Psychometrics in Context: Test Construction

Professor John Rust
http://www.psychometrics.cam.ac.uk
Types of assessment

• First impressions
• Application forms and references
• Objective tests (on or off line)
• Projective tests
• Interviews
• Essays and examinations
• Biodata
What is a test?

- A collection of items
- The responses to the items are combined in some way to create a score or scores
- Normally this is a simple sum of number of items ‘correct’
What is a psychometric test

• A psychometric test is one that:
  • has been constructed according to psychometric principles
  • for which its psychometric properties are available for evaluation purposes

• The defining criteria is that it has been:
  • carefully constructed
  • piloted
  • item analysed
  • non-fitting items have been deleted
Measurement in psychometrics

• Most psychometric tests are measurement scales
• The items can be considered statistically as variables (called ‘indicators’ in modelling)
• The sum, or other linear combination, of the items is the observed representation of a latent variable called a ‘latent trait’ or a ‘true score’
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.

Reliability

• 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
What can be measured?

- length, blood pressure, knowledge, desire, intelligence
- “Temperature is what thermometers measure”
- Measurements, decisions, the umpire, judgements, competitions ....
The Psychometric Principles
Maximizing the quality of assessment

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

Constructing a multiple choice test

- Defining the purpose
- Designing the blueprint
- Identifying choice options
- Writing items and the pilot study
- Item analysis
- Obtaining reliability and validity
- Writing the handbook
Defining the purpose

• The validity of a psychometric test is its ability to measure what it is supposed to measure.

• The key question is “Once this test has been developed, how can (and will) it be validated?” First question ‘What is it for?’

• NOT “A test of knowledge of statistics” but e.g.
  • “A test to ... select statisticians for x, y, z”
  • “A test identifying gaps in the curriculum”
  • “A test showing readiness for Part 2” etc.
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

## Manifestations

<table>
<thead>
<tr>
<th>Knowledge of Terms (25%)</th>
<th>Arithmetic</th>
<th>Geometry</th>
<th>Algebra</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Understanding (25%)</th>
<th>Arithmetic</th>
<th>Geometry</th>
<th>Algebra</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Application (25%)</th>
<th>Arithmetic</th>
<th>Geometry</th>
<th>Algebra</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Generalisation (25%)</th>
<th>Arithmetic</th>
<th>Geometry</th>
<th>Algebra</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<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>
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
The item response table

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>correct</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
The item analysis table

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
The item analysis table

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Scoring items not people

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Item score</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
The difficulty value

<table>
<thead>
<tr>
<th>Item</th>
<th>Score</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Item score</td>
<td></td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>1.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.4</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Contribution to total variance

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Item score</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>1.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.4</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>p(1-p)</td>
<td>0.0</td>
<td>0.16</td>
<td>0.24</td>
<td>0.24</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>
P vs. p(1-p)
## Item discrimination

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>r</td>
<td>-</td>
<td>0.53</td>
<td>0.87</td>
<td>0.22</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Classical Test Theory

• Choose items that maximize variance it the test’s score.

• The contribution of each item to the total test variance comes from:
  • the item variances
  • the Item covariances

• Tests and/or subtests must be unidimensional and linear.
Using difficulty (p) and discrimination (r) indices

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

1- My partner is usually sensitive to and aware of my needs.
2- I really appreciate my partner’s sense of humour.
3+ My partner doesn’t seem to listen to me any more.
4- My partner has never been disloyal to me.
5- I would be willing to give up my friends if it meant saving our relationship.
6+ I am dissatisfied with our relationship.
7+ I wish my partner was not so lazy and didn’t keep putting things off.
8+ I sometimes feel lonely even when I am with my partner.
9- If my partner left me life would not be worth living.
10- We can ‘agree to disagree’ with each other.
11+ It is useless carrying on with a marriage beyond a certain point.
– 12- We both seem to like the same things.
– 13+ I find it difficult to show my partner that I am feeling affectionate.
– 14- I never have second thoughts about our relationship.
– 15- I enjoy just sitting and talking with my partner.
– 16+ I find the idea of spending the rest of my life with my partner rather boring.
– 17- There is always plenty of ‘give and take’ in our relationship.
– 18+ We become competitive when we have to make decisions.
– 19+ I no longer feel I can really trust my partner.
– 20- Our relationship is still full of joy and excitement.
21+ One of us is continually talking and the other is usually silent.
22- Our relationship is continually evolving.
23+ Marriage is really more about security and money than about love.
24+ I wish there was more warmth and affection between us.
25- I am totally committed to my relationship with my partner.
26+ Our relationship is sometimes strained because my partner is always correcting me.
27+ I suspect we may be on the brink of separation.
28- We can always make up after an argument.
Software for classical test construction

• SPSS
  – Analysis
    • Scale
      – Reliability analysis
        » Statistics
        » Item
        » Scale if item deleted
Items in Short GRIMS

• My partner is sensitive to and aware of my needs. (P)
• My partner doesn’t listen to me any more. (N)
• I’m sometimes lonely when I’m with my partner. (N)
• Our relationship is full of joy and excitement. (P)
• I wish there was more warmth and affection between us. (N)
• I suspect we are on the brink of separation. (N)
• We can make up quickly after an argument. (P)
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%
Other tips

- Don’t ignore non responses
- Don’t ignore remarks
- Don’t ignore changes to items.
- Reduce items by about 50 – 100%
Writing the handbook

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