2018 PSYCHOLOGY R BOOTCAMP
PENNSYLVANIA STATE UNIVERSITY
AUGUST 16, 2018

FACTOR ANALYSIS
AN INTRODUCTION

https://psu-psychology.github.io/r-bootcamp-2018/index.html

WITH ADDITIONAL MATERIALS AT

https://quantdev.ssri.psu.edu/tutorials

NILAM RAM

PENNSYLVANIA STATE UNIVERSITY

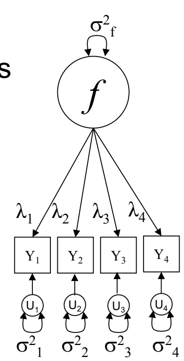
Factor Analysis: An Introduction

- What is Factor Analysis?
- Uses and Applications
- Exploratory Factor Analysis (EFA)
  - 5 Steps
  - Example
- Confirmatory Factor Analysis (CFA)
  - 5 Steps
  - Example
- Evaluating Model Fit
- Practical Issues

What is Factor Analysis?

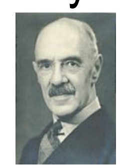
- Method for investigating the *structure* underlying variables (or people, or time)
  - a set of computational techniques widely used in research on individual differences
  - a mathematical model used to express observations in terms of latent variables

$$Y_n = \lambda f_n + u_n$$
$$\Sigma = \Lambda \Psi \Lambda' + \theta_\varepsilon$$

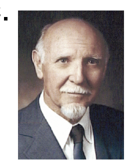


100+ years of Factor Analysis

- Beginnings: Spearman (1904)
  - "One factor theory of intelligence"



- Early Years and Transformations: C. Burt, L.L. Thurstone, H. Kaiser, R. B. Cattell, etc.
  - Methods for factor extraction
  - The number of factors
  - The meaning of factors
  - Factor rotation methods

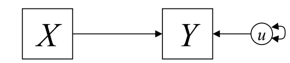


- A Revolution: Joreskog (1970s)
  - Confirmatory Factor Analysis and SEM

Response = {stimulus} + error

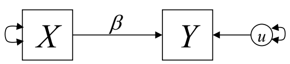
- The fundamental model of Factor Analysis can be seen as a direct descendant of other models in common usage:

In ANOVA the stimulus is <u>*fixed*</u>



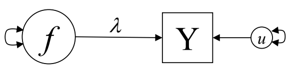
$$Y_n = X + u_n$$

In Regression the stimulus is <u>*random*</u>



$$Y_n = \beta X_n + u_n$$

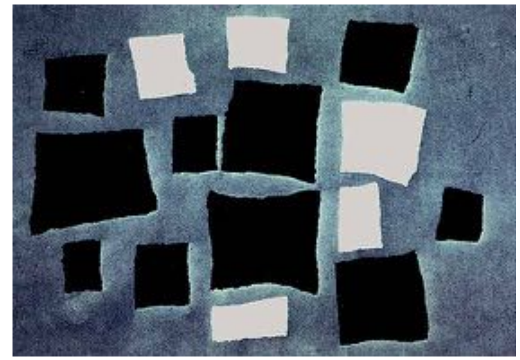
In Factor Analysis the stimulus is <u>*latent*</u>



$$Y_n = \lambda f_n + u_n$$

Observed/Manifest Variables

- A set of empirical observations – data
  - usually collected with a purpose (theory)



Arp, 1916

Factors – Abstract/Latent Variables

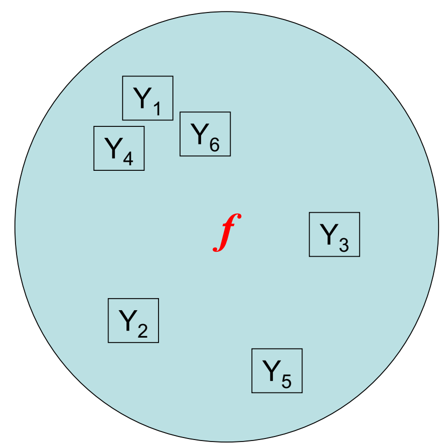
- a set of theoretical concepts used to describe hypothetical constructs

- represent testable (i.e., rejectable) hypotheses about empirical data

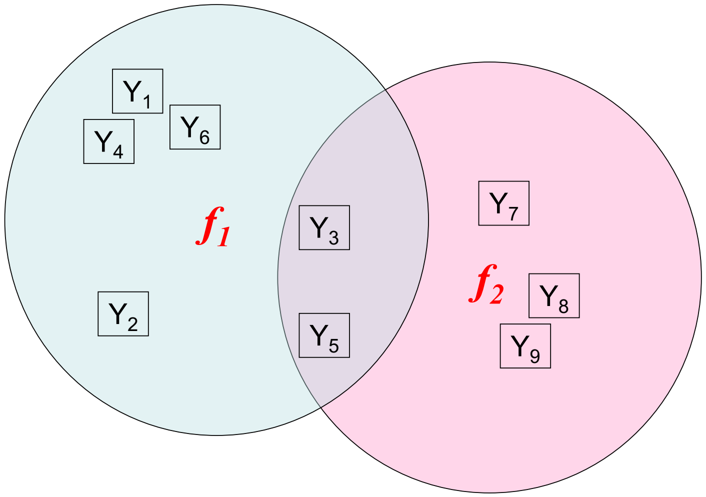
Kandinsky, 1926

- “Factors are not things – only evidence for the existence of things” (Cattell, 1966)

A Hypothetical Factor Space



A Multi-Factor Space



The Common Factor Model

- If two or more characteristics correlate they may reflect a shared underlying trait. Patterns of correlations reveal the *latent* dimensions that lie beneath the *measured qualities* (Tabachnik & Fidel, 2005)

- Aim of factor analysis is to represent the covariation among observed variables in terms of linear relations among a *smaller number* of abstract or latent variables (Cattell, 1988).

A Set of Multivariate Measurements

(Lebo & Nesselroade, 1978)

N = 103
\# of vars = 6

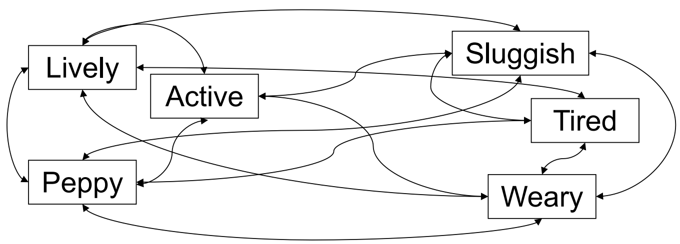
obs#	active	lively	peppy	sluggish	tired	weary
1	1	1	1	0	1	0
2	1	1	0	0	1	0
3	1	1	0	0	2	1
4	2	1	1	0	0	0
5	1	1	1	0	0	0
6	2	1	1	0	0	0
7	1	1	0	0	1	1
8	1	1	0	0	1	1
9	1	1	1	0	0	0
10	2	1	0	0	0	0

etc.

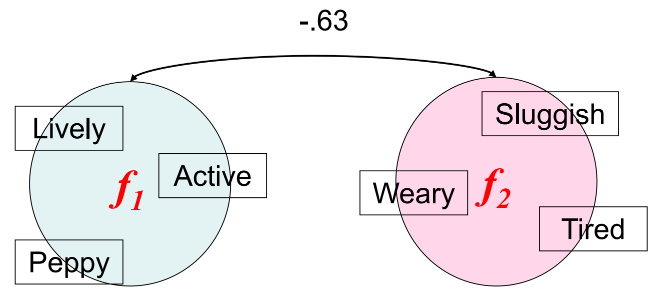
A Set of Multivariate Measurements Summarized as a Correlation Matrix

	Active	**Lively**	**Peppy**	**Slugg**	**Tired**	**Weary**
Active	**1.00**					
Lively	**.64**	**1.00**				
Peppy	**.56**	**.41**	**1.00**			
Sluggish	**-.48**	**-.35**	**-.42**	**1.00**		
Tired	**-.47**	**-.42**	**-.47**	**.72**	**1.00**	
Weary	**-.43**	**-.43**	**-.44**	**.64**	**.83**	**1.00**

A Multivariate Space



Data Reduction – Parsimonious Representation of the Data

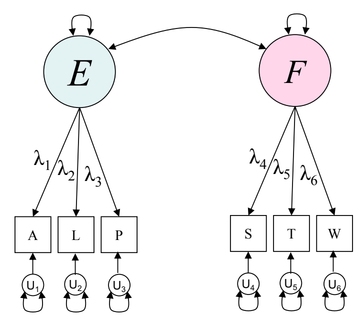


	Active	Lively	Peppy	Slugg	Tired	Weary
Active	1.00					
Lively	.64	1.00				
Peppy	.56	.41	1.00			
Sluggish	-.48	-.35	-.42	1.00		
Tired	-.47	-.42	-.47	.72	1.00	
Weary	-.43	-.43	-.44	.64	.83	1.00

	ENERGY	FATIGUE
ENERGY	1.00	
FATIGUE	-.63	1.00

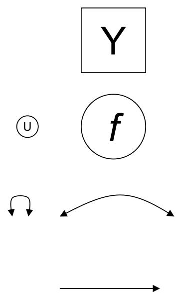
The Common Factor Model

$$Y_n = \lambda f_n + u_n$$



- The relations among these six items can be parsimoniously represented by the relation between two common factors (+ unique parts)

SEM Path Diagrams
A Key



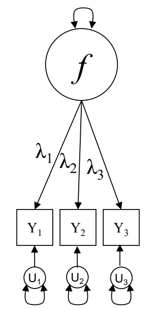
Squares = Observed Variables

Circles = Latent Variables

Double-Headed Arrows = Variances/Covariances

Single-Headed Arrows = Regressions

The Common Factor Model



$$Y_n = \lambda f_n + u_n$$

Use & Application of Factor Analysis

- Inform evaluations of construct or test validity
  - Does this set of items/variables tap into a single or multiple constructs?
  - How many constructs do we need to explain the pattern of responses in this study sample?
- Identify groups of interrelated items/variables
  - Which items are related to one another?
  - If individuals score relatively high on one item, on what other items are they also likely to score relatively high?
- Developing or testing a theory regarding hypothetical constructs
  - What underlying constructs did we measure and how do they relate to one another?
  - Did we measure the constructs we intended to measure? Do the constructs relate to one another in the hypothesized manner?
- Summarize relationships as a more parsimonious set of factors
  - that may then be used in additional analyses

EFA Steps & Example

EFA Steps

EFA Example

Exploratory Factor Analysis (EFA)

- Used to examine the dimensionality of a measurement instrument or set of variables
- Data-driven
  - *Post-hoc* examination of what structures may underlie the data
    - What factors (common and unique) were measured
    - Number of underlying factors (dimensions)
    - Inter-relations among factors
  - Finding the smallest number of interpretable factors needed to explain the correlations among a set of variables – within constraints of the model

5 Steps of EFA

1. Select data for factor analysis
2. Extract a set of factors sequentially using a set of optimization criteria
   - Principal axis
3. Select a smaller number of common factors for ease in interpretation
   - Scree test, Eigenvalues > 1
4. Rotate selected factors towards an interpretable solution
   - Orthogonal (Varimax), Oblique (promax), Target (Procrustes)
5. *Estimate factor scores using another set of criteria
   - Sum scores

Step 1: Select Data

C	Q1	Am always prepared
N	Q2	Get stressed out easily
-	**Q3**	**Have a rich vocabulary**
N	Q4	Am relaxed most of the time
C	Q5	Pay attention to details
N	Q6	Worry about things
C	Q7	Make a mess of things
N	Q8	Seldom feel blue
C	Q9	Get chores done right away
N	Q10	Am easily disturbed
C	Q11	Often forget to put things back in their proper place
N	Q12	Get upset easily
C	Q13	Like order
N	Q14	Change my mood a lot
C	Q15	Shirk my duties
N	Q16	Have frequent mood swings
C	Q17	Follow a schedule
N	Q18	Get irritated easily
C	Q19	Am exacting in my work
N	Q20	Often feel blue

A 20 item trait personality scale

N = 121

Selection of data is not "blind"

Scale intended to measure something

Q3 is filler item

Step 1: Select Data

id	q1	q2	q3	q4	q5 …
150	5	1	4	3	1
151	4	4	4	3	4
153	3	2	3	4	4
155	4	3	3	3	3
156	2	1	2	1	4
157	3	2	2	3	3
158	2	3	3	2	3
159	5	1	5	1	5
160	3	4	5	1	3
161	5	5	4	1	5
162	4	3	3	2	4
163	4	4	2	2	4

Step 2: Extract Factors

Principal Axis

- SAS

```
PROC FACTOR DATA=synpers
   METHOD=PRINIT MAXITER=100 CORR
   ROTATE=PROMAX
   SCREE NFACT=2 /*MINEIGEN=1*/ REORDER ;
   TITLE 'Exploratory 2-Factor Analysis of IPIP Items';
   VAR q1-q2 q4-q20;
RUN;
```

- SPSS
  - Analyze → Data Reduction → Factor
    - Select variables
    - **Extraction – Method: Principal Axis Factoring
    - Rotation: Promax

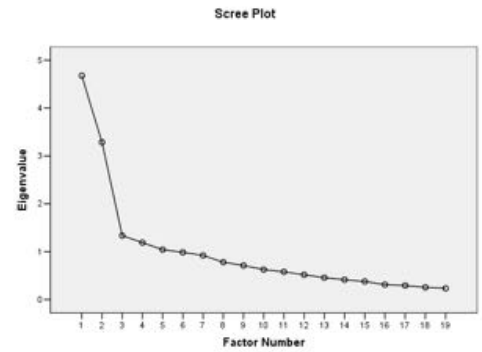
- R

```
m1 <- fa(r = synpers, nfactors=2,
         rotate="promax",
         fm="pa")
```

Principal axes factor analysis has a long history in exploratory analysis and is a straightforward procedure. Successive eigen value decompositions are done on a correlation matrix with the diagonal replaced with diag (FF’) until $\sum(diag(FF'))$ does not change (very much).

Step 3: Select Number of Factors

Scree Test, Eigenvalues > 1



Total Variance Explained

Factor	Initial Eigenvalues Total	% of Variance	Cumulative %	Rotation Total
1	4.678	24.621	24.621	3.359
2	3.287	17.299	41.919	2.408
3	1.335	7.027	48.946	3.143
4	1.190	6.261	55.207	1.952
5	1.043	5.490	60.697	2.965
6	.987	5.194	65.891	
7	.921	4.850	70.741	
8	.780	4.107	74.848	
9	.712	3.746	78.594	
10	.627	3.298	81.892	
11	.579	3.047	84.939	
12	.518	2.727	87.666	
13	.458	2.408	90.073	
14	.414	2.177	92.250	
15	.378	1.991	94.241	
16	.311	1.638	95.879	
17	.293	1.544	97.423	
18	.255	1.340	98.764	
19	.235	1.236	100.000	

Extraction Method: Principal Axis Factoring.

a. When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

Step 4: Rotate Factor Solution for Interpretation

		Factor Loadings 1	2
q1	Am always prepared	-0.111	**0.586**
q2	Get stressed out easily	**0.572**	0.038
q4	Am relaxed most of the time	**0.544**	0.088
q5	Pay attention to details	0.014	**0.325**
q6	Worry about things	**0.562**	0.051
q7	Make a mess of things	**-0.321**	**0.388**
q8	Seldom feel blue	**0.624**	-0.075
q9	Get chores done right away	-0.157	**0.666**
q10	Am easily disturbed	**0.528**	-0.044
q11	Often forget to put things …	-0.021	**0.507**
q12	Get upset easily	**0.752**	-0.004
q13	Like order	-0.014	**0.556**
q14	Change my mood a lot	**0.578**	-0.229
q15	Shirk my duties	-0.136	**0.526**
q16	Have frequent mood swings	**0.600**	-0.288
q17	Follow a schedule	-0.049	**0.707**
q18	Get irritated easily	**0.735**	-0.153
q19	Am exacting in my work	0.013	**0.494**
q20	Often feel blue	**0.646**	-0.263

Factor Correlation	
1.00	
-0.153	1.00

Conclusions:

Relations in data can be represented by 2 interpretable factors

Names of factors???

→ Evidence that scale is working in the intended manner

Step 5: *Calculate/Estimate Scores

Composite Scores

Consc = q1 + q5 + q7 + q9 + q11 + q13 + q15 + q17 + q19

Neuro = q2 + q4 + q6 + q10 + q12 + q14 + q16 + q18 + q20

id	Consc	Neuro
150	28	22
151	31	24
153	34	22
155	33	20
156	24	12
157	33	19
158	26	19
159	43	12
160	31	16
161	36	13
162	37	23
163	34	24

CFA Steps & Example

CFA Steps

CFA Example: Spearman 1904

Confirmatory Factor Analysis (CFA)

- Used to study how well a hypothesized structure fits to a sample of measurements
    - Procrustes rotation
- Hypothesis-driven
    - Explicitly test *a priori* hypotheses (theory) about the structures that underlie the data
        - Number of , characteristics of, and interrelations among underlying factors
    - Specify a common measurement base for comparisons across groups/occasions (factorial invariance)

Confirmatory Factor Analysis (CFA)

- Testing an a-priori hypothesis about the structures in the data
    - Requires specific expectations regarding
        - The number of factors
        - Which variables reflect given factors
        - How the factors are related to one another

The Common Factor Model

- Goal:
    - To represent the covariation among observed variables in terms of the linear relations between a smaller number of latent variables

$$\Sigma = \Lambda\Psi\Lambda' + \theta_\varepsilon$$

where Σ is the observed *p*-variate covariance matrix,

Λ is a *p* x *q* matrix of factor loadings,

Ψ is a *q x q* latent factor covariance matrix,

θ_ε is a *p x p* covariance matrix of unique factors

5 Steps of CFA

0. Theory-Data: Form some basic ideas of merging the common factor model and data

1. Draw a path diagram

2. Input observed covariance matrix Σ (or raw data)

3. Specify "structural expectations"
   – Number of factors
   – Relationships among factors
   – Relationships among observed variables and factors

4. Estimate parameters
   – Maximum likelihood estimation in SEM framework

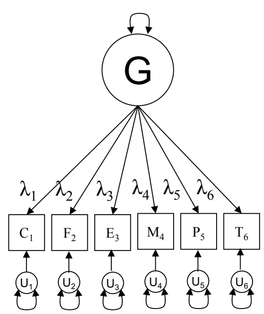
5. Evaluate parameters and fit of model

CFA Example: Step 0
The Birth of Factor Analysis, 1904

- "All branches of intellectual activity have in common one fundamental function (or group of functions) whereas the remaining or specific elements of the activity seem in every case to be wholly different from that in all others" (Spearman, 1904, p. 284)

- One-factor theory of intelligence
  – General intellectual ability (common factor)
  – Ability specific to each task or skill (unique factors)

1. A "One Factor Theory"



$$Y_n = \lambda f_n + u_n$$

2. Input Covariance Matrix

$$\Sigma = \Lambda\Psi\Lambda' + \theta_\varepsilon$$

Σ = Observed ***<u>Covariance</u>*** (Correlation) Matrix
(p x p)

N=101

	C	**F**	**E**	**M**	**P**	**T**
Classics	**1.00**					
French	**.83**	**1.00**				
English	**.78**	**.67**	**1.00**			
Math	**.70**	**.67**	**.64**	**1.00**		
Pitch	**.66**	**.65**	**.54**	**.45**	**1.00**	
Talent (Music)	**.63**	**.57**	**.51**	**.51**	**.40**	**1.00**

3. Specify Structural Expectations

$$\Sigma = \Lambda\Psi\Lambda' + \theta_\varepsilon$$

- # of Factors
  - 1 common + 6 unique
- Relations among Factors
  - Common factor is related to itself
    - Factor Covariance Matrix = Ψ
  - Common factor is unrelated to unique factors
    - By definition of the common factor model
  - Unique factors are unrelated to one another
    - Uniquenesses = θ
- Relations among observed variables and factors
  - Common factor is indicated by all six observed variables
    - Factor loading matrix = Λ

3. Specify Structural Expectations

$$\Sigma = \Lambda\Psi\Lambda' + \theta_\varepsilon$$

Λ = Factor Loading Matrix

(p x k)

	Factor1 (f_1)
Classics	λ_1
French	λ_2
English	λ_3
Math	λ_4
Pitch	λ_5
Talent	λ_6

Ψ = Factor Covariance Matrix

(k x k)

	Factor 1
Factor 1	=1.00

3. Specify Structural Expectations

$$\Sigma = \Lambda\Psi\Lambda' + \theta_\varepsilon$$

θ_ε = Uniquenesses

(p x p)

	C	F	E	M	P	T
Classics	u^2_1					
French	0	u^2_2				
English	0	0	u^2_3			
Math	0	0	0	u^2_4		
Pitch	0	0	0	0	u^2_5	
Talent	0	0	0	0	0	u^2_6

Testing "Theory" of Measurement Directly

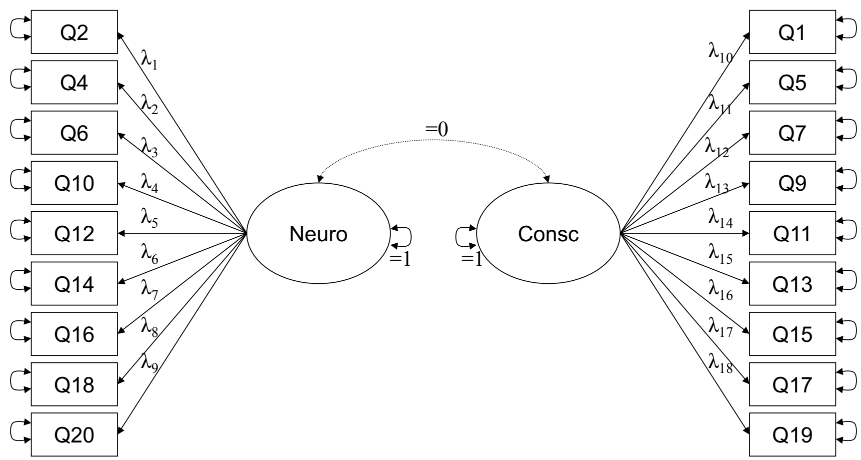
	Factor Loading Matrix	Neuro	Consc
q1	Am always prepared	---	???
q2	Get stressed out easily	???	---
q3	Filler	---	---
q4	Am relaxed most of the time	???	---
q5	Pay attention to details	---	???
q6	Worry about things	???	---
q7	Make a mess of things	---	???
q8	Seldom feel blue	???	---
q9	Get chores done right away	---	???
q10	Am easily disturbed	???	---
q11	Often forget to put things …	---	???
q12	Get upset easily	???	---
q13	Like order	---	???
q14	Change my mood a lot	???	---
q15	Shirk my duties	---	???
q16	Have frequent mood swings	???	---
q17	Follow a schedule	---	???
q18	Get irritated easily	???	---
q19	Am exacting in my work	---	???
q20	Often feel blue	???	---

Factor Covariance	
Neuro	Consc
=1.00	
0.00	=1.00

Theory:

There are two unrelated interindividual difference factors that underlie our personality scale responses: C & N.

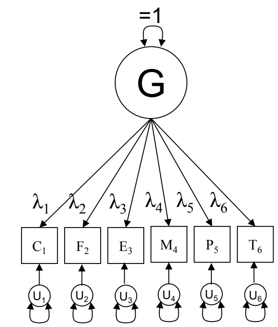
Factorial Structure of Personality Scale



3. Specify Structural Expectations

Mplus

```
TITLE: Spearman1904_corr 1 Factor

DATA:    FILE = Spearman1904_corr.dat;
         TYPE = COVARIANCE;
         NOBSERVATIONS = 101;

VARIABLE:  NAMES = c f e m p t ;
           USEVAR = c f e m p t;
           MISSING = .;

ANALYSIS: TYPE=GENERAL;

MODEL:
    g BY c* f e m p t; !Factor Loadings
    g@1;               !Factor Variance
    c f e m p t;       !Unique Variances

OUTPUT:   SAMPSTAT STANDARDIZED;
```



4. Estimate Parameters

$$\Sigma = \Lambda\Psi\Lambda' + \theta_\varepsilon$$

Λ = Factor Loading Matrix

(p x k)

	G
Classics	**.95**
French	**.87**
English	**.80**
Math	**.74**
Pitch	**.69**
Talent	**.65**

Ψ = Factor Covariance Matrix

(k x k)

	G
G	**=1.00**

4. Estimate Parameters

$$\Sigma = \Lambda\Psi\Lambda' + \theta_\varepsilon$$

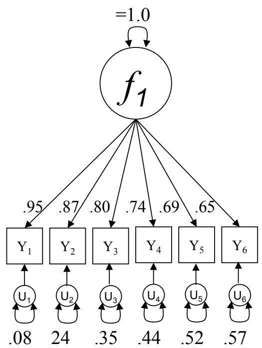
θ_ε = Uniquenesses

(p x p)

	C	F	E	M	P	T
Classics	**.08**					
French	**0**	**.24**				
English	**0**	**0**	**.35**			
Math	**0**	**0**	**0**	**.44**		
Pitch	**0**	**0**	**0**	**0**	**.52**	
Talent	**0**	**0**	**0**	**0**	**0**	**.57**

5. Evaluate Parameters & Fit of Model

Parameters of “One Factor Model”



χ^2 = 9, df = 9, RMSEA = .01

5. Evaluate Parameters & Fit of Model

$$\hat{\Sigma} = \hat{\Lambda}\Psi\hat{\Lambda}' + \hat{\theta}_{\varepsilon}$$

$\hat{\Sigma}$ = Estimated *Covariance* (Correlation) Matrix
(p x p)

	C	F	E	M	P	T
Classics	.92+.08					
French	**.82**	.76+.24				
English	**.76**	**.70**	.65+.35			
Math	**.70**	**.64**	**.59**	.56+.44		
Pitch	**.65**	**.59**	**.55**	**.51**	.48+.52	
Talent	**.62**	**.56**	**.52**	**.48**	**.45**	.43+.57

5. Evaluate Parameters & Fit of Model

Model Misfit

$\Sigma - \hat{\Sigma}$ = (Observed - Estimated)

	C	F	E	M	P	T
Classics	**.00**					
French	**-.00**	**.00**				
English	**.00**	**-.03**	**.00**			
Math	**-.01**	**.02**	**.04**	**.00**		
Pitch	**.00**	**.05**	**-.02**	**-.06**	**.00**	
Talent	**.00**	**.00**	**-.02**	**.02**	**-.05**	**.00**

Evaluating Model Fit

Basic Concepts
Fit Statistics
Relative Fit

Evaluating Model Fit

- How well does the model represent the data?
- How well does the model represent the theory?

- Fit to the data
  – Measures of how well the estimated covariance matrix derived from the model matches the observed covariance matrix (e.g., χ^2, RMSEA)
- Fit to the theory
  – Subjective interpretation

Model Fit Statistics

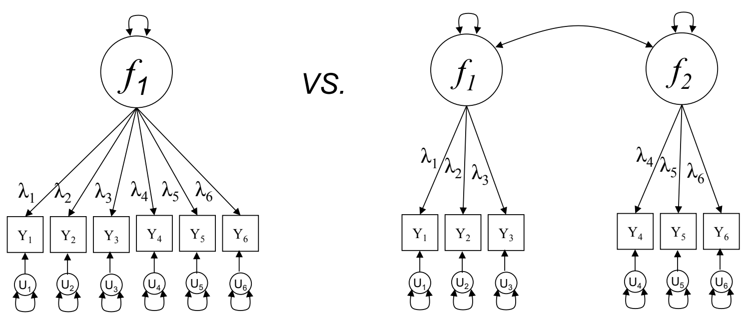
- χ^2 (or -2LL)
  – df = degrees of freedom
  – Null hypothesis – Estimated covariance matrix = Observed covariance matrix
  – (sensitive to sample size)

- RMSEA
  – Range: 0.00 to 1.00
  – lower values indicate better fit
  – Rule of thumb: RMSEA < .05 indicates good fit

- CFI (Comparative Fit Index)
- NFI (Normed Fit Index)
- TLI (Tucker-Lewis Index)
  – Range: 0.00 to 1.00+
  – higher values indicate better fit

Relative Fit

- Testing model (theory) against viable alternatives
  - e.g., fit of 1-factor model relative to 2-factor model



Relative Fit of Nested Models

- **χ^2 difference tests (for nested models)**
  – **$[(\text{Model}_B\ \chi^2) - (\text{Model}_A\ \chi^2)] / df_B - df_A$**

- Information criteria for non-nested model comparisons (using same data)
  – AIC (Aikake Information Criteria)
  – BIC (Bayes Information Criteria)
    - Lower values are better
    - **Should be used in conjunction with judgments about the theoretical interpretation of the models

Evaluating Relative Fit

- Evaluate Fit for Model A
- Add restrictions to construct Model B
- Evaluate Fit for Model B

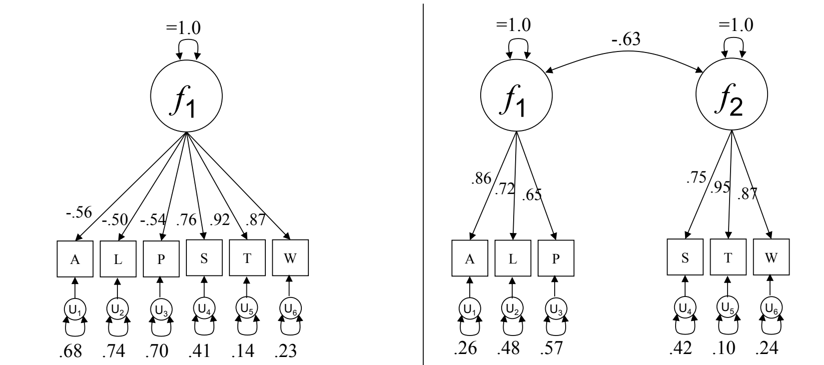
- Evaluate difference in fit = $\Delta\chi^2/\Delta df$
  - Is the restricted (parsimonious) model of significantly worse fit than the less restrictive (more complex) model – or is this complexity needed?

Relative Fit of Different Hypotheses Regarding Structure of the Data

Σ = Observed ***Covariance*** (Correlation) Matrix
(p x p)

	Active	Lively	Peppy	Slugg	Tired	Weary
Active	**1.00**					
Lively	**.64**	**1.00**				
Peppy	**.56**	**.41**	**1.00**			
Sluggish	**-.48**	**-.35**	**-.42**	**1.00**		
Tired	**-.47**	**-.42**	**-.47**	**.72**	**1.00**	
Weary	**-.43**	**-.43**	**-.44**	**.64**	**.83**	**1.00**

Relative Fit of Nested Models



χ^2 = 55, df = 9, RMSEA = .224

χ^2 = 11, df = 8, RMSEA = .053

Model Comparison: $\Delta\chi^2$ / Δdf = 44/1 p > .05

Practical Issues

Assumptions
Notes on EFA & CFA
Factor Space & Selection of Variables
Factor Analyzing Other Types of Data
CFA as base of SEM

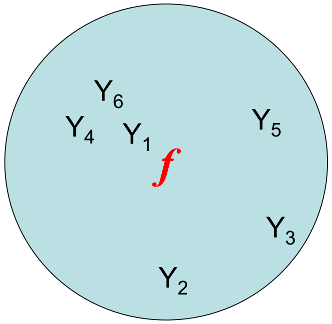
Factor Analysis Assumptions

- Continuous measures
- Multivariate normal distribution
- # of observations reasonably large
- Observations are independent

Some Practical Notes

- EFA
  - ~Large samples
  - Results influenced by the set of variables used
  - Number of factors influenced by the number of variables per factor
  - Requires interpretation of structure
- CFA
  - ~Large samples (independent from the EFA sample)
  - Results influenced by the set of variables used
  - Multiple pieces (or assumptions) needed to identify factors
  - Requires hypothesis(es) regarding structure

Factor as Centroid: Implications for Multivariate Sampling

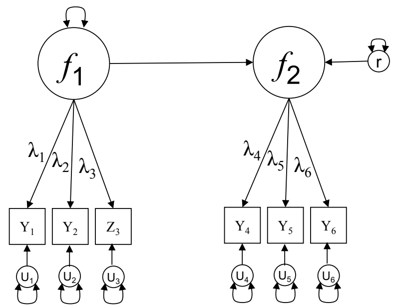


- Not always looking for factors defined by variables that are highly correlated
- Rather, looking for good coverage of factor space

Factor Analyzing Other Types of Data

- R-technique (persons x variables)
  - Relations between variables that are defined across persons
- P-technique (occasions x variables)
  - Relations between variables that are defined across occasions for a single person
- Q-technique (variables x persons)
  - Relations among persons defined across variables (How many types of people are there?)

Factor Analysis → SEM



Use & Application of Factor Analysis

Note that the method itself does not answer the theoretical question – rather, it provides evidence for careful interpretation

Richard Long, *Walking a Circle in Mist*, Scotland 1986

Selected Readings

- Gorsuch, Richard L. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.
- Loehlin, J. (1998). *Latent variable models: An introduction to factor, path, & structural analysis*. Mahwah, NJ: Erlbaum.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington DC: APA.
- Tucker, L. R, & MacCallum, R. *Exploratory factor analysis*. http://www.unc.edu/~rcm/book/factornew.htm