Principal Component Analysis and Exploratory Factor Analysis

Module 4: Data Issues, Assumptions, and Assessing Reliability

Karen Grace-Martin

Workshop Outline

1. Assumptions of PCA & EFA

2. Data requirements and issues
   - Reverse coding
   - Sample sizes
   - Normality, ordinal, and binary variables
   - Factorability of the Correlation Matrix
   - Missing Data

3. Assessing Scale Reliability and Validity

1. Assumptions

Assumptions of

Principal Component Analysis

1. The measured variables are themselves of interest
2. No measurement error
3. Variables appropriate for correlations
4. Linear relationships between all variables
5. Adequate Sample Size

Exploratory Factor Analysis

1. There are latent variables that inform the measured variables
2. Multivariate Normality (especially for ML extraction)
3. Variables appropriate for correlations
4. Linear relationships between all variables
5. Adequate Sample Size

2.1 Reverse Coding

Reverse Coding

LifeOrientBestR = 5 – LifeOrientBest.2

The General Formula:

reversed score = (minimum score) + (maximum score) – actual score

Variable Values

Value		Label
LifeOrientBest.2: In uncertain times, I usually expect the best	1	agree a lot
	2	agree a little
	3	disagree a little
	4	disagree a lot
LifeOrientWrong.2: If something can go wrong for me, it will	1	disagree a lot
	2	disagree a little
	3	agree a little
	4	agree a lot
LifeOrientOpt.2 I am always optimistic about my future	1	agree a lot
	2	agree a little
	3	disagree a little
	4	disagree a lot
LifeOrientMyWay. I am always optimistic about my future I hardly ever expect things to go my way	1	disagree a lot
	2	disagree a little
	3	agree a little
	4	agree a lot
LifeOrientCount.2 I rarely count on good things happening to me	1	disagree a lot
	2	disagree a little
	3	agree a little
	4	agree a lot
LifeOrientGood.2 Overall, I expect more good things to happen to me than bad.	1	agree a lot
	2	agree a little
	3	disagree a little
	4	disagree a lot

Reverse Coding

LifeOrientBestR = 5 – LifeOrientBest.2

The General Formula:

reversed score = (minimum score) + (maximum score) – actual score

Variable Values		
Value		Label
LifeOrientBest.2: In uncertain times, I usually expect the best	1	agree a lot
	2	agree a little
	3	disagree a little
	4	disagree a lot
LifeOrientBest.R: In uncertain times, I usually expect the best	1	disagree a lot
	2	disagree a little
	3	agree a little
	4	agree a lot

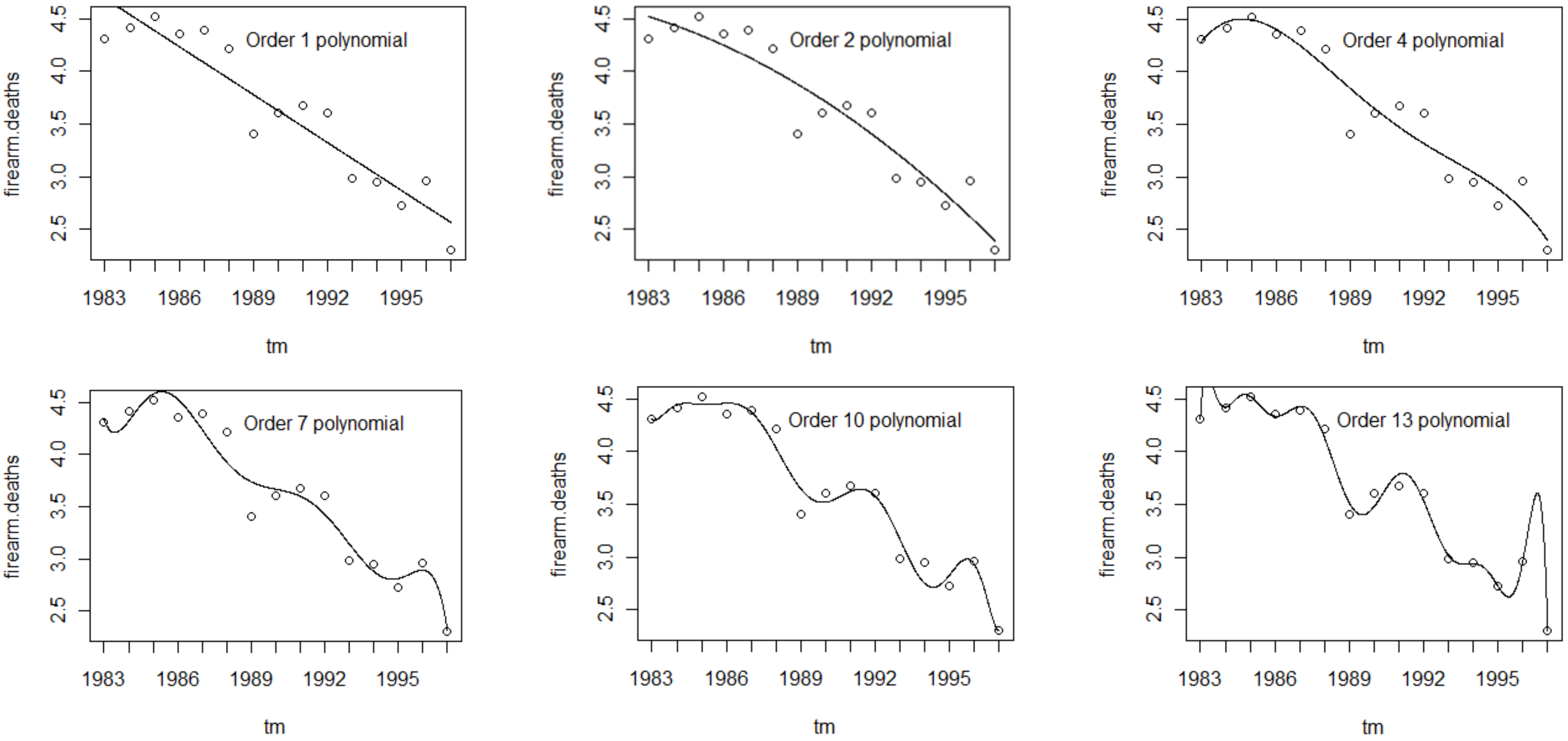
LifeOrientMyWay. I am always optimistic about my future I hardly ever expect things to go my way	1	disagree a lot
	2	disagree a little
	3	agree a little
	4	agree a lot

Correlations

	LifeOrientBest.2	LifeOrientBestR
LifeOrientMyWay.2	.374	-.374

2.2 Sample Size

Overfitting



Overfitting

What is it, exactly?

Creating a model that is too complex for the amount of data.

- Loadings are too large
- Too many loadings are non-zero

It appears to predict well with the existing data set, but…

- it does not fit future observations
- it does not replicate

Minimum Sample Size Suggestions

Observations per Variable:

- 10-15 Observations per variable (Pett, Lackey, & Sullivan)
- 10 Observations per variable (Nunnally, 1978)
- 5 Observations per variable or 100 observations, whichever is larger (Hatcher, 1994)
- 2 Observations per variable (Kline, 1994)

Observations per Factor:

- 20 Observations per factor (Arrindel & van der Ende, 1985)

Absolute number of Observations:

- 100 Observations=sufficient if clear structure; more is better (Kline, 1994)
- 100 Observations=poor; 300=good; >1000=excellent (Comrey & Lee, 1992)
- 300 Observations, though fewer ok if high correlations (Tabachnik & Fidell, 2001)

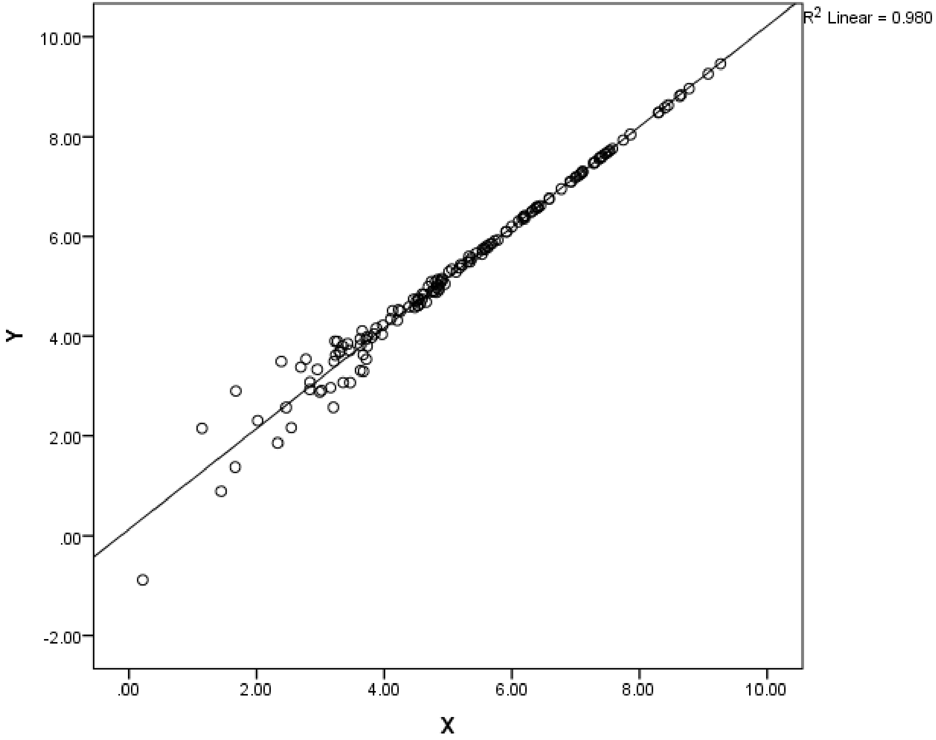
Required Sample Size is affected by:

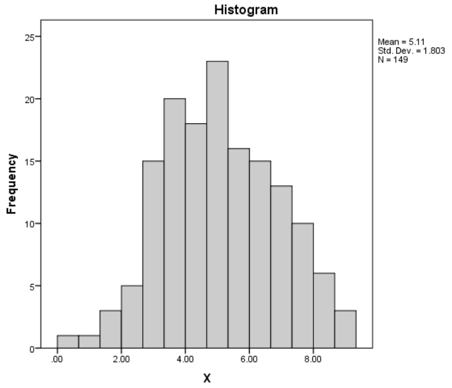
- Number of variables
- Number of factors
- Size and cleanliness of loadings onto factors
- Number of items per factor
- Missing data
- Measurement error

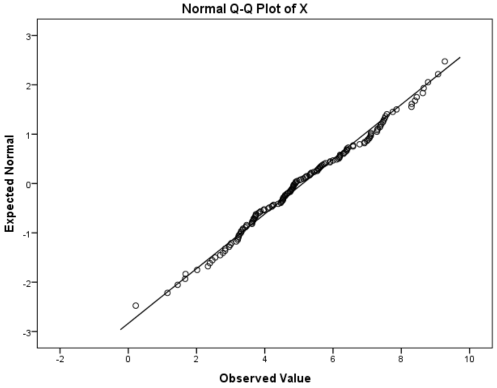
2.3 Normality, Ordinal, and Binary Variables

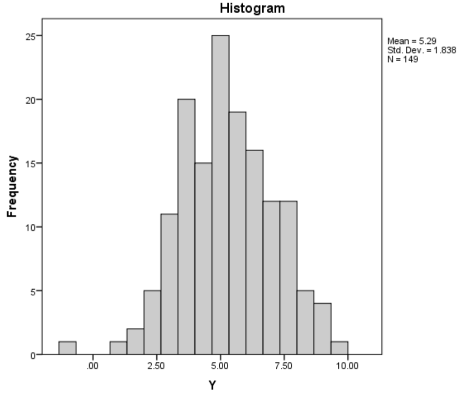
Pearson Correlation

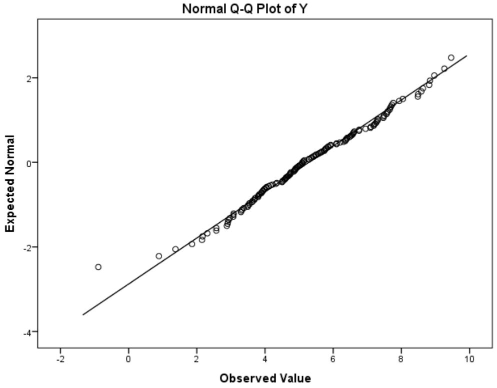
$$r = \frac{\sum_{i-1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y}$$

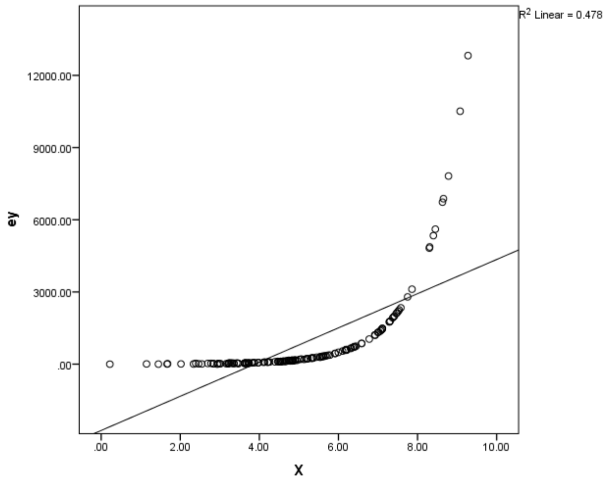



Histogram
Mean = 5.11
Std. Dev. = 1.803
N = 149
Frequency
X

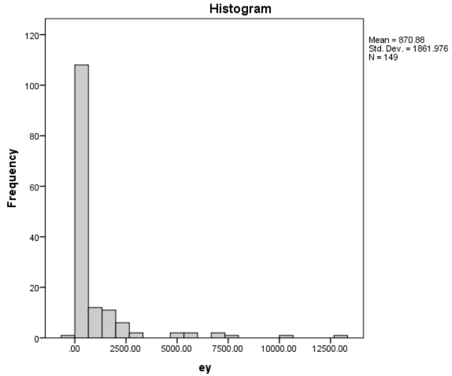

Normal Q-Q Plot of X
Expected Normal
Observed Value


Histogram
Mean = 5.29
Std. Dev. = 1.838
N = 149
Frequency
Y

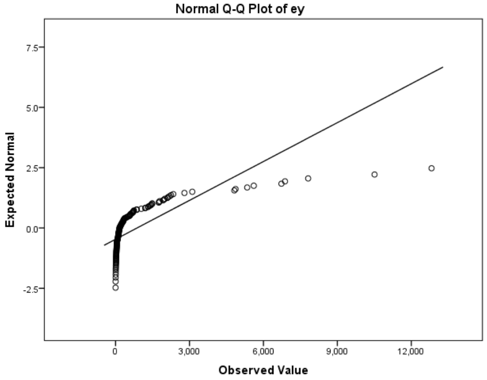

Normal Q-Q Plot of Y
Expected Normal
Observed Value



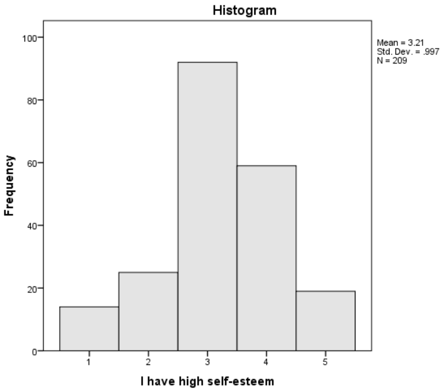
R2 Linear = 0.478
12000.00
9000.00
6000.00
3000.00
.00
ey
.00
2.00
4.00
6.00
8.00
10.00
X

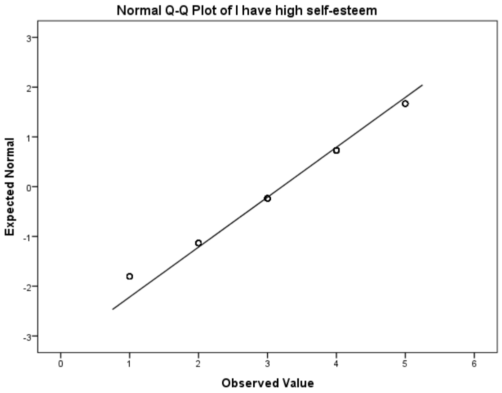


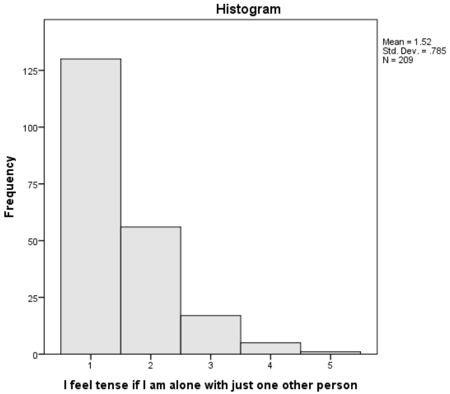
Histogram
Mean = 870.88
Std. Dev. = 1861.976
N = 149
120
100
80
60
40
20
0
Frequency
.00
2500.00
5000.00
7500.00
10000.00
12500.00
ey

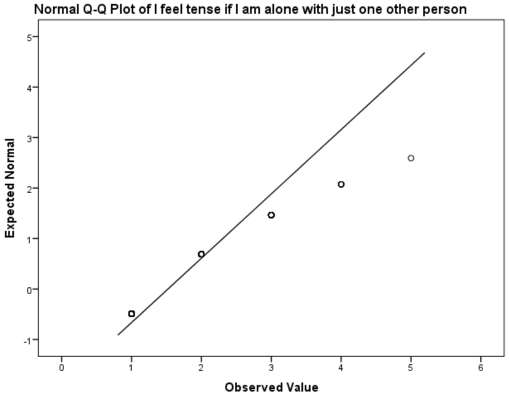


Normal Q-Q Plot of ey
7.5
5.0
2.5
0.0
-2.5
Expected Normal
0
3,000
6,000
9,000
12,000
Observed Value

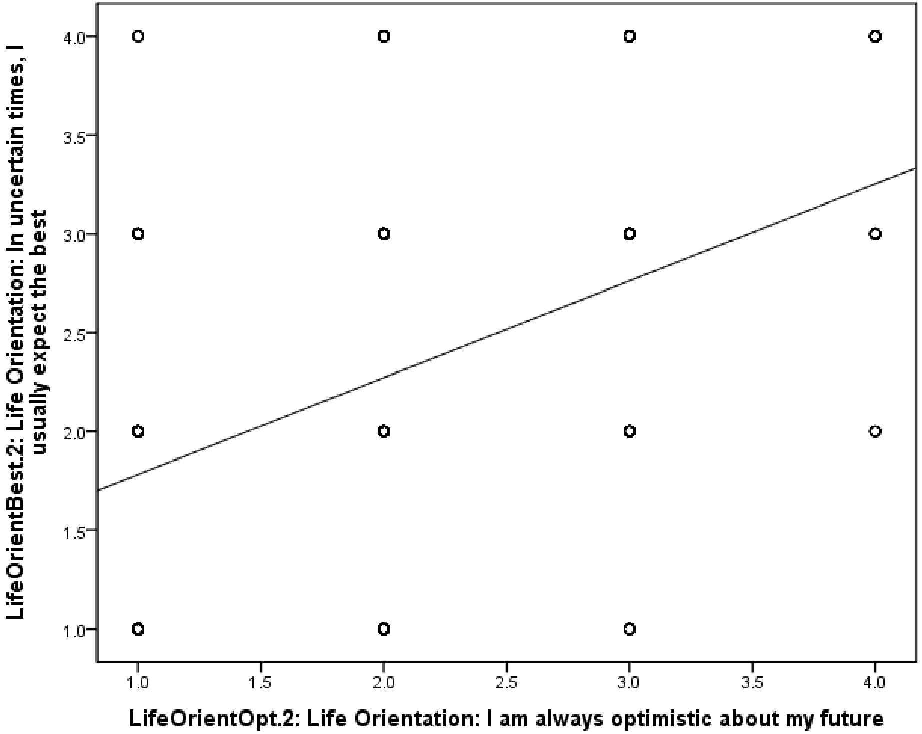

Histogram
Mean = 3.21
Std. Dev. = .997
N = 209
Frequency
I have high self-esteem


Normal Q-Q Plot of I have high self-esteem
Expected Normal
Observed Value


Histogram
Mean = 1.52
Std. Dev. = .785
N = 209
Frequency
I feel tense if I am alone with just one other person


Normal Q-Q Plot of I feel tense if I am alone with just one other person
Expected Normal
Observed Value

When Items are Ordinal or Binary



Tetrachoric and Polychoric Correlations

$$r_{tet} = \cos \frac{180^\circ}{1 + \sqrt{BC / AD}}$$

CrossTabulation			
		Highway	
		0 No	1 Yes
Rural	0 No	7 A	13 B
	1 Yes	28 C	18 D

Tetrachoric and Polychoric Correlations

Crosstabulation

Counts

		LifeOrientOpt.2			
		1 agree a lot	2 agree a little	3 disagree a little	4 disagree a lot
LifeOrientBest.2	1 agree a lot	89	63	10	0
	2 agree a little	155	225	60	6
	3 disagree a little	28	129	98	11
	4 disagree a lot	4	21	33	18

Pearson Correlations

	LifeOrientBest.2	LifeOrientOpt.2
LifeOrientBest.2	1	.473
LifeOrientOpt.2	.473	1

Polychoric Correlations

	LifeOrientBest.2	LifeOrientOpt.2
LifeOrientBest.2	1.000	.541
LifeOrientOpt.2	.541	1.000

To Get Polychoric Correlations for FA

R	*Psych* package, *fa* function with cor= “poly” option
Stata	1. user-written command *polychoric* to calculate the correlation matrix 2. Use as input for factor analysis
SAS	**Pre 9.4** *1.* *Proc freq* to calculate the polychoric correlation matrix 2. Use as input for factor analysis **v. 9.4** Outplc= option in *proc corr* saves the matrix as data
SPSS	Install R HetCor Extension into SPSS 1. HetCor R extension to calculate the correlation matrix http://www-01.ibm.com/support/docview.wss?uid=swg21477550 2. Use as input for factor analysis Or Basto & Pereira’s SPSS R-menu extension http://www.jstatsoft.org/v46/i04/paper

Input Polychoric Correlations

Pearson

Pearson Correlations					
	cHSRelief	cHSAdmir	cHSGetHelp	cHSOwn	cHSWorkOut
cHSRelief	1.000	.116	.635	.219	.056
cHSAdmir	.116	1.000	.243	.484	.294
cHSGetHelp	.635	.243	1.000	.444	.366
cHSOwn	.219	.484	.444	1.000	.541
cHSWorkOut	.056	.294	.366	.541	1.000

Polychoric

Polychoric Correlations					
	cHSRelief	cHSAdmir	cHSGetHelp	cHSOwn	cHSWorkOut
cHSRelief	1.000	.137	.744	.242	.030
cHSAdmir	.137	1.000	.315	.569	.319
cHSGetHelp	.744	.315	1.000	.509	.368
cHSOwn	.242	.569	.509	1.000	.609
cHSWorkOut	.030	.319	.368	.609	1.000

Input Polychoric Correlations

Pearson

Total Variance Explained

Factor	Initial Eigenvalues Total	Initial Eigenvalues % of Variance	Initial Eigenvalues Cumulative %	Extraction Sums of Squared Loadings Total	Extraction Sums of Squared Loadings % of Variance	Extraction Sums of Squared Loadings Cumulative %
1	2.399	47.973	47.973	1.818	36.354	36.354
2	1.194	23.877	71.850			
3	.725	14.500	86.351			
4	.393	7.856	94.207			
5	.290	5.793	100.000			

Extraction Method: Principal Axis Factoring.a

Factor Matrix[a]

	Factor 1
cHSRelief	.435
cHSAdmir	.475
cHSGetHelp	.715
cHSOwn	.763
cHSWorkOut	.556

Polychoric

Total Variance Explained

Factor	Initial Eigenvalues Total	Initial Eigenvalues % of Variance	Initial Eigenvalues Cumulative %	Extraction Sums of Squared Loadings Total	Extraction Sums of Squared Loadings % of Variance	Extraction Sums of Squared Loadings Cumulative %
1	2.584	51.685	51.685	2.064	41.270	41.270
2	1.254	25.073	76.758			
3	.690	13.791	90.549			
4	.294	5.885	96.434			
5	.178	3.566	100.000			

Extraction Method: Principal Axis Factoring.a

Factor Matrix[a]

	Factor 1
cHSRelief	.467
cHSAdmir	.532
cHSGetHelp	.772
cHSOwn	.818
cHSWorkOut	.547

2.4 Factorability of the Correlation Matrix

Factorability of the Correlation Matrix

Pearson Correlations					
	cHSRelief	cHSAdmir	cHSGetHelp	cHSOwn	cHSWorkOut
cHSRelief	1.000	.116	.635	.219	.056
cHSAdmir	.116	1.000	.243	.484	.294
cHSGetHelp	.635	.243	1.000	.444	.366
cHSOwn	.219	.484	.444	1.000	.541
cHSWorkOut	.056	.294	.366	.541	1.000

We need correlations:

- Not too low
- Not too high
- For variables that are not redundant

Factorability of the Correlation Matrix

Determinant > 0

- Matrix has an inverse
- They are important in calculating eigenvalues and eigenvectors

Positive Definite:

- The matrix contains as much information as is implied.
- The last eigenvalue will be positive
- Negative Eigenvalues: Possible in FA, not in PCA

Factorability of the Correlation Matrix

Pearson Correlations					
	cHSRelief	cHSAdmir	cHSGetHelp	cHSOwn	cHSWorkOut
cHSRelief	1.000	.116	.635	.219	.056
cHSAdmir	.116	1.000	.243	.484	.294
cHSGetHelp	.635	.243	1.000	.444	.366
cHSOwn	.219	.484	.444	1.000	.541
cHSWorkOut	.056	.294	.366	.541	1.000

Determinant = .236

Test for Basic Assumptions – Sampling Adequacy

Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy

Bartlett's Test of Sphericity

Kaiser-Meyer-Olkin (KMO)

- Marvelous - - - - - - .90s
- Meritorious - - - - - .80s
- Middling - - - - - - - .70s
- Mediocre - - - - - - - .60s
- Miserable - - - - - - .50s
- Unacceptable - - - below .50

Varies from 0 to 1

Indicates whether or not the variables are able to be grouped into a smaller set of underlying factors

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and psychological measurement*, *34*(1), 111-117.

Testing Factorability of the Correlation Matrix

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.626
Bartlett's Test of Sphericity	Approx. Chi-Square	130.587
	df	10
	Sig.	.000

- Marvelous - - - - - - .90s
- Meritorious - - - - - .80s
- Middling - - - - - - - .70s
- Mediocre - - - - - - - .60s
- Miserable - - - - - - .50s
- Unacceptable - - - below .50

Anti-Image Correlations

	cHSRelief	cHSAdmir	cHSGetHelp	cHSOwn	cHSWorkOut
cHSRelief	.504[a]	.005	-.639	-.024	.230
cHSAdmir	.005	.729[a]	-.026	-.379	-.036
cHSGetHelp	-.639	-.026	.603[a]	-.199	-.271
cHSOwn	-.024	-.379	-.199	.692[a]	-.400
cHSWorkOut	.230	-.036	-.271	-.400	.643[a]

a. Measures of Sampling Adequacy(MSA)

Testing Factorability of the Correlation Matrix

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.626
Bartlett's Test of Sphericity	Approx. Chi-Square	130.587
	df	10
	Sig.	.000

Null hypothesis: correlation matrix is an identity matrix.

Significant result indicates matrix is not an identity matrix.

Pearson Correlations

	cHSRelief	cHSAdmir	cHSGetHelp	cHSOwn	cHSWorkOut
cHSRelief	1.000	.116	.635	.219	.056
cHSAdmir	.116	1.000	.243	.484	.294
cHSGetHelp	.635	.243	1.000	.444	.366
cHSOwn	.219	.484	.444	1.000	.541
cHSWorkOut	.056	.294	.366	.541	1.000

What to do with an Ill-Conditioned Matrix

Check:

- correlations of items with each other
- for duplicate records in the data
- including item totals along with individual items
- for subjects with similar sets of responses
- that you have sufficient subjects per item

2.5 Missing Data

Missing Data

- Listwise Deletion
- Pairwise Deletion
- Base the Factor Analysis on EM Correlation Matrix
- Multiple Imputation

Don’t use: mean imputation

Missing Data

Listwise Deletion:
Drop a case if any values are missing on any variable

Pairwise Deletion:
Drop a case from each correlation if any values are missing only on one of the two variables used in that specific correlation

Pearson Correlations[a]

	cHSRelief	cHSAdmir	cHSGetHelp	cHSOwn	cHSWorkOut
cHSRelief	1.000	.116	.635	.219	.056
cHSAdmir	.116	1.000	.243	.484	.294
cHSGetHelp	.635	.243	1.000	.444	.366
cHSOwn	.219	.484	.444	1.000	.541
cHSWorkOut	.056	.294	.366	.541	1.000

a. Listwise N=94

Pearson Correlations[a]

		cHSRelief	cHSAdmir	cHSGetHelp	cHSOwn	cHSWorkOut
cHSRelief	Pearson Correlation	1	.108	.635	.206	.048
	N	96	96	95	96	95
cHSAdmir	Pearson Correlation	.108	1	.242	.486	.296
	N	96	96	95	96	95
cHSGetHelp	Pearson Correlation	.635	.242	1	.443	.366
	N	95	95	95	95	94
cHSOwn	Pearson Correlation	.206	.486	.443	1	.556
	N	96	96	95	96	95
cHSWorkOut	Pearson Correlation	.048	.296	.366	.556	1
	N	95	95	94	95	95

Missing Data

EM Algorithm: gives unbiased correlation estimates with MAR missing data (see Graham, 2009)

- in SPSS MVA

EM Correlations

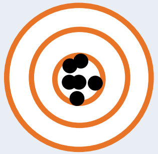
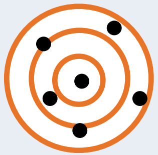
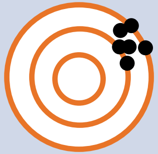
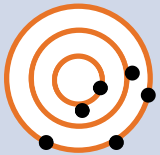
	cHSRelief	cHSAdmir	cHSGetHelp	cHSOwn	cHSWorkOut
cHSRelief	1				
cHSAdmir	.108	1			
cHSGetHelp	.634	.243	1		
cHSOwn	.206	.486	.442	1	
cHSWorkOut	.047	.296	.365	.556	1

3. Assessing Reliability and Validity

Relationship Between Reliability and Validity

If we used the scale again, would it yield the same results?

Does the scale measure what we intend to?

		Reliability (Precision)	
Validity (Accuracy)		High	Low
	High		
	Low		

38

Common Types of:

Reliability

- Test – Retest
- Alternate Form
- Split Half
- Parallel
- Inter-rater or Intra-rater Reliability
- Internal Consistency

Validity

- Face/content
- Response process
- Criterion
- Construct
- Convergent
- Discriminant

Measures of Internal Consistency

- Cronbach's alpha
- Variations on Cronbach's alpha
  - Split half correlation with Brown-Spearman adjustment
  - Kuder-Richardson 20
- Only used for composite measurements

Assessing Scale Reliability

Cronbach's α

$$\alpha = \left(\frac{N}{N-1}\right)\frac{S^2 - \Sigma s_i^2}{S^2}$$

Where S^2 = variance of summated scores and
Σs_i^2 =sum of individual variances.

Assumptions:

- All items describe a single factor
- All items contribute equally

Criticisms of Cronbach’s Alpha

- Not a substitute for other methods for assessing reliability
- Affected by number of items
- Not a measure of unidimensionality or validity
- Not useful for scale purification

Scale Purification

By convention:

.80 good

.70 adequate

.60 lenient cutoff is common in exploratory research

Reliability Statistics

Cronbach's Alpha	N of Items
.719	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
cHSRelief	10.89	8.999	.359	.720
cHSAdmir	11.12	8.900	.384	.709
cHSGetHelp	10.40	7.620	.625	.608
cHSOwn	10.82	7.913	.610	.617
cHSWorkOut	10.77	9.192	.433	.689

Cronbach's Alpha Recommendations

1. Always try to get test-retest or inter-rater reliability
2. Use confirmatory factor analysis for
    1. Unidimensionality
    2. Scale purification
3. Put it in only if you are forced to

Reporting Reliability Results

A questionnaire was employed to measure different, underlying constructs. One construct, 'Attitude towards counseling', consisted of five questions. The scale had a moderate level of internal consistency, as determined by a Cronbach's alpha of 0.719.

122 Types of Validity

One Hundred and Twenty-Two Kinds of Validity for Measurement

Administrative	Descriptive	Instructional	Rational
Artifactual	Design	Internal test	Raw
Behavior domain	Diagnostic	Internal	Relational
Cash	Differential	Interpretative	Relevant
Cluster domain	Direct	Interpretive	Representational
Cognitive	Discriminant	Intrinsic	Response
Common sense	Discriminative	Intrinsic content	Retrospective
Concept	Domain	Intrinsic correlational	Sampling
Conceptual	Domain-selection	Intrinsic rational	Scientific
Concrete	Edumetric	Item	Scoring
Concurrent	Elaborative	Job component	Self-defining
Concurrent true	Elemental	Judgmental	Semantic
Congruent	Empirical	Linguistic	Single-group
Consensual	Empirical-judgmental	Local	Site
Consequential	Etiological	Logical	Situational
Construct	External test	Longitudinal	Specific
Constructor	External	Lower-order	Structural
Content	Extratest	Manifest	Substantive
Context	Face	Natural	Summative
Contextual	Factorial	Nomological	Symptom
Convergent	Fiat	Occupational	Synthetic
Correlational	Forecast true	Operational	System
Criterion	Formative	Performance	Systemic
Cross-age	Functional	Practical	Theoretical
Cross-cultural	General	Predictive	Trait
Cross-sectional	Generalized	Predictor	Translation
Cultural	Generic	Procedural	Treatment
Curricular	Higher-order	Prospective	True
Decision	Incremental	Psychological and logical	User
Definitional	Indirect	Psychometric	Washback
Derived	Inferential		

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*(3), 301.

Types of Validity

Content-Related (appropriate content)	Criterion Related (relationship to other measures)
Face Validity: Does the scale appear to measure what it aims to?	**Concurrent Validity:** Does the measure relate to an existing similar measure?
Construct Validity: Does the measure relate to underlying theoretical concepts?	**Predictive Validity:** Does the measure predict later performance on related criterion?

Recommendations for next steps

1. Check validity
2. Check internal consistency, ideally via CFA
3. Check other forms of reliability
4. If there are any changes to be made to the items, revise, collect a new sample and re-run EFA.

   Repeat steps 1-3 until no further changes need to be made.

5. Collect a new sample and run a confirmatory factor analysis
6. Publish your scale, with results from the EFA and CFA