FACTOR ANALYSIS

What is Factor Analysis?

Factor analysis examines the interrelationships among a large number of variables and, then, attempts to explain them in terms of their common underlying dimension

- Common underlying dimensions are referred to as factors

Why do Factor Analysis?

- Data Summarization
  - Identify latent dimensions within data set
  - Identification and understanding of these underlying dimensions is the goal

- Data Reduction
  - Discover underlying dimensions to reduce data to fewer variables so all dimensions are represented in subsequent analyses
    - Surrogate variables
    - Aggregated scales
    - Factor Scores
    - Avoid multicollinearity problems
    - Improve reliability of aggregated scales

Assumptions

1. Variables must be interrelated

    20 unrelated variables=20 factors

2. Sample size

    i. Min 50, prefer 100

    ii. Min 5 observations/item, prefer 10 observations/item

Types of Factor Analysis

Exploratory Factor Analysis (EFA)

- Used to discover underlying structure

Confirmatory Factor Analysis (CFA)

- Used to test whether data fit a priori expectations for data structure
- Structural equations modeling

Purpose of EFA

- EFA is a data reduction technique
  - Scientific parsimony
  - Which items are virtually the same thing
- Objective: simplification of items into subset of concepts or measures
- Part of construct validation (what are underlying patterns in data?)
- EFA assesses dimensionality or homogeneity

Factor (or common factors) analysis

- In SPSS known as principal axis factoring
- Explain relationship between observed vars in terms of latent vars or factors
- Factor is a hypothesized construct
- Assumes error in items
- Precise math not possible, solved by iteration
- Communalities (shared var) on diagonal

Concepts and Terms

Factor - Linear composite. A way of turning multiple measures into one thing.

Factor Score - Measure of one person's score on a given factor.

Factor Loadings - Correlation of a factor score with an item. Variables with high loadings are the distinguishing features of the factor.

Concepts and Terms

Communality - (h^2) - Variance in a given item accounted for by all factors. Sum of squared factor loadings in a row from factor analysis results. These are presented in the diagonal in common factor analysis

Process

Primary methods:

- Scree rule
- Kaiser criterion (eigenvalues > 1)

How Many Factors?

Scree Plot

- Look for bend in plot
- Include factor located right at bend point

Kaiser (or Latent Root) criterion

- Eigenvalues greater than 1
- Also, 1 is the amount of variance accounted for by a single item (r^2 = 1.00). If eigenvalue < 1.00 then factor accounts for less variance than a single item.

Example

R matrix (correlation matrix)

	BlPr	LSat	Chol	LStr	BdWt	JSat	JStr
BlPr	1.00						
LSat	-.18	1.00					
Chol	.65	-.17	1.00				
LStr	.15	-.45	.22	1.00			
BdWt	.45	-.11	.52	.16	1.00		
JSat	-.21	.85	-.12	-.35	-.05	1.00	
JStr	.19	-.21	.02	.79	.19	-.35	1.00

Principal Components Analysis (PCA)

Initial Statistics:

Variable	Communality	*	Factor	Eigenval	%Var	Cum%
BLPR	1.00000	*	1	2.85034	40.7	40.7
LSAT	1.00000	*	2	1.74438	24.9	65.6
CHOL	1.00000	*	3	1.16388	16.6	82.3
LSTR	1.00000	*	4	.56098	8.0	90.3
BDWT	1.00000	*	5	.44201	6.3	96.6
JSAT	1.00000	*	6	.20235	2.9	99.5
JSTR	1.00000	*	7	.03607	.5	100.0

Example

Variable	Communality	*	Factor	Eigenval	%Var	Cum%
BLPR	1.00000	*	1	2.85034	40.7	40.7
LSAT	1.00000	*	2	1.74438	24.9	65.6
CHOL	1.00000	*	3	1.16388	16.6	82.3
LSTR	1.00000	*	4	.56098	8.0	90.3
BDWT	1.00000	*	5	.44201	6.3	96.6
JSAT	1.00000	*	6	.20235	2.9	99.5
JSTR	1.00000	*	7	.03607	.5	100.0

Factor Matrix (Unrotated):

	Factor 1	Factor 2	Factor 3	... Fac7
LSTR	.73738	-.32677	.47575	
LSAT	-.71287	.38426	.52039	
JSAT	-.70452	.42559	.48553	
JSTR	.64541	-.32867	.62912	
CHOL	.54945	.68694	-.10453	
BDWT	.48867	.60471	.13043	
BLPR	.58722	.60269	-.08534	
Eigenvalue	2.850343	1.74438	1.16388	

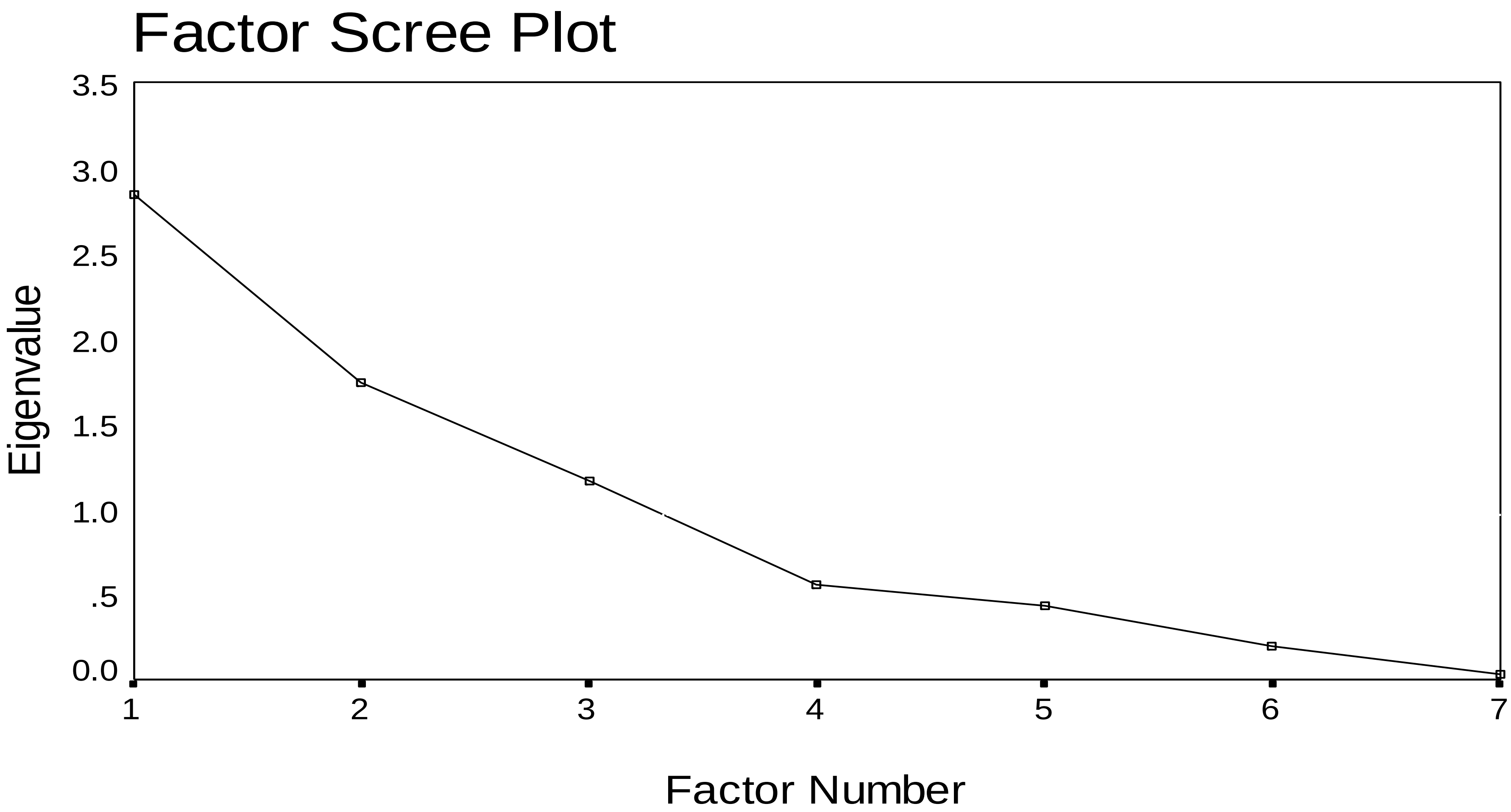
Example

Final Statistics:

Variable	Communality	*	Factor	Eigenvalue	%Var	Cum%
BLPR	.71533	*	1	2.85034	40.7	40.7
LSAT	.92665	*	2	1.74438	24.9	65.6
CHOL	.78470	*	3	1.16388	16.6	82.3
LSTR	.87684	*				
BDWT	.62149	*				
JSAT	.91321	*				
JSTR	.92037	*				

VARIMAX Rotated Factor Matrix:

	Factor 1	Factor 2	Factor 3	h^2
CHOL	*.87987*	-.10246	-.00574	.78470
BLPR	*.83043*	-.14875	.05988	.71533
BDWT	*.76940*	.05630	.16234	.62149
LSAT	-.09806	*.94430*	-.15917	.92665
JSAT	-.05790	*.93376*	-.19479	.91321
JSTR	.06542	-.10717	*.95110*	.92036
LSTR	.12381	-.26465	*.88965*	.87684
Eigenvalue	2.0883	1.8809	1.7893	



Scree comes from a word for loose rock and debris at the base of a cliff!

Information from EFA

	FACTOR			
Msr	F1	F2	F3	h^2
a	.60	-.06	.02	.36
b	.81	.12	-.03	.67
c	.77	.03	.08	.60
d	.01	.65	-.04	.42
e	.03	.80	.07	.65
f	.12	.67	-.05	.47
g	.19	-.02	.68	.50
h	.08	-.10	.53	.30
i	.26	-.13	.47	.31
Sum Sq Ldng	1.76	1.56	.98	Total
% Variance	.195	.173	.109	47.7%
	(1.76/9)	(1.56/9)	(.98/9)	

A factor loading is the correlation between a factor and an item

When factors are orthogonal, factor loadings squared are the amount of variance in one variable explained by that factor (F1 explains 36% of the variance in Msr a; F3 explains 46% of the variance in Msr g)

Information from EFA

Msr	F1	F2	F3	h^2
a	.60	-.06	.02	.36
b	.81	.12	-.03	.67
...	...	...	...	...
i	.26	-.13	.47	.31
Sum Sq Ldng	1.76	1.56	.98	Total
% Variance	.195	.173	.109	47.7%
	(1.76/9)	(1.56/9)	(.98/9)	

Eigenvalue

Eigenvalue: Sum of squared loadings down a column (associated with a factor). Total variance in all vars explained by one factor. Factors with eigenvalues less than 1 predict less than the variance of 1 item.

Communality (h^2): Variance in a given item accounted for by all factors. Sum of squared loadings across rows. Will equal 1 if you retain all possible factors.

Information from EFA

	FACTOR			
Msr	F1	F2	F3	h^2
a	.60	-.06	.02	.36
b	.81	.12	-.03	.67
...	...	...	...	...
i	.26	-.13	.47	.31
Sum Sq Ldng	1.76	1.56	.98	Total
% Variance	.195	.173	.109	47.7%
	(1.76/9)	(1.56/9)	(.98/9)	

Average of all communalities (h^2 / k) = proportion of variance in all variables explained by all factors.

If all variables reproduced perfectly by the factors, correlation between original variables equals sum of the products of factor loadings. When not perfect, gives an estimate of the correlation.

e.g. $r_{ab} \cong (.60*.81) + (-.06*.12) + (.02*-.03) \cong .48$

Information from EFA

Msr	F1	F2	F3	h^2
a	.60	-.06	.02	.36
b	.81	.12	-.03	.67
...	...	...	...	...
i	.26	-.13	.47	.31
Sum Sq Ldng	1.76	1.56	.98	Total
% Variance	.195	.173	.109	47.7%
	(1.76/9)	(1.56/9)	(.98/9)	

Eigenvalue

$1-h^2$ is the uniqueness→ variance of an item not shared with other items. Unique variance could be random error or systematic.

The factor matrix above is after rotation. Eigenvalues computed on the unrotated and unreduced factor loading matrix because we are interested in total variance accounted for in the data. Use of eigenvalues and % variance accounted for in SPSS not reordered after rotation.

Important Properties of PCA

- Each factor in turn maximizes variance explained from an **R** matrix
- For any number of factors obtained, PCs maximize variance explained
- Amount of variance explained by each PC equals the corresponding characteristic root (eigenvalue)
- All characteristic roots of PCs are positive
- Number of PCs derived equal the number of factors need to explain all the variance in **R**
- The sum of characteristic roots equals the sum of diagonal **R** elements

Rotations

- All original PC and PF solutions are orthogonal.
- Once you obtain minimal number of factors, you have to interpret them
- Interpreting original solutions is difficult. Rotation aids interpretation.
- You are looking for simple structure
  - Component loadings should be very high for a few vars and near 0 for remaining variables
  - Each variable should load highly on only 1 component

	Unrotated Matrix		Rotated Matrix	
Var	F1	F2	F1	F2
a	.75	.63	.14	.95
b	.69	.57	.14	.90
c	.80	.49	.18	.92
d	.85	-.42	.94	.09

Rotation

- After rotation, variance accounted for by a factor is spread out. First factor no longer accounts for max variance possible; others get more variance. Total variance accounted for is the same.
- Two types of rotation
  - Orthogonal (factors uncorrelated)
  - Oblique (factors correlated)

Rotation

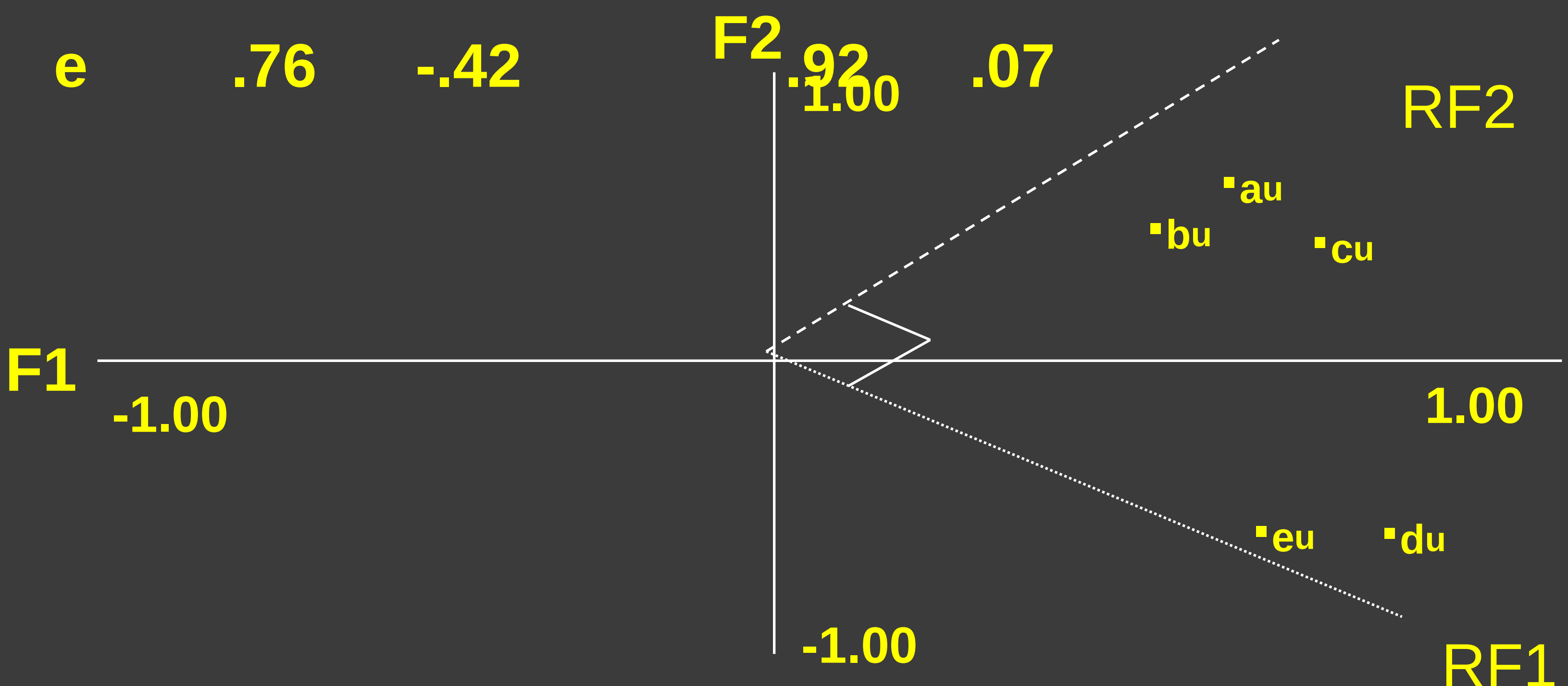
- Orthogonal rotation (rigid, 90 degrees) - PCs or PFs remain uncorrelated after transformation
  - Varimax - Simplifying column weights to 1s and 0s. Factor has items loading highly, others don't load. Not appropriate if you expect a single factor.
  - Quartimax - Simplify to 1s and 0s in a row. Item loads high on 1 factor, almost 0 on others. Appropriate if you expect single general factor.
  - Equimax. Compromise of Varimax and Quartimax rotations.
  - In practice, choice of rotation makes little difference

Rotation

- Oblique or correlated components (less or more than 90 degrees) - Account for same % var, but factors correlated
  - Some say not meaningful with PCA
  - Many factors are theoretically related, so rotation method should not "force" orthogonality
    - Allows the loadings to more closely match simple structure
    - Correlated solutions will get you closer to simple structure
    - Oblimin (Kaiser) and promax are good
  - Provides a structure matrix of loadings and a pattern matrix of partial weights – which to interpret?

Orthogonal Rotation

	Unrotated Matrix		Rotated Matrix	
Var	**F1**	**F2**	**F1**	**F2**
a	**.75**	**.63**	**.14**	**.95**
b	**.69**	**.57**	**.14**	**.90**
c	**.80**	**.49**	**.18**	**.92**
d	**.85**	**-.42**	**.94**	**.09**
e	**.76**	**-.42**	**.92**	**.07**



Simple Structure (Thurstone)

(1) Each row of factor matrix should have at least one 0 loading

(2) The number of items with 0 loadings equals the number of factors; each column has 1 or more 0 loadings

(3) Items with high loadings on one factor or the other

(4) If there are more than 4 factors, a large portion of items should have zero loadings

(5) For every pair of columns, there should be few cross-loadings

(6) Few if any negative loadings

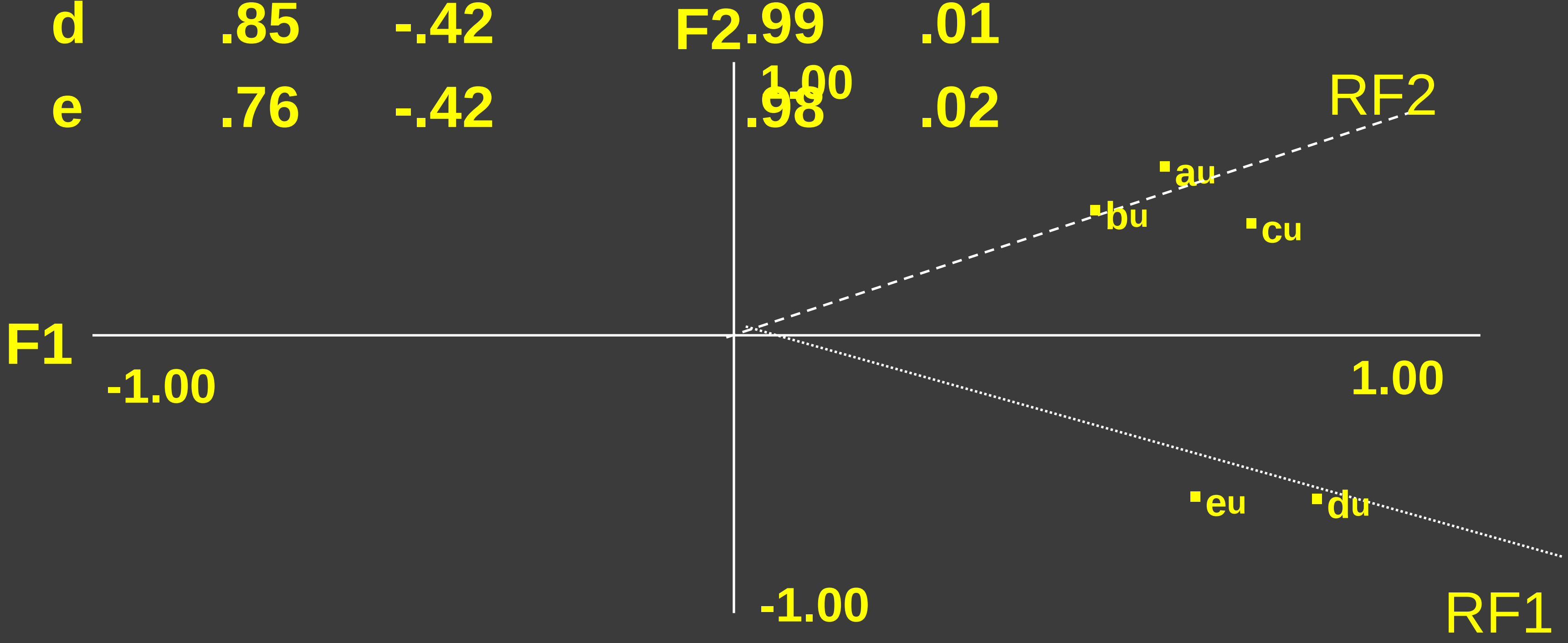
Simple Structure

	Factor		
Msr	1	2	3
a	x	0	0
b	x	0	0
c	x	0	0
d	0	x	0
e	0	x	0
f	0	x	0
g	0	0	x
h	0	0	x
i	0	0	x
j	0	0	x

Oblique Rotation

- **Example:**

	Unrotated Matrix		Rotated Matrix	
Var	**F1**	**F2**	**F1**	**F2**
a	.75	.63	.04	.98
b	.69	.57	.02	.99
c	.80	.49	.01	.97
d	.85	-.42	.99	.01
e	.76	-.42	.98	.02



Orthogonal or Oblique Rotation?

- Nunnally suggests using orthogonal as opposed to oblique rotations
  - Orthogonal is simpler
  - Leads to same conclusions
  - Oblique can be misleading
- Ford et al. suggest using oblique unless orthogonality assumption is tenable

Interpretation

- Factors usually interpreted by observing which variables load highest on each factor
  - a priori criteria for loadings (min .3+)
- Name factor. Always provide factor loading matrix in study.
- Cross-loadings are problematic
  - a priori criteria for “large” cross-loading
  - decide a priori what you will do
- Factor loadings or summated scales used to define new scale. Can go back to correlation matrix and do not only use factor loadings. Loadings can be inflated.

PCA and FA

- PCA - No constructs of theoretical meaning assumed; Simple mechanical linear combination. (1s in the diagonal of R)
- FA - assumes underlying latent constructs. Allows for measurement error (communalities in diagonal of R)
  - Also PAF or common factors analysis
- PCA uses all the variance. FA uses ONLY shared variance.
- In FA you can have indeterminant (unsolvable) solutions. Have to iterate (computer makes best “guess”) to get the solutions.

FA

- Also known as principal axis factoring or common factor analysis
- Steps
  - Estimate communalities of the variables (shared variance)
  - Substitute communalities in place of 1s on diagonal of **R**
  - Perform a principal component analysis on the reduced matrix
  - Iterated FA
    - Estimate h^2
    - Solve for factor model
    - Calculate new communalities
    - Substitute new estimates of h^2 into matrix and redo
    - Iterate until communalities don’t change much
    - Rotate for interpretation

Estimating Communalities

- Highest correlation of given variable with other variables in data set
- Squared multiple correlations (SMCs) of each variable predicted by all other variables in the data set
- Reliability of the variable
- Because you are estimating and the factors are no longer combinations of actual variables, can get funny results:
  - Communalities > 1.00
  - Negative eigenvalues
  - Negative uniqueness

Example FA

R matrix (correlation matrix with h^2)

		BlPr	LSat	Chol	LStr	BdWt	JSat	JStr
BlPr	.54							
LSat		-.18	.89					
Chol		.65	-.17	.67				
LStr	.15	-.45	.22	.87				
BdWt		.45	-.11	.52	.16	.41		
JSat		-.21	.85	-.12	-.35	-.05	.86	
JStr		.19	-.21	.02	.79	.19	-.35	.87

Principal Axis Factoring (PAF)

Initial Statistics:

Variable	Communality	*	Factor	Eigenvalue	%Var	Cum%
BLPR	.53859	*	1	2.85034	40.7	40.7
LSAT	.88573	*	2	1.74438	24.9	65.6
CHOL	.66685	*	3	1.16388	16.6	82.3
LSTR	.87187	*	4	.56098	8.0	90.3
BDWT	.41804	*	5	.44201	6.3	96.6
JSAT	.86448	*	6	.20235	2.9	99.5
JSTR	.86966	*	7	.03607	.5	100.0

FA

Principal Axis Factoring (PAF)

Initial Statistics:

Variable	Communality	*	Factor	Eigenvalue	%Var	Cum%
BLPR	.53859	*	1	2.85034	40.7	40.7
LSAT	.88573	*	2	1.74438	24.9	65.6
CHOL	.66685	*	3	1.16388	16.6	82.3
LSTR	.87187	*	4	.56098	8.0	90.3
BDWT	.41804	*	5	.44201	6.3	96.6
JSAT	.86448	*	6	.20235	2.9	99.5
JSTR	.86966	*	7	.03607	.5	100.0

Factor Matrix (Unrotated):

	Factor 1	**Factor 2**	**Factor 3**
LSAT	**-.75885**	**.31104**	**.54455**
LSTR	**.70084**	**-.20961**	**.36388**
JSAT	**-.70038**	**.31502**	**.39982**
JSTR	**.68459**	**-.29044**	**.66213**
CHOL	**.48158**	**.74399**	**-.07267**
BLPR	**.48010**	**.56066**	**-.02253**
BDWT	**.36699**	**.47668**	**.08381**

FA

Principal Axis Factoring (PAF)

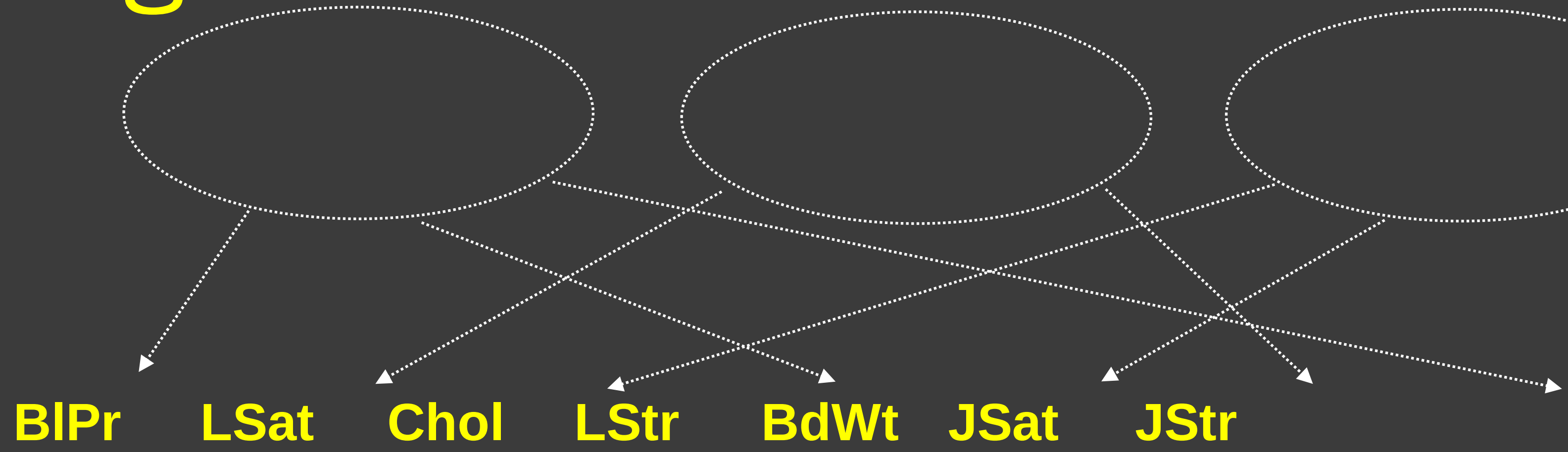
Final Statistics:

Variable	Communality	*	Factor	Eigenvalue	%Var	Cum%
BLPR	.54535	*	1	2.62331	37.5	37.5
LSAT	.96913	*	2	1.41936	20.3	57.8
CHOL	.79071	*	3	1.04004	14.9	72.6
LSTR	.66752	*				
BDWT	.36893	*				
JSAT	.74962	*				
JSTR	.99144	*				

Rotated Factor Matrix (VARIMAX):

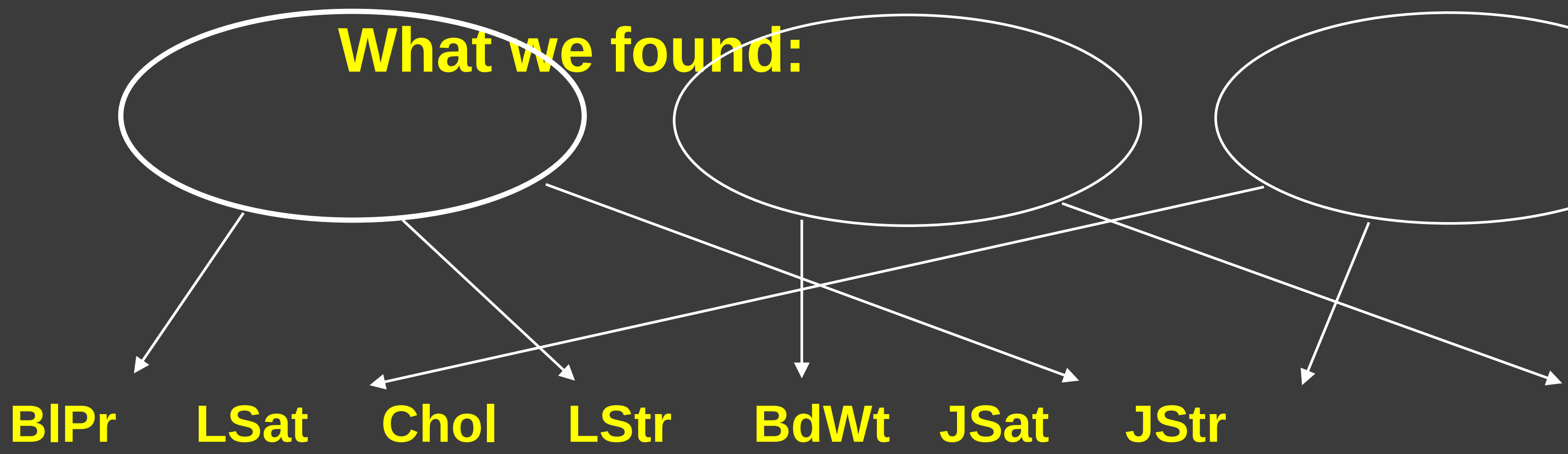
	Factor 1	**Factor 2**	**Factor 3**
LSAT	**.96846**	**-.10483**	**-.14223**
JSAT	**.83532**	**-.07092**	**-.21643**
CHOL	**-.08425**	**.88520**	**-.00547**
BLPR	**-.11739**	**.72364**	**.08898**
BDWT	**-.00430**	**.59379**	**.12778**
JSTR	**-.10474**	**.07011**	**.98770**
LSTR	**-.28514**	**.15273**	**.75026**

Logic of FA



How many? What are the factors?

What we found:



PCA vs. FA

- Pros & Cons:
  - Pro PCA: has solvable equations. “Math is right”.
  - Con PCA: Lumping garbage together. Also, no underlying concepts.
  - Pro FA: considers role of measurement error, gets at concepts.
  - Con FA: doing mathematical gymnastics.
- Practically: Usually not much difference
  - PCA will tend to converge more consistently
  - FA is more meaningful conceptually

PCA vs. FA

- Situations where you might want to use FA:
  - Where there are 12 or fewer variables (diagonal will have a large impact)
  - Where the correlations between the variables are small, then diagonals will have a large impact
- If you have clear factor structure, won’t make much difference
- Otherwise:
  - PCA will tend to overfactor
  - If doing exploratory analysis, may not mind overfactoring

Using FA Results

- Single surrogate measure – choose a single item with a high loading to represent factor
- Summated Scale*
  - Form a composite from items loading on same factor
  - Average all items that load on a factor (unit weighting)
  - Calculate the alpha for the reliability
  - Name the scale/construct
- Factor Scores
  - Composite measures for each factor were computed for each subject
  - Based on all factor loadings for all items
  - Not easily replicated

Reporting

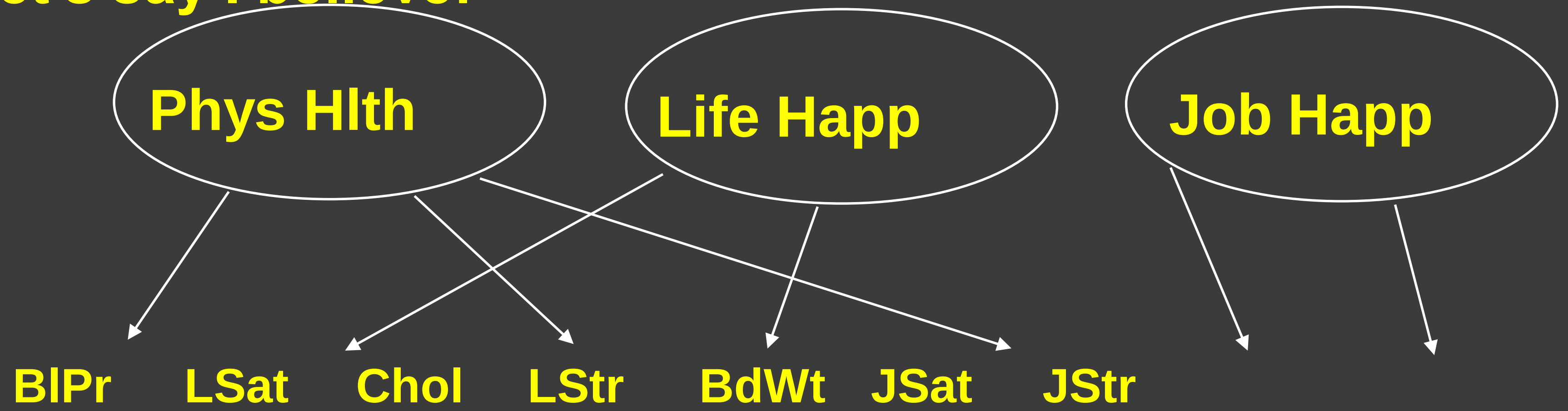
- If you create a factor based scale, describe the process
- Factor analytic study, report:
  - Theoretical rationale for EFA
  - Detailed description of subjects and items, including descriptive stats
  - Correlation matrix
  - Methods used (PCA/FA, communality estimates, factor extraction, rotation)
  - Criteria employed for number of factors and meaningful loadings
  - Factor matrix (aka pattern matrix)

Confirmatory Factor Analysis

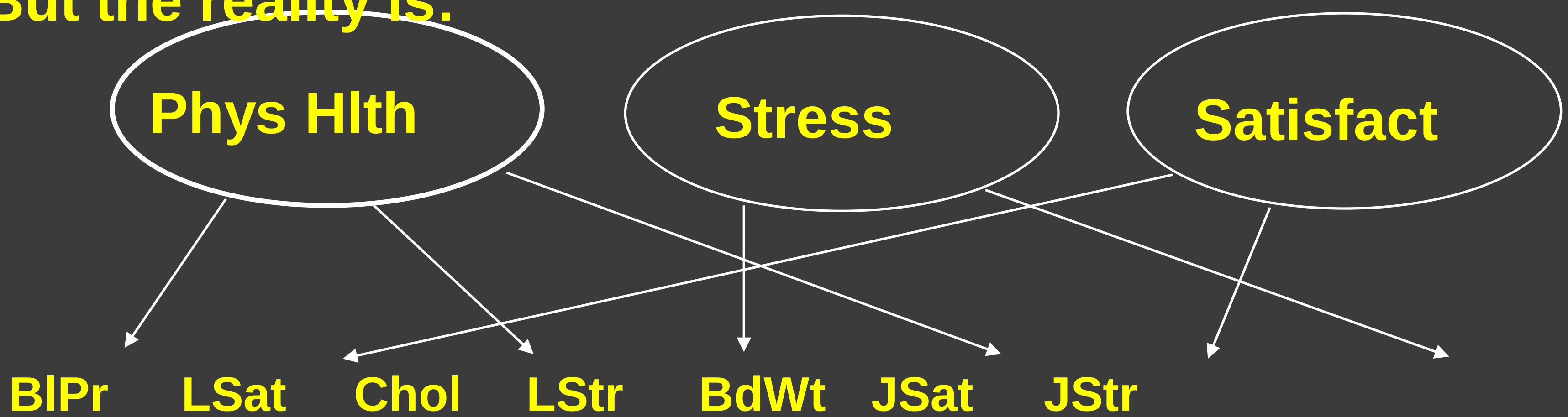
- Part of construct validation process (do the data conform to expectations regarding the underlying patterns?)
- Use SEM packages to perform CFA
- EFA with specified number of factors for a criterion is NOT a CFA
- Basically start with a correlation matrix and expected relationships
- Look at whether expected relationships can reproduce the correlation matrix well
- Tested with chi-square goodness of fit. If significant, data don’t fit expected structure. No confirmation.
- Alternative measures of fit available.

Logic of CFA

Let’s say I believe:



But the reality is:

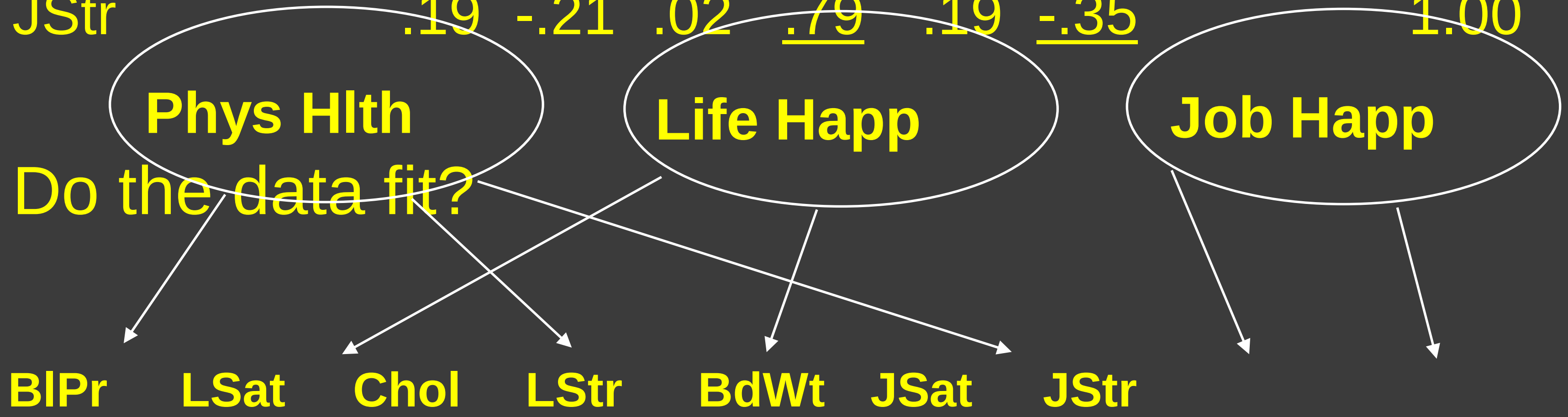


Data won’t confirm expected structure

Example

R matrix (correlation matrix)

	BlPr	LSat	Chol	LStr	BdWt	JSat	JStr	
BlPr	1.00							
LSat	-.18	1.00						
Chol	.65	-.17	1.00					
LStr	.15	-.45	.22	1.00				
BdWt	.45	-.11	.52	.16	1.00			
JSat	-.21	.85	-.12	-.35	-.05	1.00		
JStr		.19	-.21	.02	.79	.19	-.35	1.00



Do the data fit?