Introduction to exploratory common factor analysis (EFA)

Measurement in Clinical Research: Epidemiology 225

Steve Gregorich

October 29, 2009

What are the goals of the common factor model?

Provide data reduction: Represent a set of observed variables (or items) with a more parsimonious set of related constructs (AKA common factors, latent variables)

Test construct validity: Do the items measure what they are hypothesized to measure?

To provide empirical justification for creating summated composite scores, or 'scale scores,' which are more reliable than individual item scores

Overview: Common factor model

What is a common factor model?

Indirect measurement

Some constructs are not directly observable

attitudes, intelligence, economic strength, top quark

These are sometimes called *latent* variables

. Latent variables are 'everywhere' (physics, medicine, economics)

It is sometimes possible to assess latent variables indirectly,

via multiple, fallible, observed—or *manifest*—variables

A *measurement model* relates latent variables to manifest variables.

That is, the latent variables are hypothesized to directly cause responses to corresponding manifest variables

With multiple manifest variables per latent variable, the measurement model can be empirically evaluated, via *common factor analysis*

(define 'common')

Conceptual example

Suppose I want to measure two dimensions of consumer confidence

Expectations for 6-months hence

. Business conditions (1 = worse; 2 = same; 3 = better)

. Employment (1 = fewer jobs; 2 = same; 3 = more jobs)

. Income (1 = decrease; 2 = same; 3 = increase)

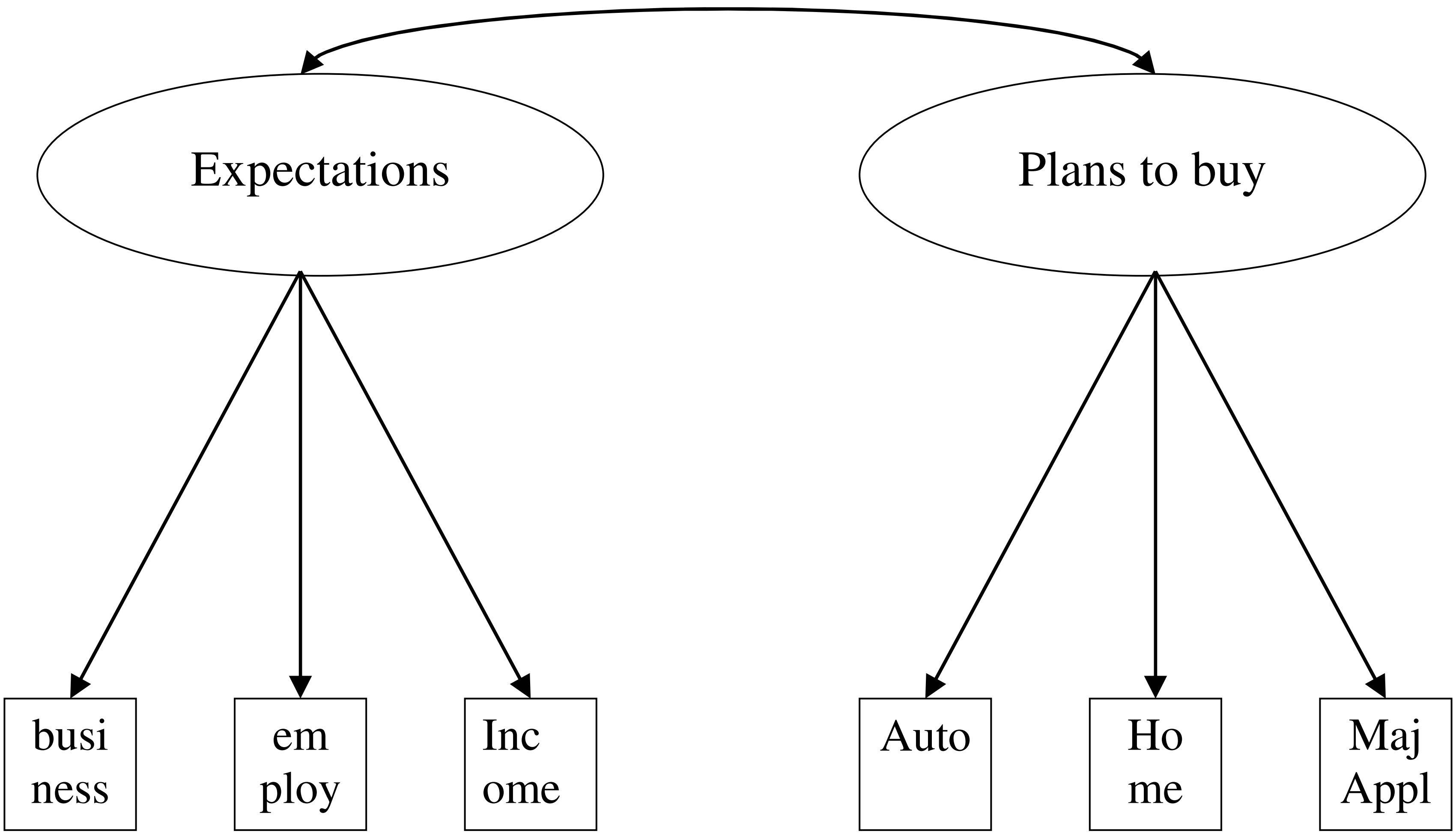
Plans to buy within 6-months

. Automobile

. Home

. Major appliances

Conceptual example: Hypothesized measurement model



(define single- and double-headed arrows)

Overview: Made-up example of a factor model

A generic representation of a factor pattern matrix
with 2 common factors and 6 manifest variables

	Expectations	Plans to buy
business	.67	.12
employment	.54	.11
income	.55	.07
auto	.05	.77
house	.09	.89
major appl.	.10	.57

The factor pattern matrix holds estimated correlations between
latent and manifest variables

The latent variables are estimated from the observed data

Correlations between latent and manifest variables aid interpretation

Question: Is the interpretation consistent with the motivating theory?

Wait a minute…

How is it possible to estimate the relationship between something measured (items) and something not measured (factors)?

<u>Start with input data</u>

The input data for a factor analysis are usually the observed correlations or covariances among the observed items

<u>Estimate factor loadings for your hypothesized model (an iterative search)</u>

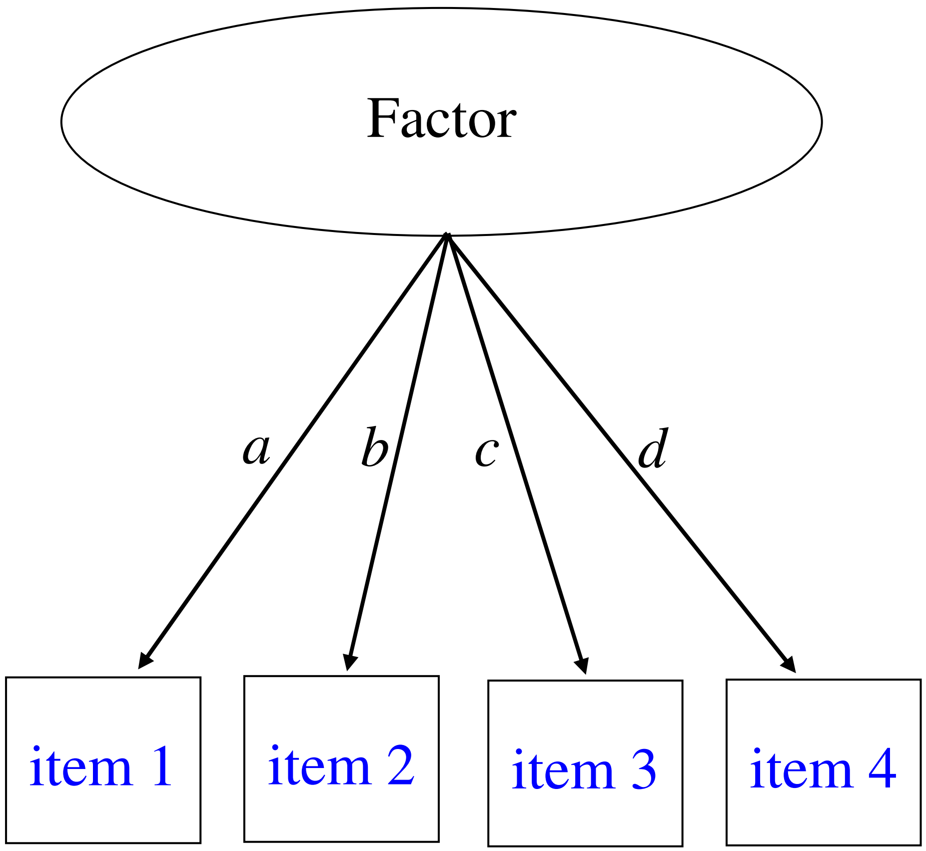
A well-fitting factor model and estimates can be used to accurately reproduce the input data

<u>Compare the model-reproduced data to the original data</u>

Good correspondence between the two suggests that the model has 'good fit' and we have more confidence in the model and estimates

Relationship between standardized factor loadings and item correlations

Factor model and loading estimates



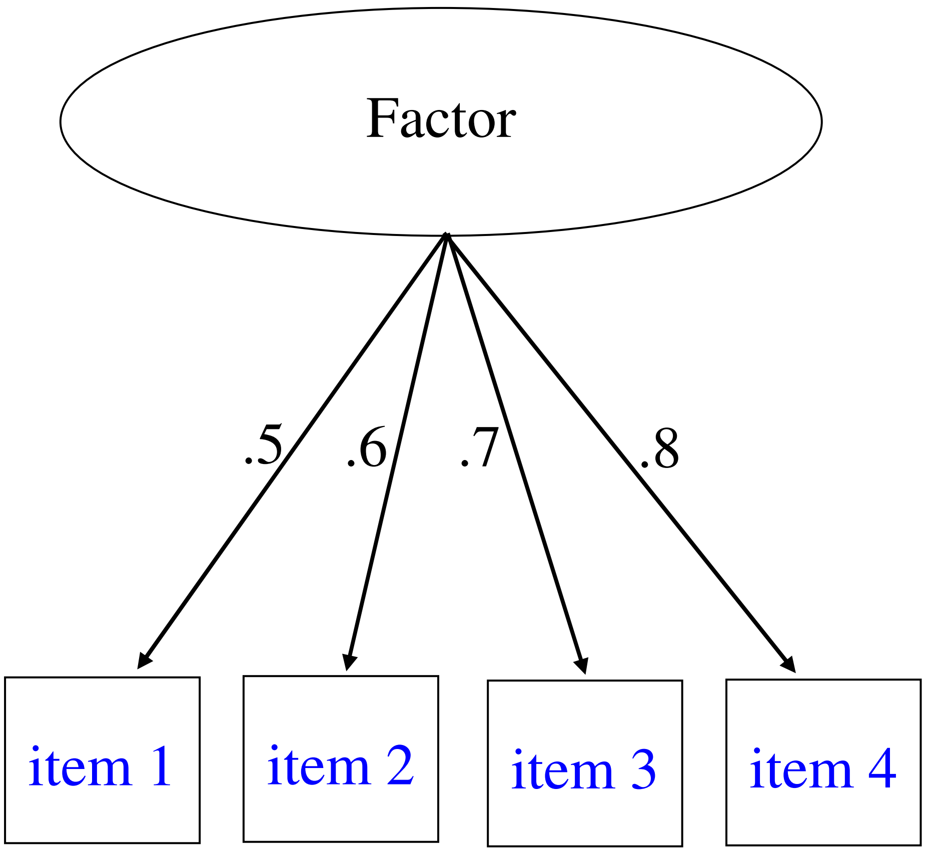
Model-implied item correlations

	item 1	item 2	item 3	item 4
item 1	1.0			
item 2	$a \times b$	1.0		
item 3	$a \times c$	$b \times c$	1.0	
item 4	$a \times d$	$b \times d$	$c \times d$	1.0

. 4 (a, b, c, and d) factor loadings attempt to explain 6 inter-item correlations

Relationship between standardized factor loadings and item correlations

Factor model and loading estimates



Model-implied item correlations

	item 1	item 2	item 3	item 4
item 1	1.0			
item 2	.30	1.0		
item 3	.35	.42	1.0	
item 4	.40	.48	.56	1.0

Empirical question

Do the model-implied correlations approximate the observed correlations?

Introduction: Steps in exploratory common factor analysis (EFA)

(1) initial choice of items to factor analyze

(2) respondent sampling and data collection

(3) compute matrix of inter-item correlations

(4) specify number of factors

(5) specify method of factor extraction

(6) specify method of factor rotation

(7) interpret/assess model:
- . Is the model substantively appealing?
- . Is the specified number of factors reasonable?
- . Are any items questionable?
- . Possibly re-specify number of factors and/or drop items and re-fit model

Introduction: Step 1. Initial choice of items to factor analyze

Choose
- . the constructs (common factors, latent variables) you want to measure
- . the items (observed or manifest variables) representing each construct

The basic structure of the hypothesized measurement model is represented by the hypothesized correspondence between
(a) items (manifest variables) and
(b) common factors (latent variables)

Should be based upon theory, or previous empirical findings

Introduction: Step 2. Sampling and data collection

Factor analysis requires that responses to a set of 'items' are collected from a sample of (usually) individuals

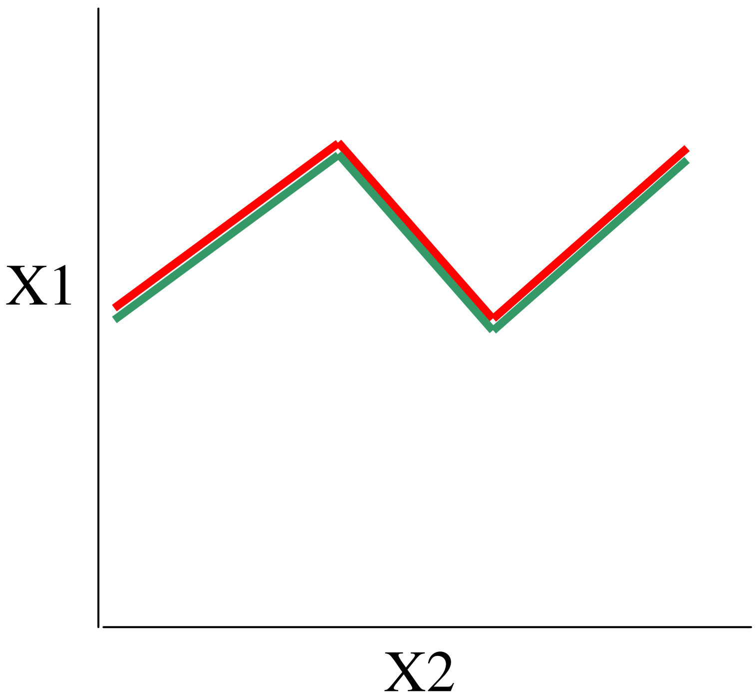
The sampling method should select individuals who are representative of the targeted population, i.e., generalizability of findings

Introduction: Step 3. Compute matrix of inter-item correlations

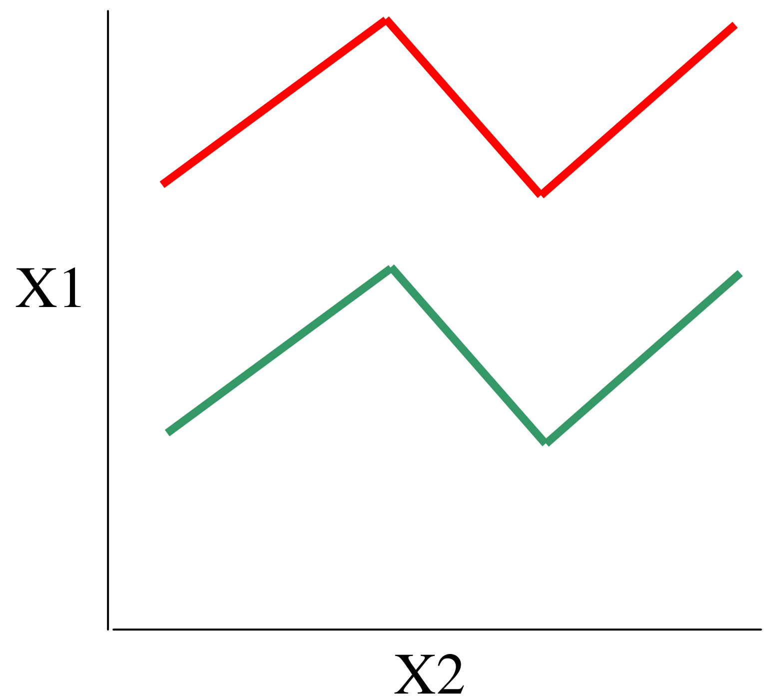
The Paul Kline reading misleadingly states that correlations assess the level of agreement between two variables.

There is a distinction between correlation and agreement

Correlation = 1.0
Agreement = 1.0



Correlation = 1.0
Agreement < 1.0



Introduction: Step 3. Compute matrix of inter-item correlations

Inter-item correlations between 4 made-up items

	x1	x2	x3	x4
x1	**1.0**	.56	.44	.35
x2	.56	**1.0**	.60	-.12
x3	.44	.60	**1.0**	.32
x4	.35	-.12	.32	**1.0**

Introduction: Step 4. Specify number of factors

Again, should be based upon theory, experience, previous findings

<u>Exploratory methods to empirically determine the number of factors</u>

. Eigenvalue > 1.0 rule

. Scree plot of eigenvalues

. χ^2 tests of model fit and fit indices (requires ML factor analysis)
- ML chi-square
- Fit indices: TLI, CFI
- others

So, what is an eigenvalue?
The variance that is explained by an eigenvector

OK, so what is an eigenvector?
The item correlation matrix can be transformed into
a set of mutually uncorrelated eigenvectors
(i.e., this is a fairly advanced topic in matrix algebra)

Introduction: Step 5. Choose factor extraction method

Principal components versus common factors

Both start with an item correlation or covariance matrix

Principal components

Decomposes *total* variation in the item correlation matrix
(the principal components *are* the eigenvectors)

Common factor analysis

Decomposes *common* variation in the item correlation matrix

. What is common variation/communality?

shared variance b/t a manifest variable & all other manifest variables

Communality estimation

. Squared multiple correlations (SMC)

. Iterated communality estimation

. ML communality estimation

Introduction: Step 6. Specify method of factor rotation

Extracted factors are uncorrelated and are usually difficult to interpret

Factor rotation hopefully allows for easier interpretation

<u>Orthogonal rotation: uncorrelated factors</u>

. VARIMAX (all factor analysis programs)
. many others

<u>Oblique rotation: correlated factors</u>

. PROMAX (SAS)
. Harris-Kaiser (SAS)
. Direct Oblimin (SPSS)
. many others

Example data

NHANES 1982-84 Epi Follow-up

Center for Epidemiologic Studies Depression scale (CES-D)

. White men aged 50+ with complete data on all 20 CES-D items
. N = 2004

The items of the CES-D are generally believed to represent 4 factors

factor	items
depressive affect	*blues, depressed, failure, fearful, lonely, cry, sad*
somatic symptoms	*bothered, appetite, mind, effort, sleep, talk, get going*
inter-personal	*unfriendly, dislike*
positive affect	*good, hopeful, happy, enjoy*

Example: Steps 1, 2, and 3.

Initial choice of items to factor analyze

How did Radloff chose items?

. She collected a list of common depressive symptoms

. Is this a good approach?

How you might choose from among the 20 CES-D items

. Previous research findings

. Your own theory

Respondent sampling and data collection (secondary data)

Compute matrix of inter-item correlations

Example: Step 4. specify number of factors

Options

. a priori choice: theory, prior empirical findings

. Eigenvalue > 1.0

. Scree plot

. Model fit tests, indices

Example: Step 4. specify number of factors

a priori choice

. Many investigators, but not all, have reported 4 factors

Example: Step 4. specify number of factors

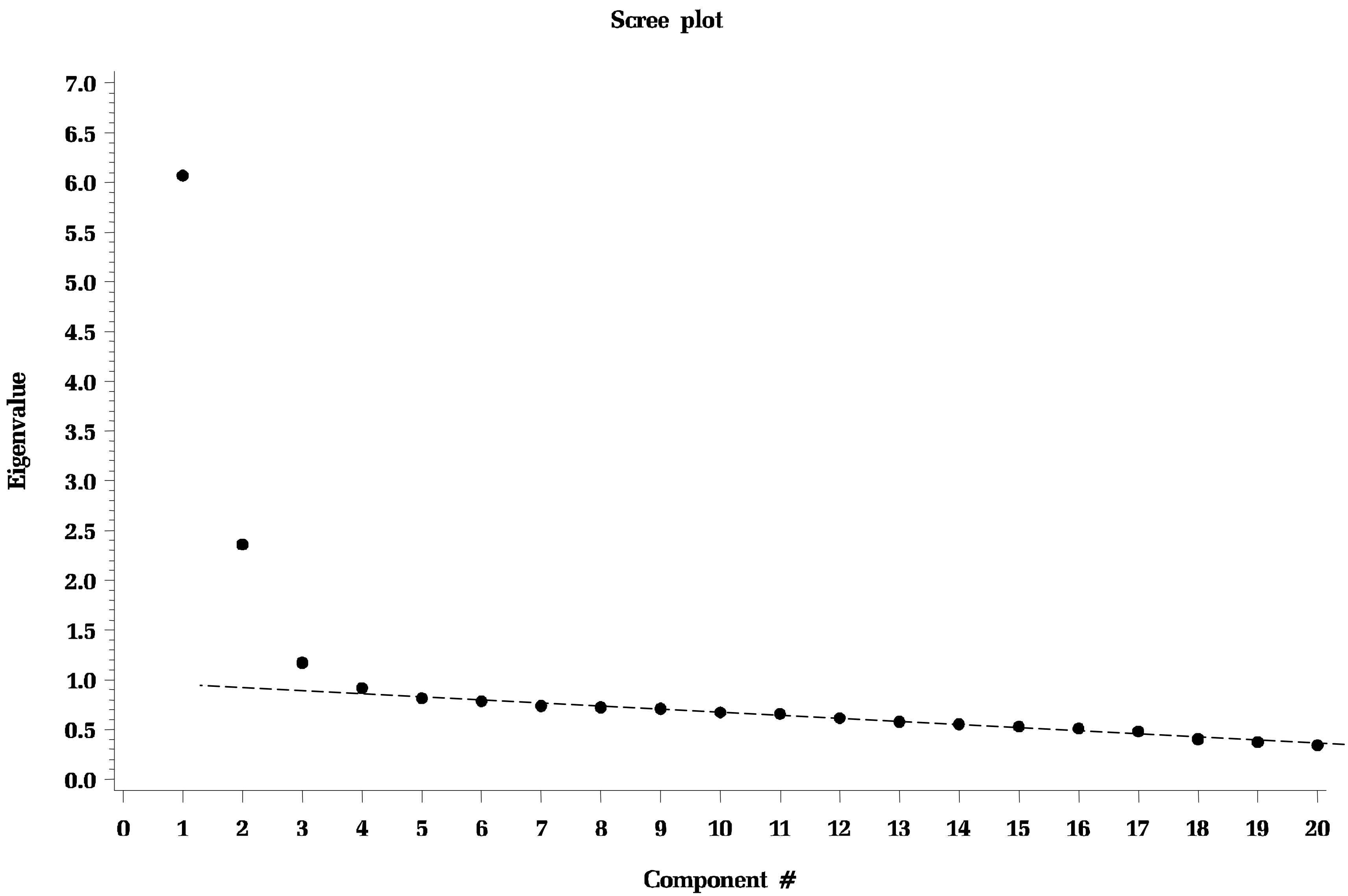
Choose number of factors to equal number of eigenvalues > 1.0

Eigenvalues of the item correlation matrix

	Eigenvalue	Proportion	Cumulative
1	6.06974392	0.3035	0.3035
2	2.36134193	0.1181	0.4216
3	1.17053740	0.0585	0.4801
4	0.91717920	0.0459	0.5259
5	0.81328821	0.0407	0.5666
6	0.78490631	0.0392	0.6058
7	0.73721336	0.0369	0.6427
8	0.72236274	0.0361	0.6788
9	0.70879857	0.0354	0.7143
10	0.67176187	0.0336	0.7479
11	0.65731560	0.0329	0.7807
12	0.61666653	0.0308	0.8116
13	0.57798797	0.0289	0.8405
14	0.55424913	0.0277	0.8682
15	0.52913967	0.0265	0.8946
16	0.50941923	0.0255	0.9201
17	0.48174892	0.0241	0.9442
18	0.40293658	0.0201	0.9643
19	0.37224703	0.0186	0.9829
20	0.34115581	0.0171	1.0000

Example: Step 4. specify number of factors

Scree plot



Example: Step 4. Specify number of factors

<u>Model fit</u>

This is only available with maximum likelihood (ML) factor analysis

There is a chi-square test that the number of factors is sufficient

If the test is significant, it suggests that the model fit is poor and that more factors need to be extracted

There are also various fit indices: common are the TLI and CFI

. TLI and CFI values generally range between 0 and 1

. Usually CFI/TLI values > .95 are thought to suggest approximate fit

# of factors	χ^2	*df*	*p*	TLI	CFI
1	8226	170	<.0001	.74	.77
2	3155	151	<.0001	.89	.91
3	1941	133	<.0001	.93	.95
4	1053	116	<.0001	.96	.97
5	745	100	<.0001	.96	.98

Example: Step 4. Specify number of factors

Summary

. Radloff and many others suggested 4 factors

. Eigenvalue > 1.0 suggested 3 factors

. Scree plot suggested 3 factors

. Model fit suggested *at least* 3 or 4 factors

We will consider models with 2 through 4 factors

Example: Steps 5 and 6

<u>Specify method of factor extraction</u>
I selected ML factor extraction
    It allows for the tests/indices of model fit, described above

<u>Specify method of factor rotation</u>
I selected Harris-Kaiser (oblique) rotation

For comparison, I also present a VARIMAX (orthogonal) rotation

Example: Step 7: assess the model

2 factors

		(Negative) Factor1	(Pos Aff) Factor2
cesd06	depressed	77	.
cesd18	sad	72	.
cesd03	blues	70	.
cesd14	lonely	68	.
cesd07	effort	61	.
cesd10	fearful	60	.
cesd20	get going	59	.
cesd05	mind	59	.
cesd17	cry	56	.
cesd01	bothered	53	.
cesd09	failure	52	.
cesd11	restless	49	.
cesd19	dislike	48	.
cesd02	appetite	48	.
cesd13	talk	47	.
cesd15	unfriendly	42	.
cesd12	happy	.	76
cesd16	enjoy	.	76
cesd08	hopeful	.	60
cesd04	good	17	55

Example: Step 7: assess the model

Inter-Factor Correlations

	Factor1	Factor2
Factor1	100	
Factor2	−30	100

Example: Step 7: assess the model

3 factors

		(Dep+Som) Factor1	(InterPers) Factor2	(Pos Aff) Factor3
cesd06	depressed	82	.	.
cesd03	blues	75	.	.
cesd07	effort	70	-10	.
cesd01	bothered	66	-16	.
cesd20	get going	61	.	.
cesd18	sad	59	17	.
cesd05	mind	58	.	.
cesd02	appetite	57	-11	.
cesd14	lonely	54	18	.
cesd11	restless	52	.	.
cesd10	fearful	46	19	.
cesd17	cry	45	15	.
cesd13	talk	38	11	.
cesd19	dislike	.	74	.
cesd15	unfriendly	.	62	.
cesd09	failure	22	40	.
cesd12	happy	.	.	76
cesd16	enjoy	.	.	76
cesd08	hopeful	.	.	60
cesd04	good	24	.	56

Example: Step 7: assess the model

Inter-Factor Correlations

	Factor1	Factor2	Factor3
Factor1	100		
Factor2	67	100	
Factor3	−30	−21	100

Example: Step 7: assess the model

4 factors

		(Somatic) Factor1	(InterPers.) Factor2	(Pos Aff) Factor3	(Dep Aff) Factor4
cesd07	effort	80	.	.	-11
cesd20	get going	74	.	.	-14
cesd05	mind	47	.	.	11
cesd11	sleep	44	.	.	.
cesd02	appetite	42	.	.	14
cesd01	bothered	35	-14	.	31
cesd13	talk	35	15	.	.
cesd19	dislike	.	69	.	.
cesd15	unfriendly	.	61	.	.
cesd09	failure	16	39	.	10
cesd16	enjoy	.	.	76	.
cesd12	happy	.	.	76	.
cesd08	hopeful	.	.	60	.
cesd04	good	11	.	55	13
cesd18	sad	-15	.	.	86
cesd17	cry	-27	.	.	84
cesd06	depressed	22	.	.	65
cesd03	blues	17	.	.	63
cesd14	lonely	.	11	.	60
cesd10	fearful	13	15	.	37

Example: Step 7: assess the model

Inter-Factor Correlations

	Factor1	Factor2	Factor3	Factor4
Factor1	100			
Factor2	54	100		
Factor3	-24	-18	100	
Factor4	82	63	-30	100

Example: Effects of other options

Unrotated factors: 4-factor model

		Factor1	Factor2	Factor3	Factor4
cesd06	depressed	77	.	-13	-11
cesd18	sad	73	.	.	-24
cesd03	blues	70	.	-11	-12
cesd14	lonely	67	.	.	-11
cesd07	effort	59	13	-23	26
cesd10	fearful	59	.	.	.
cesd20	get going	58	12	-15	27
cesd05	mind	56	12	.	13
cesd17	cry	56	.	.	-28
cesd01	bothered	53	.	-19	.
cesd09	failure	53	.	22	11
cesd19	dislike	49	.	47	13
cesd11	sleep	49	.	-11	12
cesd02	appetite	46	.	-16	.
cesd13	talk	44	10	.	13
cesd15	unfriendly	41	.	39	16
cesd16	enjoy	-32	69	.	.
cesd12	happy	-36	69	.	.
cesd08	hopeful	-19	55	.	.
cesd04	good	.	52	.	.

Example: Effects of other options: VARIMAX (orthogonal rotation)

		Factor1	Factor2	Factor3	Factor4
cesd07	effort	67	12	.	14
cesd20	get going	63	10	.	20
cesd06	depressed	57	52	-10	16
cesd05	mind	52	21	.	21
cesd03	blues	50	49	.	15
cesd01	bothered	48	29	.	.
cesd11	sleep	46	17	.	14
cesd02	appetite	45	19	.	.
cesd10	fearful	38	35	.	29
cesd13	talk	38	15	.	23
cesd18	sad	37	61	.	28
cesd17	cry	23	55	.	21
cesd14	lonely	40	48	.	29
cesd12	happy	-14	-13	75	.
cesd16	enjoy	-12	.	75	.
cesd08	hopeful	.	.	58	.
cesd04	good	.	.	52	.
cesd19	dislike	17	19	.	65
cesd15	unfriendly	18	12	.	56
cesd09	failure	31	22	.	43

17 cross-loadings > .20

Conclusions

Choice of number of factors should be based upon

Theoretical appeal, parsimony, clinical experience *as much as*
empirical model fit

Exploratory factor analysis can be confusing/trying/squishy

Sometimes the initial selection of items requires modification
(e.g., elimination of poor items).

This means that selection of both
(1) the items to be modeled *and*
(2) the number of factors
can be 'in play' at the same time

Confirmatory factor models also exist

. Specify both the number of factors and the item-to-factor configuration
. Also allows tests of whether factor model is invariant across groups

Conclusions

To assess the properties of multi-item measurement instruments

Always use factor analysis, not principal components analysis

Always use oblique rotation

Use ML factor analysis.

This allows a test of whether the specified number of factors is sufficient.

Be thoughtful about the measurement model

. Item creation and selection should be a deliberate process

. In many ways, Radloff got 'lucky'

Start with 'small' models

Conclusions

Creation of a measurement instrument with good psychometric properties represents programmatic work—not project work.

. Iterative modification with new samples

. Replication

. Testing in new population groups

Example: SAS PROC FACTOR code

The following SAS code will estimate the eigenvalues and create a scree plot

```
proc factor method=prin priors=one scree;
  var cesd01-cesd20;
run;
```

The following SAS code will fit a common factor model

(1) n=4: extraction of 4 factors
(2) method=ml: maximum likelihood factor extraction
(3) priors=smc: SMCs as initial communality estimates
(4) rotate=hk: Harris-Kaiser oblique factor rotation
(5) other options to enhance output (re, fuzz, round, flag)

```
proc factor n=4 method=ml priors=smc rotate=hk
            re fuzz=.1 round flag=99;
  var cesd01-cesd20;
run;
```