THE UNIVERSITY
of EDINBURGH

WEEK 5
Exploratory Factor Analysis 2

Data Analysis for Psychology in R 3

dapR3 Team

Department of Psychology
The University of Edinburgh

Learning Objectives

1. Understand and apply criteria for identifying good EFA solutions
2. Apply and interpret factor congruence for testing the replicability of factor solution
3. Estimate different measures of reliability
4. Compute and save factor scores from EFA solutions.

Part 1: Evaluating and modifying a solution

Part 2: Factor congruence

Part 3: Reliability

Part 4: Validity

Part 5: Factor Scores

Practical Steps

1. Check the appropriateness of the data and decide of the appropriate estimator.
2. Decide which methods to use to select a number of factors.
3. Decide conceptually whether to apply rotation and how to do so.
4. Decide on the criteria to assess and modify your solution.
5. Run the analysis.
6. Evaluate the solution (apply 4)
7. Select a final solution and interpret the model, labelling the factors.
8. Report your results.

Evaluating results

- Good idea to start by examining how much variance each factor accounts for and the total amount of variance

- You interpret the meaning of the factors based on the size and the sign of the loadings that you deem to be "salient"

    - What's salient is defined by research question
    - In personality, and most other, research, salient loadings are those $\geq |.3| or |.4|$

Check results

- Need to also look for signs of trouble
- Heywood cases
    - If present, something is **wrong**; you should not trust these results
    - Try different rotation, eliminate item, rethink whether factor analysis is right "tool"
- Are there items that do not have any salient loadings?
    - Could signal a problem item, which should be removed
    - Could signal presence of another factor; can pursue this when revising the questionnaire
- Do some items have multiple salient loadings (cross-loadings)?
    - Indicated by the item complexity values.
- Do any factors load on only two or three items?
    - Minimum number of items should = 3
    - May have over-extracted
    - May be that you're trying to measure too many things with too few items

Good list of criteria

1. All factors load on 3+ items at salient level.
2. All items have at least one loading above salient cut off.
3. No heywood cases
4. Complex items removed in accordance with goals.
5. Solution accounts for an acceptable level of variance given goals.
6. Item content of factors is coherent and substantively meaningful.

Check results

Remember: If you deleted one or more items, you **must** re-run your factor analysis starting at trying to determine how many factors you should extract

Most important is this: If one or more factors don't make sense, then either your items are bad, your theory is bad, your analysis is bad, or all three are bad!

End of Part 1

Part 1: Evaluating and modifying a solution

Part 2: Factor congruence

Part 3: Reliability

Part 4: Validity

Part 5: Factor Scores

Replicability

- After conducting a factor analysis and developing a questionnaire, it's a good idea to test whether it replicates

- One way to do this is to see whether similar factors appear when similar data are collected

    - Examples: Arguably the Big Five (some caveats)
    - The "positive manifold" of mental abilities

- Another way is to test this formally by collecting data on another sample.

    - Or split one large sample into two (exploratory vs confirmatory)

- Based on these two samples, we can then use a number of approaches:

    - Compute **congruence coefficients** between the factors
    - 'Targeted' rotations that try to rotated one set of factors towards another set
    - **Confirmatory factor analysis** where you specify what factors load on what items

Congruence Coefficients

- Congruence coefficients, or Tucker's Congruence Coefficients, are essentially the correlations between vectors of factor loadings across samples.

$$r_c = \frac{\Sigma x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

- To calculate congruence:
    - Run the factor model in sample 1.
    - Run the factor model in sample 2
    - Ensuring the same items are included, same number of factors specified etc.
    - Calculate congruence (very simple in R).

Congruence Coefficients R

```r
library(psych)
library(tidyverse)
bfi <- na.omit(bfi) #drop missing data for ease

expl <- bfi %>%
  sample_frac(.5) # randomly select one half

conf <- anti_join(bfi, expl) # select the non-matching cases

res1 <- fa(expl[1:25], nfactors = 5, rotate = "oblimin") # run EFA on expl
res2 <- fa(conf[1:25], nfactors = 5, rotate = "oblimin") # run same analysis on conf

fa.congruence(res1, res2) # test the congruence
```

```
##       MR2   MR1   MR4   MR5   MR3
## MR2  0.99 -0.10 -0.01 -0.03 -0.03
## MR3 -0.10  0.12  0.98  0.06  0.07
## MR5 -0.01  0.27  0.11  0.99  0.04
## MR1  0.13 -0.97 -0.06 -0.19  0.00
## MR4 -0.02  0.21  0.08  0.05  0.99
```

Congruence Coefficients

- Lorenzo-Seva & ten Berge (2006) discuss nice properties of the Tucker coefficient:
    - It measures similarity independent of the mean size of the loadings.
    - It is insensitive to a change in the sign of any pair of loadings.
- MacCallum et al. (1999) suggest the following criteria following Tucker:
    - < 0.68 = terrible
    - 0.68 to 0.82 = poor
    - 0.82 to 0.92 = borderline
    - 0.92 to 0.98 = good
    - 0.98 to 1.00 = excellent

Confirmatory factor analysis

- In EFA, all factors load on all items
    - These loadings, as you have seen, are purely data driven
    - However, if we have idea about which items should group, we may want to test this explicitly

- In CFA, we specify a model and test how well it fits the data
    - We specify a model by indicating what loadings we believe will be zero
    - We then try to reject this model

- CFA is powerful and can be used for other purposes.
    - Test whether people 'use' a test in the same way across samples
    - Control for measurement error
    - Separate our method and trait variance

Exploratory

$$\begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \\ \lambda_{61} & \lambda_{62} \\ \lambda_{71} & \lambda_{72} \\ \lambda_{81} & \lambda_{82} \end{bmatrix}$$

Confirmatory

$$\begin{bmatrix} 0 & \lambda_{12} \\ 0 & \lambda_{22} \\ 0 & \lambda_{32} \\ 0 & \lambda_{42} \\ \lambda_{51} & 0 \\ \lambda_{61} & 0 \\ \lambda_{71} & 0 \\ \lambda_{81} & 0 \end{bmatrix}$$

- If the correlations derived from the CFA don't match the correlations from your data, it *probably* means that parameters set to zero should not have been set to zero

Confirmatory factor analysis

- Requires strong theory
- Requires that the constructs you're studying exhibit simple structure, i.e., *not* personality
- Never conduct a CFA on the same data you used for your EFA
    - Inflates the Type I error rate (as other statistical methods)
- If you have an interest in CFA, feel free to email me.

End of Part 2

Part 1: Evaluating and modifying a solution

Part 2: Factor congruence

Part 3: Reliability

Part 4: Validity

Part 5: Factor scores

Acknowledgement

- Slides *heavily* influenced by work of Bill Revelle and his book on psychometrics

Measurement

- The aim of measurement is to develop and use measures of constructs to test psychological theories

- 'Classical test theory' describes data from any measure as a combination of
    - The signal of the construct, or the 'true score'
    - Noise or 'error', that is, measure of other, unintended things

$$\text{Observed score} = \text{True score} + \text{Error}$$

- This should remind you of the factor analysis formula that you learned in the last lecture

True score theory

- If we assume of our test that:
  1. It measures some ability or trait
  2. In the world, there is a "true" value or score for this test for each individual

- Then the reliability of the test is a measure of how well it reflects the true score

Parallel tests

- Charles Spearman was the first to note that, under certain assumptions, the correlations between two **parallel tests** of the same construct provided an estimate of reliability

- What are the assumptions?

    - **Parallelism**
    - Both tests have the same relationship to the true score and the same error variance
    - **Tau equivalence**
    - Same relation to true score, error variance can differ.
    - **Congeneric**
    - 4+ tests, we can relax the relation assumption and just assume each is an imperfect measure.

Where do parallel tests come from?

- Previous, older definitions of "parallel tests" were somewhat abstract
- Parallel tests can come from several sources
  - Time tests were administered
  - Raters
  - Items

Alternate forms reliability

- Correlation between two variants of a test
  - Same items in different order (randomization of stimuli)
  - Tests with similar, but not identical, content, e.g., tests with a fixed number of basic addition, subtraction, division, and multiplication problems
  - Ideally, alternate tests have equal means and variances
- Assumption is that, if the tests are perfectly reliable, then they should correlate perfectly
  - They won't
  - To the extent that they don't, we have a measure of reliability
- When we have 4+ tests, we can use factor loadings to estimate each test's relation to the true score (the latent variable)
  - Hence why it's nice to have more than three items per factor

Alternate forms reliability

- Developing alternate forms is becoming much easier

    1. Write a large item bank
    2. Get many respondents
    3. Use item-response theory (IRT) to assess how "difficult" items are
    4. Create your tests with items matched on difficulty

- This is what computerized adaptive testing is all about

Test-retest reliability

- Correlation between tests taken at 2+ points in time (think again about parallel, tau equivalent, and congeneric reliability)

- Corner-stone of test assessment and appears in many test manuals, but poses some tricky questions

    - What's the appropriate time between when measures are taken?
    - How stable should the construct be if we are to consider it a trait?

- **Remember**: Even if you have within-individual changes in mean scores, so long as rank ordering is consistent, correlations can stay high

Split-half reliability

- This measure of reliable indicates how internally consistent the test is

  1. Split test into a pair of equal subsets of n items
  2. Score the two subsets
  3. Correlate these scores

- With an increasing number of items, the number of possible splits gets very large, the relationship is:

$$\frac{n!}{2\left(\frac{n}{2}\right)!^2}$$

Cronbach's alpha

- If we take the idea of correlating subsets of items to its logical conclusion
    - Split-half reliability is a special case of reliability among all test items
    - The best known estimate of this is formula 2 from Cronbach (1951)

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum\limits_{i=1}^{n} V_i}{V_t}\right)$$

- V_t is the variance of test scores (total variance)
- V_i is the variance of the n item scores after they've been weighted (error variance)

Cronbach's alpha

- It does **not** indicate whether items measure one unidimensional construct
    - This is clear when one recognizes that Cronbach's alpha increases as you add items, so it can be high even if there's no underlying factor
    - Known as 'Spearman-Brown prophecy formula'

$$\rho^*_{xx'} = \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}}$$

- $\rho^*_{xx'}$ is the predicted (prophesized) Cronbach's alpha

- $\rho_{xx'}$ is the original Cronbach's alpha

- n ratio of number of new to old measures, so, e.g., if $n = 2$, the new test has twice as many items; can also be a fraction

McDonald's omega

- Any item may measure
  - A "general" factors that load on all items
  - A "group" or "specific" factor that loads on a subset of items
- Given this, we can derive two internal consistency measures
  - Omega hierarchical (ω_h), the proportion of item variance that is general
  - Omega total (ω_t), the total proportion of reliable item variance
- Estimating these values
  - Use `omega` function in the psych package
  - Use CFA and compute it manually

Interrater reliability

- Ask a set of judges to rate a set of targets
  - Get friends to rate the personality of a family member
  - Get zoo keepers to rate the subjective well-being of an animal

- We can determine how consistent raters are by means of intraclass correlation coefficients
  - How reliable are their individual estimates
  - How reliable is the average estimate based on the judges' ratings

Intraclass correlations

- Splits variance of a set of ratings into multiple components
    - Variance between subjects (across targets)
    - Variance within subjects (across raters, same target)
    - Variance due to raters (across targets, same rater)

- Depending on what we want to know and the design of our study, we can calculate intraclass correlations from these variance components

Intraclass correlations

ICC	Description
1,1	Targets rated by a different set of randomly selected raters and reliability is based on one measurement
1,k	As (1,1), but with reliability calculated as the average of k raters' measurements
2,1	Each target measured by each rater. Raters are considered representative of larger pool of raters. Reliability calculated from a single measurement
2,k	As (2,1), but with reliability calculated as the average of k raters' measurements
3,1	Each target measured by each rater. Raters only raters of interest. Reliability calculated from a single measurement
3,k	As (3,1), but with reliability calculated as the average of k raters' measurements

Shrout and Fleiss (1979)

Uses of reliability

- Good to know how reliable a measure is
    - Implications for validity (will discuss it soon)
    - Also allows us to 'correct for attenuation'

$$r^*_{xy} = \frac{r_{xy}}{\sqrt{\rho^2_{\theta x}\rho^2_{\theta y}}}$$

r^*_{xy} is the correlation between x and y after correcting for attenuation

r_{xy} is the correlation before correcting for attenuation

$\rho^2_{\theta x}$ is the reliability of x

$\rho^2_{\theta y}$ is the reliability of y

Reliability in R

Reliability over	Estimate	Functions in R
Forms	Alternate forms	cor
Time	Test-retest	cor
		rptR
Split-half	Random split	splitHalf
	Worst split	splitHalf
	Best split	splitHalf
Items	General factor	omega
	Average	alpha
Raters	All variants	ICC

End of Part 3

Part 1: Evaluating and modifying a solution

Part 2: Factor congruence

Part 3: Reliability

Part 4: Validity

Part 5: Factor Scores

Validity

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of the test scores for the proposed uses that are valuated, not the test itself.
>
> Standard for Educational and Psychological Testing

Debates about the definition

[W]hether a test really measures what it purports to measure (Kelley, 1927)

[H]ow well a test does the job it is employed to do. The same may be used for ... different purposes and its validity may be high for one, moderate for another and low for a third (Cureton, 1951)

Validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (Messick, 1989)

A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes (Borsboon et al., 2004)

[V]alidity means that the information yielded by a test is appropriate, meaningful, and useful for decision making -- the purpose of mental measurement (Osterlind, 2010)

Evidence for validity

- Debates about how to define validity lead to questions about what constitutes evidence for validity
- Sources of evidence align to what may be viewed as "classical" concepts reported in textbooks, studies, and test manuals

Evidence related to content

- Content validity
    - A test should contain only content relevant to the intended construct
    - It should measure what it was intended to measure

- Face validity
    - i.e., for those taking the test, does the test "appear to" measure what it was designed to measure?

Evidence related to the scale

- Do the items measure a single "intended" construct?
- Factor analysis only provides very limited information towards this
- How else might we assess it?

Evidence related to the scale

- Do the items measure a single "intended" construct?
- Factor analysis only provides very limited information towards this
- How else might we assess it?
    - Construct validity, or convergent and discriminant validity (Cronbach & Meehl, 1955)

Relationships with other constructs

- **Convergent**:
  - Measure should have high correlations with other measures of the same construct
- **Discriminant**:
  - Measure should have low correlations with measures of different constructs
- **Nomological Net**
  - Measure should have expected patterns (positive/negative) correlations with different sets of constructs
    - Also, some measures should vary depending on manipulations,
    - e.g., a measure of "stress" should be higher among students who are told that a test is "high stakes" than among students told that a test is "low stakes"

Relationships with other constructs

- Consider relations in terms of temporal sequence.
- Concurrent validity: Correlations with contemporaneous measures
  - Neuroticism and subjective well-being
  - Extraversion and leadership
- Predictive validity: Related to expected future outcomes
  - IQ and health
  - Agreeableness and future income

Evidence related to response processes

- Discussed in a paper by Karabenick et al. (2007)

- Not commonly considered in validation studies

    - Is how people "process" the items belonging to a scale the way we think they ought to?
    - Do tests of intelligence engage problem-solving behaviors?
    - Do extraversion items lead people to reflect on related past behaviors?

> Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., ... & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean?. Educational Psychologist, 42(3), 139-151.

Evidence related to consequences

- Perhaps most controversial aspect of current validity discussions
- Should potential consequences of test use be considered part of the evidence for test's validity?
- Important questions for the use of tests
    - Is my measure systematically biased or fair for all groups of test takers?
    - Does bias have social ramifications?

Example: Implicit association test

- Reliability is okay, but not great
- Weakly (if at all) predicts discriminatory behavior
- Used to label people, decide who gets certain jobs, etc.

Definition redux (Hughes, in press)

- Validity has many meanings, some markedly different

- Validation is on-going process concerning accuracy and appropriateness of a test

Accuracy

- Content
- Response processes
- Structural (within and across groups)
- Convergent and discriminant

Appropriateness

- Predictive, concurrent, incremental
- Know groups who test is designed for
- Consequences (fairness, bias)
- Feasibility (cost, length, etc.)

> Hughes, D. (2018). Psychometric Validity: Establishing the Accuracy and Appropriateness of psychometric measures. In The Wiley Handbook of Psychometric Testing: A Multidisciplinary Approach to Survey, Scale and Test Development John Wiley & Sons Ltd.

Relationship between reliability and validity

- Reliability: relation of true score with observed score
- Validity: correlations with other measures play a key role
- Logically, a score or measure cannot correlate with anything more than it correlates with itself, so reliability is the limit on validity

Importance of test reliability and validity

- Fundamental first step in measurement
  - If we cannot measure well variables of interest, then we cannot study them
  - Large, tricky problem in psychology; many variables/constructs are not directly accessible

- Important for later research
  - Poor reliability and validity may lead to erroneous conclusions due to measurement problems
  - If we know reliability of test, we can sometimes make adjustments (correction for attenuation)
  - Not "glamorous" research, but can be interesting in and of itself

Where can you find this information?

- Test manuals

- Papers describing new tests and papers investigating exisitng measures in different groups, languages, contexts, etc.
    - *Assessment*
    - *Psychological Assessment*
    - *European Journal of Psychological Assessment*
    - *Organisational Research Methods*
    - Personality journals

- Papers describing new ways to establish reliability, validity, etc.
    - *Methdology*
    - *Psychometrika*
    - *Journal of Educational Statistics*

End of Part 4

Part 1: Evaluating and modifying a solution

Factor congruence

Part 3: Reliability

Part 4: Validity

Part 5: Factor scores

Factor scores

- Sometimes EFA is an end in and of itself: research question may concern the structure of a set of items
- However, you usually want to "do something" with your factors
  - Test whether your construct is related to other constructs in ways that you would predict
  - Test whether your construct is related systematically to other variables, including those that you manipulate
- This pursuit was key to the resurgence of trait theory
- To do these things, you need variables that represent what you've measured

Factor scores

- Unit-weighting
    - Sum raw scores on the observed variables which have primary loadings on each factor
    - Which items to sum is a matter of defining what loadings are salient
    - Need to reverse score items with negative loadings
- Regression-based methods
    - Thurstone or Thompson method
        - Ordinary least squares approach
        - Computes scores from observed item correlations and loadings
    - Bartlett method focuses on minimizing sums of squares for unique factors
- How do they fare?
    - Simple, produce estimates that are highly correlated with more exact techniques (especially true for unit-weighting)
    - Negative consequence of unit-weighting or regression method is they can produce correlated scores for orthogonal factors

Factor scores

- Anderson-Rubin and Ten Berge methods
    - Negative consequence of unit-weighting or regression method is they can produce correlated scores for orthogonal factors
    - The Anderson-Rubin method preserves orthogonality of factors
    - The Ten Berge method generalizes the Anderson-Rubin method to preserve correlations (or lack thereof) between factors

Choosing factor scores

- Simple sum scores (unit-weighting) require strict properties in the data, but these are rarely tested, and do not often hold
- If the goal is to try to find higher-order factors, the correlations between the scores are important, and so ten Berge scores are preferable
    - Not doing this can lead to really biased results
    - An alternative is to factor the Phi matrix
- Alternative is to use structural equations modeling, which includes a measurement component (a CFA) and a structural component (regression)
    - Doesn't require you to compute factor scores
    - Requires good theory of measurement and structure
    - If your constructs don't approximate simple structure, you may have to turn to other alternatives

Sample size

- In the past, rules of thumb have guided sample size decisions for factor analysis (reviewed by MacCallum et al., 1999)
    - Rules based on minimum number of participants: 100, 200, 250, 500
    - Rules based on the participant-to-item (*N*:*p*) ratio: 3 to 6, 5, 10

- MacCallum et al. (1999) tested what was important using simulated datasets
    - Tested how well sample sizes of 60, 100, 200, and 400 would recover factors from each of nine population correlation
    - The populations varied in two ways
        - Items to factor (*p*:*m*) ratio (10:3, 20:7, 20:3)
        - Communalities: low (.2, .3, .4), wide (.2 to .8 in .1 increments), high (.6, .7, .8)

Sample size

- As MacCallum et al. hypothesized, the crucial determinants of minimum sample size was not the number of items, but communalities and *p*:*m*
    - Fewer subjects were needed if communalities were wide or high
    - Fewer subjects were needed if *p*:*m* was high
    - Communalities were even more important when *p*:*m* = 20:7 (an interaction effect)
- Subsequent studies support these findings; all reject that the *N*:*p* ratio should be used

Sample size

- Thus, when planning a study you should do the following to determine your minimum sample size
  - Think of how many factors you expect and get many items measuring each
  - Use pilot data and previous studies to make an "educated guess" about what communalities you're going to expect

GIGO

- Make sure to check the quality of your data
- PCA and factor analysis cannot turn bad data into good data
- 'Garbage in, garbage out'

End