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

Recap: Exploratory Factor Analysis
What is factor analysis?

- Factor analysis is:
  - a family of multivariate correlational methods used to identify clusters of covariance (called factors)
- Two main purposes:
  - Theoretical (PAF)
  - Data reduction (PC)
- Two main types (extraction methods):
  - Exploratory factor analysis (EFA)
  - Confirmatory factor analysis (CFA)

EFA steps

1 Test assumptions
- Sample size
  - 5+ cases x no. of variables (min.)
  - 20+ cases x no. of variables (ideal)
  - Another guideline: N > 200
- Outliers & linearity
- Factorability - Use any of:
  - Correlation matrix: Some > .3?
  - Anti-image correlation matrix diags > .5
  - Measures of Sampling Adequacy:
    - KMO > ~ .5 to 6
    - Bartlett's sig?

EFA steps

2 Select type of analysis
- Extraction
  - Principal Components (PC)
  - Principal Axis Factoring (PAF)
- Rotation
  - Orthogonal (Varimax)
  - Oblique (Oblimin)
EFA steps
3. Determine no. of factors
   – Theory?
   – Kaiser's criterion?
   – Eigen Values and Scree plot?
   – % variance explained?
   – Interpretability of weakest factor?

EFA steps
4. Select items
   – Use factor loadings to help identify which items belong in which factor
   – Drop items one at a time if they don't belong to any factor e.g., consider any items for which
     • primary (highest) loading is low? (< .5 ?)
     • cross- (other) loading(s) are high? (> .3 ?)
     • item wording doesn't match the meaning of the factor

EFA steps
5. Name and describe factors
6. Examine correlations amongst factors
7. Analyse internal reliability
8. Compute composite scores
9. Check factor structure across sub-groups
EFA example 4: University student motivation

Example EFA: University student motivation

- 271 UC students responded to 24 student motivation statements in 2008
- 8-point Likert scale (False to True)
- For example:
  - “I study at university … ”
  - to enhance my job prospects.
  - because other people have told me I should.
- EFA PC Oblimin revealed 5 factors

Cross-loadings are all below .3.

Example EFA: Pattern matrix

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>0.85</td>
<td>-0.20</td>
<td>0.35</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>Item 2</td>
<td>0.75</td>
<td>0.30</td>
<td>0.25</td>
<td>0.15</td>
<td>0.05</td>
</tr>
<tr>
<td>Item 3</td>
<td>0.80</td>
<td>-0.15</td>
<td>0.30</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>Item 4</td>
<td>0.70</td>
<td>0.20</td>
<td>0.25</td>
<td>0.15</td>
<td>0.05</td>
</tr>
</tbody>
</table>

This is a pattern matrix showing factor loadings and shows a simple factor structure. Primary loadings for each item are above .5.
Example EFA:
University student motivation
1. Career & Qualifications
   (6 items; \( \alpha = .92 \))
2. Self Development
   (5 items; \( \alpha = .81 \))
3. Social Opportunities
   (3 items; \( \alpha = .90 \))
4. Altruism
   (5 items; \( \alpha = .90 \))
5. Social Pressure
   (5 items; \( \alpha = .94 \))

Example EFA:
University student motivation
Factor means and confidence intervals (error-bar graph)

Example EFA:
University student motivation
Factor correlations
Exploratory factor analysis: Q&A
Questions?

Psychometric Instrument Development

Readings: Psychometrics
1 Bryman & Cramer (1997).
   Concepts and their measurement. [UCLearn Reading List]
2 DeCoster, J. (2000).
   Scale construction notes. [Online]
   Reliability and validity: Evaluating the value of tests and measures. [Textbook/UCLearn Reading List]
   Ch 37: Reliability in scales and measurement: Consistency and measurement. [Textbook/UCLearn Reading List]
5 Wikiversity.
   Composite scores. [Online]
   Measurement error. [Online]
   Reliability and validity. [Online]
Operationalising fuzzy concepts

Concepts and their measurement

Bryman & Cramer (1997)

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

Bryman & Cramer (1997)

Hypotheses
• specify expected relations between concepts
Operationalisation

- A concept needs to be *operationally defined* in order to be systematically researched.
- “An operational definition specifies the procedures (operations) that will permit differences between individuals in respect of the concept(s) concerned to be precisely specified ...”

Bryman & Cramer (1997)

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

“... What we are in reality talking about here is *measurement*, that is, the assignment of numbers to the units of analysis - be they people, organizations, or nations - to which a concept refers.”

Bryman & Cramer (1997)

Operationalisation

- The act of making a *fuzzy concept* measurable.
- Social science often uses *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

**Operationalisation Example (Brainstorming indicators)**

- Played position well
- Achieved the desired game result
- Personaly implemented the team's game plan
- Played consistently well throughout the game (as opposed to playing in patches)
- Happy with fitness during game
- Experienced 'flow' states during game (i.e. optimal arousal/performance)

**Nurse empowerment**

- Trust
- Organisational Commitment
- Job Satisfaction
- Self-Efficacy
- Influence of Leadership Behaviours
- Work Teams
- Immediate Supervisor's Power
- Occupational Mental Health
- Work Effectiveness
- Job Performance

- Autonomy / Control Over Nursing Practice / Accountability
Measurement error

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

- **Observed score**
  - true score +/- measurement error
  - **Measurement error** = systematic error +/- random error
- Systematic error = sampling error +/- non-sampling error
Measurement error — e.g., bathroom scales aren’t calibrated properly, and every measurement is 0.5kg too high. This error occurs for each measurement.

Random error — e.g., measure your weight 3 times using the same scales but get three slightly different readings. The amount of error differs for each measurement.

Sources of systematic error

Non-sampling

Test reliability & validity
(e.g., unreliable or invalid tests)

Respondent bias
(e.g., social desirability)

Researcher bias
(e.g., researcher favours a hypothesis)

Paradigm
(e.g., focus on positivism)

Sampling
(non-representative sample)

Measurement precision & noise

- The lower the measurement precision, the more participants are needed to make up for the "noise" in the measurements.
- Even with a larger sample, noisy data can be hard to interpret.
- Especially when testing and assessing individual clients, special care is needed when interpreting results of noisy tests.

http://www.sportsci.org/resource/stats/precision.htm
Minimising measurement error

• Standardise administration conditions with clear instructions and questions
• Minimise potential demand characteristics (e.g., train interviewers)
• Use multiple indicators for fuzzy constructs

Minimising measurement error

• Obtain a representative sample:
  – Use probability-sampling, if possible
  – For non-probability sampling, use strategies to minimise selection bias
• Maximise response rate:
  – Pre-survey contact
  – Minimise length / time / hassle
  – Rewards / incentives
  – Coloured paper
  – Call backs / reminders

Minimising 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: Tasks

• Develop approaches and procedures (theory and practice) for measuring psychological phenomena
• Design and test psychological measurement instrumentation (e.g., examine and improve reliability and validity of psychological tests)
Psychometrics: In demand

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

Psychometricians are in demand due to increased testing of educational and psychological capacity and performance.

Psychometrics: Methods

• Factor analysis
  – Exploratory
  – Confirmatory
• Classical test theory
  – Reliability
  – Validity

Reliability and Validity
Reliability and validity
(Howitt & Cramer, 2005)

Reliability and validity (“classical test theory”) are ways of evaluating the accuracy of psychological tests and measures.

• Reliability is about consistency of
  – items within the measure
  – the measure over time
• Validity is about whether the measure actually measures what it is intended to measure.

Reliability and validity

In classical test theory, reliability is generally thought to be necessary for validity, but it does not guarantee validity.

In practice, a test of a relatively changeable psychological construct such as suicide ideation, may be valid (i.e., accurate), but not particularly reliable over time (because suicide ideation is likely to fluctuate).

Reliability vs. validity

Reliability
• A car which starts every time is **reliable**.
• A car which only starts sometimes is **unreliable**.

Validity
• A car which always reaches the desired destination is **valid**.
• A car which misses the desired destination is **not valid**.

Image source: [https://commons.wikimedia.org/wiki/File:Aiga_carrental_cropped.svg](https://commons.wikimedia.org/wiki/File:Aiga_carrental_cropped.svg)
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
Reproducibility of a measurement

Reliability: Types

- **Internal consistency**: Correlation among multiple items in a factor
  - Cronbach's Alpha (α)
- **Test-retest reliability**: Correlation between test at one time and another
  - Product-moment correlation (r)
- **Inter-rater reliability**: Correlation between one observer and another:
  - Kappa
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 depends on the nature what is being measured

Reliability: Rule of thumb

<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. .95

Internal consistency
(or internal reliability)

Internal consistency refers to:
- 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 can measured by:
- Split-half reliability
- Odd-even reliability
- Cronbach's Alpha (α)
If dealing with a mixture of positively and negatively scored items, remember to recode so that all items are measured in the same direction.

Internal consistency: Split-half reliability

- Sum the scores for the first half (e.g., 1, 2, 3) of the items.
- Sum the scores for the second half (e.g., 4, 5, 6) of the items.
- Compute a correlation between the sums of the two halves.

Internal consistency: Odd-even reliability

- Sum the scores for odd items (e.g., 1, 3, 5)
- Sum the scores for even items (e.g., 2, 4, 6)
- Compute a correlation between the sums of the two halves.
**Internal consistency: Cronbach’s alpha (α)**

- Averages all possible split-half reliability coefficients.
- Akin to a single score which represents the degree of intercorrelation amongst the items.
- Most commonly used indicator of internal reliability.

**How many items per factor?**

- More items → greater reliability (The more items, the more “rounded” the measure)
- Minimum items to create a factor is 1.
- No maximum. Law of diminishing returns = each additional item will add less and less to the reliability.
- Typically ~ 3 to 10 items per factor are used.
- 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

Internal reliability example

Quality of maths teaching

Internal reliability example

SPSS: Corrected Item-total correlation

Item-total correlations should be > ~.5
SPSS: Cronbach’s α

If “Cronbach’s α if item deleted” is higher than the α, consider removing item.

A measure for examining the relationship between individual items and the total scale; this is the value of Cronbach’s Alpha for the remaining items if the given item is not included in the scale.

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>Variance</th>
<th>Correlation Deleted</th>
<th>Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>Variance</th>
<th>Correlation Deleted</th>
<th>Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MATHS1    | 25.2749 | 25.5752 | .6614               | .8629                |
MATHS2    | 25.0333 | 26.5322 | .6235               | .8661                |
MATHS3    | 25.0192 | 30.5174 | .0994               | .9081                |
MATHS4    | 24.9786 | 25.8671 | .7255               | .8622                |
MATHS5    | 25.4664 | 25.6455 | .6707               | .8671                |
MATHS6    | 25.0813 | 24.9830 | .7114               | .8662                |
MATHS7    | 25.0909 | 26.4215 | .6208               | .8662                |
MATHS8    | 25.8699 | 25.7345 | .6513               | .8637                |
MATHS9    | 25.0340 | 26.1201 | .6762               | .8623                |
MATHS10   | 25.4642 | 25.7578 | .6495               | .8638                |

N of Cases = 1353.0
Alpha = .8790

Internal reliability example

SPSS: Reliability output

Remove this item. Math3 does not correlate well with the other items and the Cronbach’s alpha would increase without this item.

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>Variance</th>
<th>Correlation Deleted</th>
<th>Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>Variance</th>
<th>Correlation Deleted</th>
<th>Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MATHS1    | 22.2694 | 24.0698 | .6614               | .8629                |
MATHS2    | 22.0280 | 25.2730 | .6078               | .8641                |
MATHS4    | 21.9727 | 24.4372 | .7365               | .8871                |
MATHS5    | 22.4605 | 24.2325 | .6051               | .8871                |
MATHS6    | 22.0753 | 23.5423 | .7255               | .8873                |
MATHS7    | 22.8849 | 25.0777 | .6166               | .8935                |
MATHS8    | 22.8442 | 24.3669 | .6562               | .9027                |
MATHS9    | 22.9280 | 24.5812 | .7015               | .8959                |
MATHS10   | 22.4500 | 24.3850 | .6524               | .8930                |

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

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

Validity = does the test measure what its meant to measure?

Validity

- Validity is multifaceted and includes:
  - Comparing wording of the items with theory and expert opinion
  - Examining correlations with similar and dissimilar measures
  - Testing how well the measure predicts the future
Validity: Types

- Face validity
- Content validity
- Criterion validity
  - Concurrent validity
  - Predictive validity
- Construct validity
  - Convergent validity
  - Discriminant validity

Face validity

(low-level of importance overall)

- **Asks:**
  "At face-value, 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 item generation (lit. review, theory, interviews, expert review)
**Criterion validity**
(high importance)

- **Asks:**
  "Can a test score predict real world outcomes?"

- **Important for:**
  Test relevance and usefulness

- **How assessed:**
  Concurrent validity: Correlate test scores with recognised external criteria such as performance appraisal scores
  Predictive validity: Correlate test scores with future outcome e.g., offender risk rating with recidivism

---

**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:**
  - **Theoretical**
    (is the theory about the construct valid?)
  - **Statistical**
    Convergent – correlation with similar measures
    Discriminant – not correlated with other constructs
Composite Scores

Composite scores
Combine item-scores into an overall factor score which represents individual differences for the target construct. The new composite score can then be used for:
- Descriptive statistics and histograms
- Correlations
- As IVs and/or DVs in inferential analyses such as MLR and ANOVA

Ways of creating composite scores:
- Unit weighting
- Regression weighting
Unit weighting

Average (or total) of item scores within a factor.
(each variable is equally weighted)

\[ X = \text{mean}(y_1 \ldots y_p) \]

Unit weighting

.25
.25
.25
.25

Composite scores: Missing data

To maximise the sample size, consider computing composite scores in a way that allows for some missing data.

Composite scores: Missing data

SPSS syntax:

- Compute \( X = \text{mean}(v_1, v_2, v_3, v_4, v_5, v_6) \)
- Compute \( X = \text{mean}.4(v_1, v_2, v_3, v_4, v_5, v_6) \)

Specifies a min. # of items. If the min. isn't available, the composite score will be missing.

In this example, \( X \) will be computed for a case when the case has responses to at least 4 of the 6 items.

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

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

\[ X = 0.20a + 0.19b + 0.27c + 0.34d \]

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

Regression weighting

Two calculation methods:

- Manual (use **Compute** – New variable name = MEAN.*([list of variable names separated by commas]) - Unit weighted

- Automatic (use **Factor Analysis** – Factor Scores – Save as variables - Regression) - Regression weighted

Regression weighting

**Variable view:** of variables auto-calculated through SPSS factor analysis

**Data view:** Data are standardised, centred
Writing up instrument development

• Introduction
  – Review previous literature about the construct's underlying factors – consider both theory and research
  – Generate a research question e.g., “What are the underlying factors of X?”.  
  – Could also make a hypothesis about the number of factors and what they will represent.

• Method
  – Materials – summarise the design and development of the measures and the expected factor structure  
    e.g., present a table of the expected factors and their operational definitions.
### Results
- Factor analysis
  - Assumption testing
  - Extraction method & rotation
  - # of factors, with names and definitions
  - # of items removed and rationale
  - Item factor loadings & communalities
  - Factor correlations
- Reliability for each factor
- Composite scores for each factor
- Correlations between factors

### 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
- Download examples: http://goo.gl/fD2qby

### Summary
**Summary: Psychometrics**

1. Science of psychological measurement
2. Goal: Validly measure individual psychosocial differences
3. Design and test psychological measures e.g., using
   1. Factor analysis
   2. Reliability and validity

**Summary: Concepts & their measurement**

1. Concepts name common elements
2. Hypotheses identify relations between concepts
3. Brainstorm indicators of a concept
4. Define the concept
5. Draft measurement items
6. Pre-test and pilot test
7. Examine psychometric properties
8. Redraft/refine and re-test

**Summary: Measurement error**

1. Deviation of measure from true score
2. Sources:
   1. Non-sampling (e.g., paradigm, respondent bias, researcher bias)
   2. Sampling (e.g., non-representativeness)
3. How to minimise:
   1. Well-designed measures
   2. Representative sampling
   3. Reduce demand effects
   4. Maximise response rate
   5. Ensure administrative accuracy
Summary: Reliability
1 Consistency or reproducibility
2 Types
   1 Internal consistency
   2 Test-retest reliability
3 Rule of thumb
   1 > .6 OK
   2 > .8 Very good
4 Internal consistency
   1 Split-half
   2 Odd-even
   3 Cronbach’s Alpha

Summary: Validity
1 Extent to which a measure measures what it is intended to measure
2 Multifaceted
   1 Compare with theory and expert opinion
   2 Correlations with similar and dissimilar measures
   3 Predicts future

Summary: Composite scores
Ways of creating composite (factor) scores:
1. Unit weighting
   1. Total of items or
   2. Average of items
      (recommended for lab report)
2. Regression weighting
   1. Each item is weighted by its importance to measuring the underlying factor (based on regression weights)
Summary: Writing up instrument development

1. Introduction
   1. Review constructs & previous structures
   2. Generate research question or hypothesis

2. Method
   1. Explain measures and their development

3. Results
   1. Factor analysis
   2. Reliability of factors
   3. Descriptive statistics for composite scores
   4. Correlations between factors

4. Discussion
   1. Theory? / Measure? / Recommendations?

References


Next lecture

Multiple linear regression I
• Correlation (Review)
• Simple linear regression
• Multiple linear regression