Psychometric Instrument Development

Lecture 6
Survey Research & Design in Psychology
James Neill, 2012

Overview
1. Recap: Exploratory factor analysis
2. Concepts & their measurement
3. Measurement error
4. Psychometrics
5. Reliability & validity
6. Composite scores
7. Writing up instrument development

Readings: Psychometrics
What is factor analysis?

- FA is:
  - a family of multivariate correlational data analysis methods
  - used to identify clusters of covariance (called factors)
- Two main types:
  - Exploratory factor analysis (EFA)
  - Confirmatory factor analysis (CFA)

EFA assumptions

- Sample size
  - 5+ cases per variables (min.)
  - 20+ cases per variable (ideal)
  - Another guideline: Or N > 200
- Check bivariate outliers & linearity
- Factorability: check any of:
  - Correlation matrix: Some over .3?
  - Anti-image correlation matrix diags > .5
  - Measures of Sampling Adequacy
    - KMO > ~ .5 to 6; Bartlett's sig?
**Summary of EFA steps / process**

1. Test assumptions
   - Sample size, Outliers & linearity, Factorability

2. Select type of analysis
   - PC/PAF, Orthogonal/Oblique rotation

3. Determine no. of factors
   - Theory, Kaiser's criterion, Eigen Values, Scree plot, % variance explained, interpretability of weakest factor

4. Select items
   - Check factor loadings to identify which items belong in which factor; drop items 1-by-1 if primarily loading low and/or cross-loadings high and/or item wording doesn't belong to meaning of factor.

5. Name and define factors

6. Examine correlations amongst factors

7. Check factor structure for subgroups

8. Analyse internal reliability

9. Compute composite scores

Covered in this lecture
271 UC students responded to 24 university student motivation statements in 2008 using an 8-point Likert scale (False to True), e.g., “I study at university…”
– to enhance my job prospects.
– because other people have told me I should.
• EFA PC Oblimin revealed 5 factors

Example EFA:
University student motivation

Career & Qualifications
(6 items; $\alpha = .92$

Self Development
(5 items; $\alpha = .81$

Social Opportunities
(3 items; $\alpha = .90$

Altruism
(5 items; $\alpha = .90$

Social Pressure
(5 items; $\alpha = .94$
### Example EFA: Factor correlations

<table>
<thead>
<tr>
<th>Motivation</th>
<th>CQ</th>
<th>SD</th>
<th>SO</th>
<th>AL</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Career &amp; Qualif.</td>
<td>.26</td>
<td>.25</td>
<td>.24</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>Self Develop.</td>
<td></td>
<td>.33</td>
<td>.55</td>
<td>-.18</td>
<td></td>
</tr>
<tr>
<td>Social Enjoyment</td>
<td></td>
<td></td>
<td>.26</td>
<td>.33</td>
<td></td>
</tr>
<tr>
<td>Altruism</td>
<td></td>
<td></td>
<td></td>
<td>.11</td>
<td></td>
</tr>
<tr>
<td>Social Pressure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Exploratory factor analysis: Q & A

?
Concepts & their measurement

**Operationalising fuzzy concepts**

**Concepts & their measurement: Bryman & Cramer (1997)**

**Concepts**
- form a linchpin in the process of social research
- express common elements in the world (to which we give a name)

**Hypotheses**
- express relations between concepts

“Once formulated, a concept … will need to be operationally defined, in order for systematic research to be conducted in relation to it..."

"...An operationale definition specifies the procedures (operations) that will permit differences between individuals in respect of the concept(s) concerned to be precisely specified..."

Operationalisation

- ...is the act of making a fuzzy concept measurable.
- Social sciences often use multi-item measures to assess related but distinct aspects of a fuzzy concept.

Operationalisation steps

1. Brainstorm indicators of a concept
2. Define the concept
3. Draft measurement items
4. Pre-test and pilot test
5. Examine psychometric properties – how precise are the measures?
6. Redraft/refine and re-test
Operationalising a fuzzy concept:
Example (Brainstorming indicators)

Personal skill level was at or above desired standard
Communicated well with teammates in game situation
Played with good sportsmanship
Played consistently well throughout the game (as opposed to playing in patches)
Experienced "flow" states during game (i.e., optimal arousal/performance)
Achieved the desired game result
Personally implemented the team's game plan
Happy with fitness during game

Trust
Organisational Commitment
Job Satisfaction
Self-Efficacy
Influence of Leadership Behaviours

Empowerment

Autonomy/Control Over Nursing Practice/Accountability
Job Performance
Occupational Mental Health
Work Effectiveness
Immediate Supervisor's Power
Work Teams
Measurement precision & noise

“The lower the precision, the more subjects you’ll need in your study to make up for the "noise" in your measurements. Even with a larger sample, noisy data can be hard to interpret. And if you are an applied scientist in the business of testing and assessing clients, you need special care when interpreting results of noisy tests.”

http://www.sportsci.org/resource/stats/precision.html

Measurement error

Measurement error is any deviation from the true value caused by the measurement procedure.

- Observed score = true score + measurement error
- Measurement error = systematic error + random error
To minimise measurement error

Use well designed measures:
- Multiple indicators for fuzzy constructs
- Sensitive to target constructs
- Clear instructions and questions

To minimise measurement error

Reduce demand effects:
- Train interviewers
- Use standard administration survey protocol
To minimise measurement error

Obtain a representative sample:
• Use probability-sampling if possible
• Minimise bias in selection for non-probability sampling

Maximise response rate:
• Pre-survey contact
• Minimise length / time / hassle
• Offer rewards / incentives
• Coloured paper
• Call backs / reminders

To minimise measurement error

Ensure administrative accuracy:
• Set up efficient coding, with well-labelled variables
• Check data (double-check at least a portion of the data)

Psychometrics
Psychometrics: Goal

To validly measure differences between individuals and groups in psychosocial qualities such as attitudes and personality.

Psychometrics: As test-taking grows, test-makers grow rarer

"Psychometrics, one of the most obscure, esoteric and cerebral professions in America, is now also one of the hottest."

e.g., due to increased testing of educational and psychological capacity and performance

Psychometric tasks

- Develop approaches and procedures (theory and practice) for measurement of psychological phenomena
- Design and test psychological measurement instrumentation
e.g., examine and improve reliability and validity
But remember

**Psychometric methods**

- Factor analysis
  - Exploratory
  - Confirmatory
- Classical test theory:
  - Reliability
  - Validity
- Item response modeling

**Reliability & Validity**
Types of reliability

• **Internal consistency**
  —correlations amongst multiple items in a factor
  • Split-half reliability
  • Odd-even reliability
  • Cronbach’s Alpha ($\alpha$)
  • Alternate forms reliability

• **Test-retest reliability**
  —correlation over time
  • Product-moment correlation ($r$)

---

Reliability vs. validity

Reliability is generally thought to be necessary for validity, but it does not guarantee validity.

---

Reliability

Reproducibility of a measurement
Reliability and validity (Howitt & Cramer, 2005)
Reliability and validity are the means by which we evaluate the value of psychological tests and measures.

- Reliability is about
  - the consistency of the items within the measure
  - the consistency of a measure over time
- Validity concerns the evidence that the measure actually measures what it is intended to measure.

Reliability and validity (Howitt & Cramer, 2005)

- Reliability and validity are not inherent characteristics of measures. They are affected by the context and purpose of the measurement → a measure that is valid for one purpose may not be valid for another purpose.

Reliability rule of thumb

<.6 = Unreliable
.6 = OK
.7 = Good
.8 = Very good, strong
.9 = Excellent
>.95 = may be overly reliable or redundant – this is subjective and whether a scale is overly reliable depends also on the nature what is being measured
Reliability rule of thumb

Table 7 Fabrigar et al. (1999)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>Average reliability of variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than .60</td>
<td>3</td>
<td>1.9</td>
</tr>
<tr>
<td>.60-.69</td>
<td>6</td>
<td>3.8</td>
</tr>
<tr>
<td>.70-.79</td>
<td>33</td>
<td>20.8</td>
</tr>
<tr>
<td>.80-.89</td>
<td>33</td>
<td>20.8</td>
</tr>
<tr>
<td>.90-1.00</td>
<td>14</td>
<td>8.8</td>
</tr>
<tr>
<td>Unknown</td>
<td>70</td>
<td>44.0</td>
</tr>
</tbody>
</table>

Rule of thumb - reliability coefficients should be over .70, up to approx. .90

Internal consistency
(or internal reliability)

Internal consistency is about:
• How well multiple items combine as a measure of a single concept
• The extent to which responses to multiple items are consistent with one another

Internal consistency
(Recoding)

Remember to:
• Ensure that negatively-worded items are recoded
### Types of internal consistency: Split-half reliability
- Sum the first half of the items.
- Sum the second half of the items.
- Compute a correlation between the sums of the two halves.

### Types of internal consistency - Odd-even reliability
- Sum items 1, 3, 5, etc.
- Sum items 2, 4, 6, etc.
- Compute a correlation between the sums of the two halves.

### Types of internal reliability: Alpha reliability (Cronbach's $\alpha$)
- Averages all possible split-half reliability coefficients.
- Akin to a single score which represents the degree of intercorrelation amongst the items.
How many items per factor?

- More items → greater reliability
  (The more items, the more ‘rounded’ the measure)
- Law of diminishing returns
- Min. = 2?
- Max. = unlimited?
- Typically ~ 4 to 12 items per factor
- Final decision is subjective and depends on research context

Internal reliability example: Student-rated quality of maths teaching

- 10-item scale measuring students’ assessment of the educational quality of their maths classes
- 4-point Likert scale ranging from: strongly disagree to strongly agree

Quality of mathematics teaching

1. My maths teacher is friendly and cares about me
2. The work we do in our maths class is well organised.
3. My maths teacher expects high standards of work from everyone.
4. My maths teacher helps me to learn.
5. I enjoy the work I do in maths classes.
   + 5 more
Internal reliability example: Quality of maths teaching

SPSS: Corrected Item-total correlation

<table>
<thead>
<tr>
<th>School</th>
<th>Scale Mean if Item Deleted</th>
<th>Scale Variance if Item Deleted</th>
<th>Corrected Item Total Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>School 1</td>
<td>41.15</td>
<td>98.608</td>
<td>.458</td>
<td>.888</td>
</tr>
<tr>
<td>School 2</td>
<td>40.04</td>
<td>91.500</td>
<td>.648</td>
<td>.872</td>
</tr>
</tbody>
</table>

SPSS: Cronbach's α

<table>
<thead>
<tr>
<th>School</th>
<th>Scale Mean if Item Deleted</th>
<th>Scale Variance if Item Deleted</th>
<th>Corrected Item Total Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>School 1</td>
<td>41.15</td>
<td>98.608</td>
<td>.438</td>
<td>.866</td>
</tr>
<tr>
<td>School 2</td>
<td>40.04</td>
<td>91.500</td>
<td>.648</td>
<td>.872</td>
</tr>
</tbody>
</table>
SPSS: Reliability output

Item-total Statistics

<table>
<thead>
<tr>
<th>Scale</th>
<th>Mean</th>
<th>Variance</th>
<th>Corrected Mean</th>
<th>Variance</th>
<th>Total Mean</th>
<th>Total Variance</th>
<th>Item-Deleted Mean</th>
<th>Item-Deleted Variance</th>
<th>Alpha</th>
<th>Alpha-Deleted</th>
<th>Item-Deleted Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATHS1</td>
<td>25.2749</td>
<td>25.5752</td>
<td>.6414</td>
<td>.8629</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS2</td>
<td>25.0333</td>
<td>26.5322</td>
<td>.6155</td>
<td>.8661</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS3</td>
<td>25.0192</td>
<td>30.5174</td>
<td>.0996</td>
<td>.9021</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS4</td>
<td>25.0786</td>
<td>25.8671</td>
<td>.7255</td>
<td>.8590</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS5</td>
<td>25.4664</td>
<td>25.6455</td>
<td>.7075</td>
<td>.8622</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS6</td>
<td>25.0813</td>
<td>24.9830</td>
<td>.7114</td>
<td>.8587</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS7</td>
<td>25.0909</td>
<td>26.4215</td>
<td>.6208</td>
<td>.8662</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS8</td>
<td>25.8699</td>
<td>25.7345</td>
<td>.6513</td>
<td>.8637</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS9</td>
<td>25.0340</td>
<td>26.1201</td>
<td>.6762</td>
<td>.8623</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS10</td>
<td>25.4642</td>
<td>25.7578</td>
<td>.6495</td>
<td>.8638</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reliability Coefficients

N of Cases = 1353.0
N of Items = 10
Alpha = .8790

SPSS: Reliability output

Item-total Statistics

<table>
<thead>
<tr>
<th>Scale</th>
<th>Mean</th>
<th>Variance</th>
<th>Corrected Mean</th>
<th>Variance</th>
<th>Total Mean</th>
<th>Total Variance</th>
<th>Item-Deleted Mean</th>
<th>Item-Deleted Variance</th>
<th>Alpha</th>
<th>Alpha-Deleted</th>
<th>Item-Deleted Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATHS1</td>
<td>22.2694</td>
<td>24.0699</td>
<td>.6821</td>
<td>.8907</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS2</td>
<td>22.0280</td>
<td>25.2710</td>
<td>.6978</td>
<td>.8961</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS4</td>
<td>21.9727</td>
<td>24.4372</td>
<td>.7305</td>
<td>.8870</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS5</td>
<td>22.4605</td>
<td>24.2235</td>
<td>.6801</td>
<td>.8909</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS6</td>
<td>22.7573</td>
<td>25.9862</td>
<td>.7255</td>
<td>.8873</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS7</td>
<td>22.8449</td>
<td>25.0777</td>
<td>.6106</td>
<td>.8955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS9</td>
<td>22.0280</td>
<td>24.5812</td>
<td>.7015</td>
<td>.8895</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATHS10</td>
<td>22.4530</td>
<td>24.8365</td>
<td>.6524</td>
<td>.8930</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reliability Coefficients

N of Cases = 1355.0
N of Items = 9
Alpha = .9024

Table: Definitions of the Life Effortlessness Questionnaire dimensions, with Internal Consistency and Test-Retest Correlation

<table>
<thead>
<tr>
<th>Life Effortlessness Dimension</th>
<th>Description</th>
<th>Items per scale</th>
<th>Test-Retest Correlation</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achievement Motivation</td>
<td>Motivation to achieve excellence and put the required effort into actions to attain it.</td>
<td>.88</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>Active Initiative</td>
<td>Initiating actions in new situations.</td>
<td>.73</td>
<td>.81</td>
<td></td>
</tr>
<tr>
<td>Emotional Control</td>
<td>Maintaining emotional control when faced with potentially stressful situations.</td>
<td>.73</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>Intellectual Flexibility</td>
<td>Adapting thinking and overcoming new information from changing conditions and different perspectives.</td>
<td>.60</td>
<td>.78</td>
<td></td>
</tr>
<tr>
<td>Self-Confidence</td>
<td>Confidence in abilities and the success of actions.</td>
<td>.70</td>
<td>.84</td>
<td></td>
</tr>
<tr>
<td>Social Competence</td>
<td>Ability and success of social interactions.</td>
<td>.79</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>Task Leadership</td>
<td>Ability to lead other people effectively when a task needs to be done and productivity is the primary requirement.</td>
<td>.81</td>
<td>.82</td>
<td></td>
</tr>
<tr>
<td>Time Management</td>
<td>Making optimal use of time.</td>
<td>.75</td>
<td>.84</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Effective in generic life skills.</td>
<td>.72</td>
<td>.84</td>
<td></td>
</tr>
</tbody>
</table>

N = 10
Validity
Validity is the extent to which an instrument actually measures what it purports to measure.

Validity – does the test measure what it is meant to measure?

Validity
• Validity is multifaceted and includes:
  – Correlations with similar measures
  – How the measure performs in relation to other variables
  – How well the measure helps to predict the future

Types of validity
• Face validity
• Content validity
• Construct validity
• Criterion validity
### Face validity
(low-level of importance overall)
- **Asks:**
  "Do the questions appear to measure what the test purports to measure?"
- **Important for:**
  Respondent buy-in
- **How assessed:**
  Read the test items

### Content validity
(next level of importance)
- **Asks:**
  "Are questions measuring the complete construct?"
- **Important for:**
  Ensuring holistic assessment
- **How assessed:**
  Diverse means of item generation (lit. review, theory, interviews, expert review)

### Criterion validity
(high importance)
- **Asks:**
  Concurrent validity & predictive validity
  "Can a test score predict real world outcomes?"
- **Important for:**
  Test relevance and usefulness
- **How assessed:**
  Correlate with external criteria such as performance appraisal scores
Construct validity (high importance)

• **Asks:** Does the test assess the construct it purports to? ("the truth, the whole truth and nothing but the truth.")

• **Important for:** Making inferences from operationalisations to theoretical constructs

• **How assessed:** Statistical (common factor underlying several measurements using different observable indicators?) and theoretical (is the theory about the construct valid?)

---

Composite Scores
Composite scores (Factor scores)

Combine item-scores into overall scores which represent individual differences in the target constructs. These new 'continuous' variables can then be used for:

- Descriptive statistics
- As IVs and/or DVs in inferential analyses such as MLR and ANOVA

There are two ways of creating composite scores:

- Unit weighting
- Regression weighting

Unit weighting

Average (or total) of all variables in a factor.
(each variable is equally weighted)

\[ X = \text{mean}(y_1 \ldots y_p) \]
Creating composite scores: Dealing with missing data

It can be helpful to maximise sample size by allowing for some missing data.

Reliability rule of thumb

\(<.6 = \text{Unreliable}
\)
\(0.6 = \text{OK}
\)
\(0.7 = \text{Good}
\)
\(0.8 = \text{Very good, strong}
\)
\(0.9 = \text{Excellent}
\)
\(>.95 = \text{may be overly reliable or redundant} – \text{this is subjective and whether a scale is overly reliable depends also on the nature what is being measured}

Composite scores: Missing data

SPSS syntax:
Compute \(X = \text{mean (v1, v2, v3, v4, v5, v6)}\)
You can specify a min. # of items. If the min. isn’t available, the composite score will be missing: e.g.,
Compute \(X = \text{mean (v1, v2, v3, v4, v5, v6)}\)
How many items can be missed? Depends on overall reliability. A rule of thumb:
- Allow 1 missing per 4 to 5 items
- Allow 2 missing per 6 to 8 items
- Allow 3+ missing per 9+ items

A researcher may decide to be more or less conservative depending on the factors’ reliability, sample size, and the nature of the study.
Regression weighting

Factor score regression weighting

The contribution of each item to the composite score is weighted to reflect some items more than other items.

\[ X = 20a + .19b + .27c + .34d \]

This is arguably more valid, but it may be marginal, and it makes factor scores difficult to compare.

Multiple regression weighting

Two calculation methods:

- Manual (use Compute)
- Automatic (use Factor Analysis – Factor Scores)

**Data view**
Writing up instrument development

• Introduction
  – Lit. review of underlying factors – theory and research

• Method
  – Materials/Instrumentation – summarise how the measures were developed and their expected factor structure
  e.g., present a table of the expected factors and their operational definitions.

• Results
  – Factor analysis
    • Assumption testing/ factorability
    • Extraction method & rotation
    • # of factors & items removed
    • Names & definitions of factors
    • Item factor loadings & communalities
    • Factor correlations
  – Reliability & composite scores
Writing up instrument development

• Discussion
  – Theoretical underpinning – Was it supported by the data? What adaptations should be made to the theory?
  – Quality / usefulness of measure – Provide an objective, critical assessment, reflecting the measures’ strengths and weaknesses
  – Recommendations for further improvement

• Writing up a factor analysis
  – See downloadable example

Summary

1. Operationally define concepts
2. Brainstorm measurement items
3. Draft measure – aiming to minimise measurement error
4. Pre-test & pilot
5. Use EFA, reliability, and validity
6. Create composite scores

Questions
References


