standard error of measurement is $10 \sqrt{1-0.66}$, which is approximately 7.

• The 95% confidence interval is this value $\pm 1.96$, = approx 14

• The 95% confidence interval of her mark is 67 $\pm$ 14. That is, her ‘true score’ could be anything between 53 and 81

• What!!!
Changing test length

• The Spearman Brown Prophesy Formula:
  • making the test longer (or shorter)

• The New reliability = \( \frac{n*r}{1+(n-1)r} \). Where:
  • \( n \) = ratio by which test length has changed
    – \( (N = 2 \) means double length, \( N =0.5 \) means length halved)
  • \( r \) = old reliability
Example 3

• If Emma completed 9 essays, then the new reliability = \( \frac{9 \times 0.66}{1 + (9-1) \times 0.66} \) = 0.95.

• The SEM is \( 10 \times \sqrt{1-0.95} \) = 2.3

• This gives a confidence interval of 67 ± 2 i.e. between 65 and 69

• With her final degree result dependent on several papers, the level of confidence in the accuracy of her degree class increases.
CTT vs IRT

• Classical Test Theory assumes that the reliability of a test is a fixed value.
• Yet we know that it depends on the sample from which the data is drawn
• Also the reliability is not necessarily the same at different points of the scale.
• In diagnostic terms, it should be optimal at the threshold.
• IRT methods provide an ‘Information Function’
Tests and Items

• There is a continuum between broad traits (eg “g”), narrow traits and subtests (eg “Arithmetic”, “Vocabulary”) and test items

• Some tests are very short (eg 4 items)

• Reliability is a function of test length, but what about single items. Can they be reliable or unreliable?
Types of scale

• Revision of Measurement Theory.
• There are four types of scale:
  – Nominal e.g. gender
  – Ordinal e.g. Ranking of candidates
  – Interval e.g. Essay marks or test scores
  – Ratio e.g. length or temperature
Types of questionnaire response

- Categorical: England, Wales, Scotland; Gender
- Nominal: True, False; Right, Wrong
- Ordinal: Always, Often, Seldom, Never
- Interval: Test scores (where test is long)

• While test scores may be treated as interval, test items are either nominal or ordinal
Types of dichotomy

• Nominal dichotomy
  – E.g. There were 16 men and 84 women in the class

• Categorical dichotomy
  – E.g. 16 people answered an item correctly, while 84 answered it incorrectly
  – Because the item’s difficulty on an underlying scale is 1 standard deviation above the mean
With more than two categories

- Categorical (Jaguar, Porsche, Ford)
- Ordinal (e.g. “I like talking to strangers”)
  - 16% Strongly Agreed
  - 32% Agreed
  - 32% Disagreed
  - 16% Strongly Disagreed

- Note that there are three thresholds
  - (In agree/disagree only one):
Thresholds in Ordinal data

• “I like talking to strangers” trait of Extraversion

• Three thresholds:
  • Between Strongly Agree and Agree
    • divides extraversion latent trait at 1 s.d. Above mean
  • Between Agree and Disagree
    • divides extraversion latent trait at the mean
  • Between Disagree and Strongly Disagree
    • divides latent trait at 1.s.d. Below the mean

• Partial Credit Model in IRT
Types of correlation for these data

• For genuine dichotomies (eg Male, Female)
  – Phi Coefficient (categorical, categorical)
  – Point-Biserial Correlation (categorical, continuous)

• Where dichotomy reflects a division on an assumed underlying normal distribution (E.g. Agree/Disagree)
  – Tetrachoric Correlation (nominal, nominal)
  – Biserial Correlation (nominal, continuous)

• Where ordinal data represents divisions on an assumed underlying normal distribution (E.g. Always, Sometimes, Rarely, Never)
  – Polychoric correlation
Reliability of items

• Percentage agreement “Is there a problem”

  E.g.  Yes  No
  Yes  3    1
  No   2    57

  Percentage agreement > 90%

• Therefore Cohen’s Kappa

• Adjusts for expected levels of agreement
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
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
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 the same thing?

• Multi-trait/Multi-method approach (Campbell & Fiske)
  – Convergent validity
  – Discriminant validity
Divergent Validity

• Does the test measure the trait it purports to measure?
• ‘Anxiety’ but not ‘Depression’
• ‘Potential’ but not ‘Ability’
• ‘Critical Thinking’ but not ‘Intelligence’
• ‘Self Monitoring’ but not ‘Impression Management’
Predictive validity

• Validates the test against its ability to predict
• Performance at work
• Performance at university
• Success in a career
• Potential to learn
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? Do they ‘do better’:
  • (a) In their medical school exams?
  • (b) as doctors?
Accuracy of Predictors

- AC (prom)
- WS Tests
- Ability Ts
- AC (perf)
- Biodata
- Pers Tests
- Interviews
- References
- Astrology
- Graphology

Prediction

0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
AC (prom) WS Tests Ability Ts AC (perf) Biodata Pers Tests Interviews References Astrology Graphology
Construct Validity

• What is
  • Electricity
  • Intelligence
  • Competency
  • Creativity
  • Self-esteem

• How do we know that our criterion is also a measure of the same construct?
Requirements for validity

• Clear definitions of purpose
• Purposes within purposes
• Ways of assessing each criterion of success
• Unambiguous way of explaining success
• The coriolis force (mixed criteria).
Theories of personality

• What is integrity?
• Personality and personality tests
• ORPHEUS
• Ethical theory and integrity testing
• GIOTTO
• Conclusion
The psychological assessment of personality

• Psychometric tests of personality have been based on:
  – The Medical Model
  – The Ethical Model
  – The Factor Analysis Model
  – The “Big Five” Model
The Medical Model
Galen of Pergamon (c.130 – c.200.)

- **Phlegmatic** *(placid, indifferent, unconcerned)*
  - Cold/Moist, Phlegm, Water, North, Winter
- **Melancholic** *(morose, gloomy, moody)*
  - Cold/Dry, Black Bile, Earth, East, Autumn
- **Choleric** *(hot-tempered, irritable)*
  - Dry/Hot, Yellow Bile, Fire, South, Summer
- **Sanguine** *(confident, cheerful, hopeful)*
  - Moist/Hot, Blood, Air, West, Spring
What is integrity?

• The word ‘integrity’ is derived from the word ‘integer’, meaning ‘wholeness’

• Systems with integrity are balanced, in step, and working to a common purpose

• Signs are:
  – Reliability and Dependability
  – Openness and Transparency
  – Confidence and Optimism
Prudentius saw human life as a series of choices between (i) indulging our animal nature, and (ii) using our powers of reason. As we grow from childhood to adulthood the choices we make become habits and determine our adult personality.

“In our hearts conflicting affections fight hard in successive combats, and savage war rages hotly within our bones. Man’s complex nature is an uproar of rebellion.”
The Factor Analysis Model
Galton’s Lexical Hypothesis

“Any important individual differences between people would have become encoded throughout history in single linguistic terms that would occur in all the world’s languages”
(Sir Francis Galton, 1881)
The ‘Big Five’ Model

The results of meta-analysis

- **F**ellowship
  - Extraversion/Introversion
- **A**uthority
  - Tough-mindedness/Agreeableness
- **C**onformity
  - Conservatism/Openness-to-experience
- **E**motion
  - Stability/Neuroticism
- **D**etail
  - Conscientiousness/(opposite not named)
Characteristics of Orpheus (1)

• Pilot contained a lie scale
• Major scales were constructed to be as independent from lying as possible \((r < 0.30)\)
• Reliable lie scale in final version (as P6)
• This provides for Dissimulation and Despondency audit
Characteristics of Orpheus (2)

• Scales constructed to be as independent from each other as possible (r < .30)
• Much wider potential for sophisticated interpretation (120 different combinations of scale profiles)
• Less risk of misinterpretation through differential validity failure.
• Narrative based on reliable and valid combinations of scores
Characteristics of Orpheus (3)

• Acquiescence addressed by
  • Balanced positive and negative items for all scales
  • Within subject standardisation
  • Eliminates acquiescence response bias and enhances validity of scales

• WS standardisation
  • Subtraction from mean eliminates yea/nay saying
  • Division by s.d. Eliminates range effects
  • Test constructed from WS standardised pilot data
  • Requires computer scoring.
Characteristics of Orpheus (4)

- Response Audit Scales
  - Ambivalence
  - Inattention
  - Dissimulation
  - Despondency
- Ambivalence (contradiction)
- Inattention (chains and patterns)
- Alerts at 5%, 2% and 1% extremes
Reliability and validity of the Orpheus Major Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Reliability (N=423)</th>
<th>Validity (N=240)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fellowship</td>
<td>0.73</td>
<td>0.26</td>
</tr>
<tr>
<td>Authority</td>
<td>0.77</td>
<td>0.24</td>
</tr>
<tr>
<td>Conformity</td>
<td>0.76</td>
<td>0.36</td>
</tr>
<tr>
<td>Emotion</td>
<td>0.81</td>
<td>0.23</td>
</tr>
<tr>
<td>Detail</td>
<td>0.73</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Validation through correlation with supervisor’s ratings.
Interviewing for integrity

• Carelessness, sloppiness
• Lateness, absenteeism
• Hostility, intimidation, racist/sexist attitudes
• Disciplinary problems, ‘chip on shoulder’
• Disrespect for seniors, overbearing behavior
• Theft of company property, unwillingness to share
• Inability to cope with change

• SLOTH
• INDULGENCE
• ANGER
• ENVY
• PRIDE
• GREED
• DESPAIR
Reliability and validity of the Orpheus Minor Scales

Validation through correlation with supervisor’s ratings (Security Guards).

<table>
<thead>
<tr>
<th></th>
<th>Reliability (N=423)</th>
<th>Validity (N=61)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence</td>
<td>0.70</td>
<td>0.33</td>
</tr>
<tr>
<td>Work-orientation</td>
<td>0.70</td>
<td>0.28</td>
</tr>
<tr>
<td>Patience</td>
<td>0.73</td>
<td>0.27</td>
</tr>
<tr>
<td>Fair-mindedness</td>
<td>0.72</td>
<td>0.29</td>
</tr>
<tr>
<td>Loyalty</td>
<td>0.73</td>
<td>0.26</td>
</tr>
<tr>
<td>Disclosure (Lie)</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>Initiative</td>
<td>0.72</td>
<td>0.28</td>
</tr>
</tbody>
</table>
Integrity vs. personality

- INTEGRITY
  - Evaluative
  - Can change
  - Responsible
  - By choice

- PERSONALITY
  - Non-evaluative
  - Cannot change
  - Not responsible
  - By nature
“The Psychomachia” of Prudentius

• Literally ‘Battle for the mind’
• A metaphorical allegory portraying human development as a battle between:
  – Our animal instincts (The ‘passions’ or ‘vices’)
  – Our ability to reason (The ‘sentiments’ or ‘virtues’)

53
Influence of the Psychomachia

• Dante Alighieri (1263-1321)
  – The divine comedy: Inferno, Purgatory, Paradise

• Giotto di Bondoni, (1267-1377)
  – Frescoes at the Arena Chapel in Padua

• John Bunyan (1628-1688)
  – “The pilgrim’s progress: from this world to that which is to come”.
Reliability and validity of Giotto
(correlations with Orpheus Minor Scales)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Reliability (N=701)</th>
<th>Validity (N=504)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prudence</td>
<td>0.72</td>
<td>0.40</td>
</tr>
<tr>
<td>Fortitude</td>
<td>0.71</td>
<td>0.56</td>
</tr>
<tr>
<td>Temperance</td>
<td>0.75</td>
<td>0.58</td>
</tr>
<tr>
<td>Justice</td>
<td>0.71</td>
<td>0.47</td>
</tr>
<tr>
<td>Faith</td>
<td>0.73</td>
<td>0.57</td>
</tr>
<tr>
<td>Charity</td>
<td>0.75</td>
<td>0.21*</td>
</tr>
<tr>
<td>Hope</td>
<td>0.76</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Next week

• Standardisation
• How to interpret test results
• Norm and Criterion referencing
• Equivalence
• Test Equivalence
• Differential Item Functioning
• Test bias and equality of opportunity