Principal Components and Factor Analysis

PSY 5102: Advanced Statistics for Psychological and Behavioral Research 2

Goals

- What is factor analysis?
- What are factors?
- Representing factors
  - Graphs and equations
- Extracting factors
  - Methods and criteria
- Interpreting factor structures
  - Factor rotation
- Reliability
  - Cronbach's alpha

When and Why Do We Use Factor Analysis?

- Take many variables and explain them with a few "factors" or "components"
- To see whether different measures are tapping aspects of a common dimension

General Steps in Factor Analysis

- Step 1: Select and measure a set of variables in a given domain
- Step 2: Screen data in order to prepare the correlation matrix
- Step 3: Factor extraction
- Step 4: Factor rotation to increase interpretability
- Step 5: Interpretation of factors
- Further Steps: Validate and determine reliability of the scales

Correlation Matrix

	Talk 1	Social Skills	Interest	Talk 2	Selfish	Liar
Talk 1	1.000					
Social Skills	.772	1.000				
Interest	.646	.879	1.000			
Talk 2	.074	−.120	.054	1.000		
Selfish	−.131	.031	−.101	.441	1.000	
Liar	.068	.012	.110	.361	.277	1.000

Factor 1

Factor 2

- In factor analysis, we look to reduce the correlation matrix into smaller sets of *uncorrelated* dimensions

What is a Factor?

- If several variables correlate highly, they might measure aspects of a common underlying dimension
  - These dimensions are called factors
- Factors are classification axes along which the measures can be plotted
  - The greater the loading of variables on a factor, the more that factor explains relationships between those variables

What is a "Good" Factor?

- A good factor should…
  - Make sense
  - Be easy to interpret
  - Have a simple structure
  - Lack complex loadings

Problems with Factor Analysis

- Unlike many of the other analyses we have covered, there is no statistical criterion to serve as a comparison for the linear combination
- It is more art than science
  - There are a number of extraction methods (e.g., principal-components analysis)
  - There are a number of rotation methods (e.g., Orthogonal, Oblique)
  - Choice of the number of factors to extract
  - Communality estimates

Types of Factor Analysis

- Exploratory factor analysis
  - Summarizing data by grouping correlated variables
  - Investigating sets of measured variables related to theoretical constructs
  - Usually done near the onset of research
  - This is the type of factor analysis that we will address

Types of Factor Analysis

- Confirmatory Factor Analysis
  - More advanced technique
  - When factor structure is known…or at least theorized
  - This basically involves testing the generalization of a factor structure to new data
  - This is tested through Structural Equation Modeling which we will discuss later

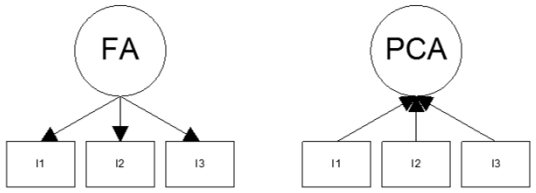
Basic Terminology

- Orthogonal Rotation
  - Loading Matrix: correlation between each variable and the factor
- Oblique Rotation
- Factor Correlation Matrix: correlations between the factors
  - Structure Matrix: correlation between factors and variables
  - Pattern Matrix: unique relationship between each factor and variable uncontaminated by overlap between the factors
- Factor Coefficient matrix: coefficients used to calculate factor scores (like regression coefficients)

Factor Analysis vs. Principal-Components Analysis

- FA produces factors
  - PCA produces components
- Factors cause variables
  - Components are aggregates of the variables
- FA analyzes only the variance shared among the variables (common variance without error or unique variance)
  - PCA analyzes all of the variance
- FA: "What are the underlying processes that could produce these correlations?"
  - PCA: Just summarize empirical associations, very data driven

Conceptual FA and PCA



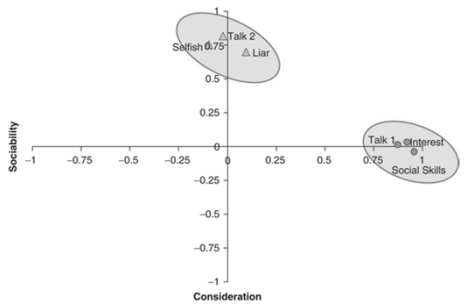
Principal-Components Analysis

- The goal of principal-components analysis is to identify a new set of a few variables that explain all (or nearly all) of the total variance
  - The goal is parsimony
- These principal-components are a linear function of the original variables
  - The first principal-component maximizes the amount of variance that is explained
- Eigenvector: the linear function
  - Their goal is to explain as much variability as possible
  - The number varies between analyses
- Eigenvalue: the total amount of variance that is explained by an eigenvector

How Many Factors to Extract?

- There are a number of strategies for deciding when to stop extracting factors or components
  - Percentage of variance that is explained
  - A specific number of factors may be extracted
  - Kaiser's stopping rule: only extracts (and retains) those factors with eigenvalues of at least 1
  - Scree test is a graphical procedure that depicts the eigenvalues

Graphical Representation



Mathematical Representation

$$Y = b_1X_1 + b_2X_2 \ldots b_nX_n$$
$$Factor_i = b_1Variable_1 + b_2Variable_2 \ldots b_nVariable_n$$

$$\begin{aligned} Y &= b_1X_1 + b_2X_2 \ldots b_nX_n \\ \text{Sociability} &= b_1\text{Talk1} + b_2\text{Social Skills} + b_3\text{Interest} \\ &\quad + b_4\text{Talk2} + b_5\text{Selfish} + b_6\text{Liar} \\ \text{Consideration} &= b_1\text{Talk1} + b_2\text{Social Skills} + b_3\text{Interest} \\ &\quad + b_4\text{Talk2} + b_5\text{Selfish} + b_6\text{Liar} \end{aligned}$$

Factor Loadings

- The *b* values in the equation represent the weights of a variable on a factor
- These values are the same as the co-ordinates on a factor plot
- They are called Factor Loadings
- These values are stored in a *Factor pattern matrix* (A)

$$A = \begin{pmatrix} 0.87 & 0.01 \\ 0.96 & -0.03 \\ 0.92 & 0.04 \\ 0.00 & 0.82 \\ -0.10 & 0.75 \\ 0.09 & 0.70 \end{pmatrix}$$

Practical Concerns

- How many participants?
  - Subjects-to-variables ratio should be at least 5-to-1 (5 participants for every variable)

The SPSS Anxiety Questionnaire

23 items that are intended to capture aspects of anxiety concerning statistical analyses

SD = Strongly Disagree, D = Disagree, N = Neither, A = Agree, SA = Strongly Agree

		SD	D	N	A	SA
1	Statistics make me cry	○	○	○	○	○
2	My friends will think I'm stupid for not being able to cope with SPSS	○	○	○	○	○
3	Standard deviations excite me	○	○	○	○	○
4	I dream that Pearson is attacking me with correlation coefficients	○	○	○	○	○
5	I don't understand statistics	○	○	○	○	○
6	I have little experience of computers	○	○	○	○	○
7	All computers hate me	○	○	○	○	○
8	I have never been good at mathematics	○	○	○	○	○
9	My friends are better at statistics than me	○	○	○	○	○
10	Computers are useful only for playing games	○	○	○	○	○
11	I did badly at mathematics at school	○	○	○	○	○
12	People try to tell you that SPSS makes statistics easier to understand but it doesn't	○	○	○	○	○
13	I worry that I will cause irreparable damage because of my incompetence with computers	○	○	○	○	○
14	Computers have minds of their own and deliberately go wrong whenever I use them	○	○	○	○	○
15	Computers are out to get me	○	○	○	○	○
16	I weep openly at the mention of central tendency	○	○	○	○	○
17	I slip into a coma whenever I see an equation	○	○	○	○	○
18	SPSS always crashes when I try to use it	○	○	○	○	○
19	Everybody looks at me when I use SPSS	○	○	○	○	○
20	I can't sleep for thoughts of eigenvectors	○	○	○	○	○
21	I wake up under my duvet thinking that I am trapped under a normal distribution	○	○	○	○	○
22	My friends are better at SPSS than I am	○	○	○	○	○
23	If I am good at statistics people will think I am a nerd	○	○	○	○	○

Initial Considerations

- The quality of analysis depends upon the quality of the data ("Garbage In" → "Garbage Out")
- Test variables should correlate quite well
  - $r > .3$
- Avoid Multicollinearity:
  - several variables highly correlated, $r > .80$
- Avoid Singularity:
  - some variables perfectly correlated, $r = 1$
- Screen the correlation matrix and eliminate any variables that obviously cause concern

Further Considerations

- Determinant: Indicator of multicollinearity
  - Should be greater than 0.00001
- Kaiser-Meyer-Olkin (KMO): Measures sampling adequacy
  - Should be greater than 0.5
- Bartlett's Test of Sphericity: Tests whether the R-matrix is an identity matrix
  - Should be significant at $p < .05$
- Anti-Image Matrix: Measures of sampling adequacy on diagonal
  - Off-diagonal elements should be small
- Reproduced: Correlation matrix after rotation
  - Most residuals should be < |0.05|

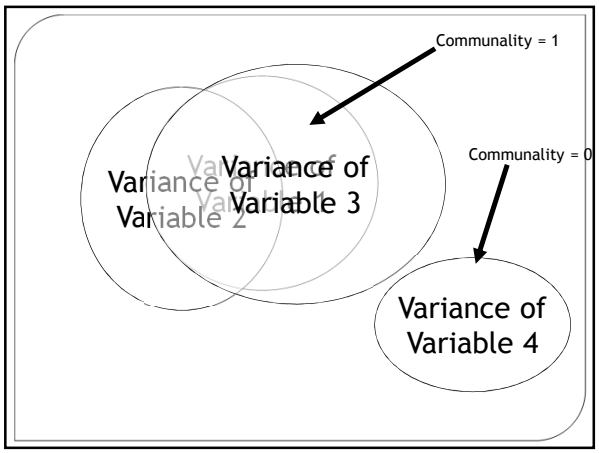
Correlation Matrix[a]

		Q01	Q02	Q03	Q04	Q08	Q19	Q20	Q21	Q22	Q23
Correlation	Q01	1.000	-.099	-.337	.436	.402	-.189	.214	.329	-.104	-.004
	Q02	-.099	1.000	.318	-.112	-.119	.203	-.202	-.208	.231	.100
	Q03	-.337	.318	1.000	-.380	-.310	.342	-.328	-.417	.224	.180
	Q04	.436	-.112	-.380	1.000	.401	-.188	.243	.410	-.098	-.034
	Q08	.402	-.119	-.310	.401	1.000	-.168	.200	.336	-.133	-.042
	Q06	.217	-.074	-.227	.278	.287	-.167	.101	.272	-.168	-.069
	Q07	.305	-.159	-.382	.409	.339	-.269	.221	.483	-.168	-.070
	Q08	.331	-.080	-.289	.349	.289	-.189	.178	.296	-.079	-.080
	Q09	-.092	.316	.300	-.128	-.096	.249	-.189	-.136	.287	.171
	Q10	.214	-.084	-.193	.216	.258	-.127	.084	.193	-.131	-.062
	Q11	.367	-.144	-.351	.369	.298	-.200	.255	.346	-.162	-.086
	Q12	.348	-.198	-.410	.442	.347	-.267	.286	.441	-.187	-.048
	Q13	.386	-.143	-.318	.344	.302	-.227	.204	.374	-.198	-.083
	Q14	.338	-.168	-.371	.381	.315	-.254	.226	.399	-.170	-.043
	Q15	.246	-.165	-.312	.334	.261	-.210	.206	.300	-.168	-.062
	Q16	.499	-.168	-.419	.416	.396	-.267	.266	.421	-.186	-.082
	Q17	.371	-.087	-.327	.383	.310	-.163	.205	.362	-.128	-.092
	Q18	.347	-.184	-.375	.382	.322	-.287	.238	.430	-.160	-.080
	Q19	-.189	.203	.342	-.196	-.168	1.000	-.249	-.275	.234	.122
	Q20	.214	-.202	-.328	.243	.200	-.249	1.000	.468	-.100	-.038
	Q21	.329	-.208	-.417	.410	.336	-.275	.468	1.000	-.129	-.068
	Q22	-.104	.231	.224	-.098	-.133	.234	-.100	-.129	1.000	.230
	Q23	-.004	.100	.180	-.034	-.042	.122	-.038	-.068	.230	1.000
Sig. (1-tailed)	Q01		.000	.000	.000	.000	.000	.000	.000	.000	.410
	Q02	.000		.000	.000	.000	.000	.000	.000	.000	.000
	Q03	.000	.000		.000	.000	.000	.000	.000	.000	.000
	Q04	.000	.000	.000		.000	.000	.000	.000	.000	.043
	Q08	.000	.000	.000	.000		.000	.000	.000	.000	.017
	Q06	.000	.000	.000	.000	.000		.000	.000	.000	.000
	Q07	.000	.000	.000	.000	.000	.000		.000	.000	.000
	Q08	.000	.000	.000	.000	.000	.000	.000		.000	.000
	Q09	.000	.000	.000	.000	.000	.000	.000	.000		.000
	Q10	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001
	Q11	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Q12	.000	.000	.000	.000	.000	.000	.000	.000	.000	.009
	Q13	.000	.000	.000	.000	.000	.000	.000	.000	.000	.004
	Q14	.000	.000	.000	.000	.000	.000	.000	.000	.000	.007
	Q15	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001
	Q16	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Q17	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Q18	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Q19	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Q20	.000	.000	.000	.000	.000	.000	.000	.000	.000	.039
	Q21	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Q22	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Q23	.410	.000	.000	.043	.017	.000	.039	.000	.000	

a. Determinant = 8.271E-04

Finding Factors: Communality

- Common Variance:
  - Variance that a variable shares with other variables
- Unique Variance:
  - Variance that is unique to a particular variable
- The proportion of common variance in a variable is called the communality
  - Communality = 1, All variance shared
  - Communality = 0, No variance shared
  - 0 < Communality < 1 = Some variance shared



Finding Factors

- We find factors by calculating the amount of common variance
  - Circularity
- Principal Components Analysis:
  - Assume all variance is shared
  - All communalities = 1
- Factor Analysis
  - Estimate communality
  - Use squared multiple correlation (SMC)
- Principal Components and Factor Analysis will identify similar factors when there are a large number of variables (i.e., more than 30) and the communalities are high (i.e., greater than .7)

Communalities

	Initial	Extraction
Q01	1.000	.435
Q02	1.000	.414
Q03	1.000	.530
Q04	1.000	.469
Q05	1.000	.343
Q06	1.000	.654
Q07	1.000	.545
Q08	1.000	.739
Q09	1.000	.484
Q10	1.000	.335
Q11	1.000	.690
Q12	1.000	.513
Q13	1.000	.536
Q14	1.000	.488
Q15	1.000	.378
Q16	1.000	.487
Q17	1.000	.683
Q18	1.000	.597
Q19	1.000	.343
Q20	1.000	.484
Q21	1.000	.550
Q22	1.000	.464
Q23	1.000	.412

Extraction Method: Principal Component

Factor Extraction

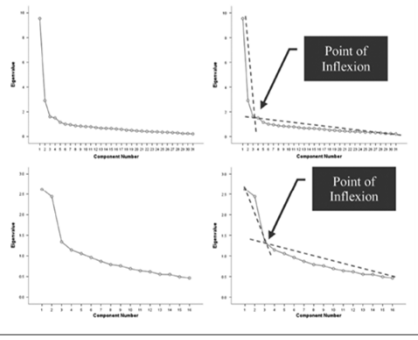
- Kaiser's Extraction
  - Kaiser (1960): retain factors with Eigenvalues > 1
- Scree Plot
  - Cattell (1966): use 'point of inflexion' of the scree plot
- Which Rule?
  - Use Kaiser's Extraction when
    - less than 30 variables, communalities after extraction > 0.7
    - sample size > 250 and mean communality ≥ 0.6
  - Scree plot is good if sample size is > 200

Total Variance Explained

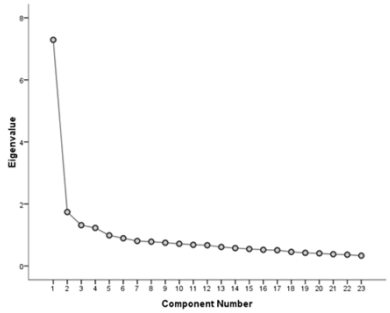
	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.290	31.696	31.696	7.290	31.696	31.696	3.730	16.219	16.219
2	1.739	7.560	39.256	1.739	7.560	39.256	3.340	14.523	30.742
3	1.317	5.725	44.981	1.317	5.725	44.981	2.553	11.099	41.842
4	1.227	5.336	50.317	1.227	5.336	50.317	1.949	8.475	50.317
5	.988	4.295	54.612						
6	.895	3.893	58.504						
7	.806	3.502	62.007						
8	.783	3.404	65.410						
9	.751	3.265	68.676						
10	.717	3.117	71.793						
11	.684	2.972	74.765						
12	.670	2.911	77.676						
13	.612	2.661	80.337						
14	.578	2.512	82.849						
15	.549	2.388	85.236						
16	.523	2.275	87.511						
17	.508	2.210	89.721						
18	.456	1.982	91.704						
19	.424	1.843	93.546						
20	.408	1.773	95.319						
21	.379	1.650	96.969						
22	.364	1.583	98.552						
23	.333	1.448	100.000						

Extraction Method: Principal Component Analysis.

Scree Plots



The Scree Plot For SAQ Data

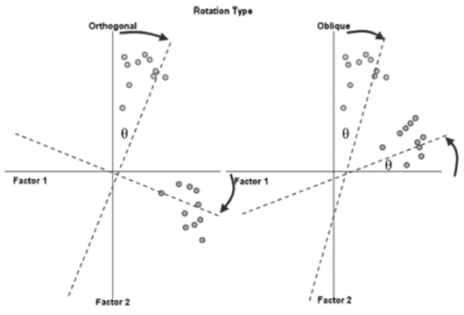


Rotation

- To aid interpretation it is possible to maximize the loading of a variable on one factor while minimizing its loading on all other factors
- This is known as Factor Rotation
- There are two types
  - Orthogonal (factors are uncorrelated)
  - Oblique (factors intercorrelate)

Orthogonal Rotation **(Uncorrelated)**

Oblique Rotation **(Correlated)**



Before Rotation

Component Matrix[a]

	Component			
	1	2	3	4
Q18	.701			
Q07	.685			
Q16	.679			
Q13	.673			
Q12	.669			
Q21	.658			
Q14	.656			
Q11	.652			-.400
Q17	.643			
Q04	.634			
Q03	-.629			
Q15	.593			
Q01	.586			
Q05	.556			
Q08	.549	.401		-.417
Q10	.437			
Q20	.436		-.404	
Q19	-.427			
Q09		.627		
Q02		.548		
Q22		.465		
Q06	.582		.571	
Q23				.507

Extraction Method: Principal Component Analysis.
a. 4 components extracted.

Orthogonal Rotation

Rotated Component Matrix[a]

	Component			
	1	2	3	4
I have little experience of computers	.800			
SPSS always crashes when I try to use it	.684			
I worry that I will cause irreparable damage because of my incompetence with computers	.647			
All computers hate me	.638			
Computers have minds of their own and deliberately go wrong whenever I use them	.579			
Computers are useful only for playing games	.560			
Computers are out to get me	.489			
I can't sleep for thoughts of eigen vectors		.677		
I wake up under my duvet thinking that I am trapped under a normal distribution		.661		
Standard deviations excite me		-.567		
People try to tell you that SPSS makes statistics easier to understand but it doesn't	.473	.523		
I dream that Pearson is attacking me with correlation coefficients		.516		
I weep openly at the mention of central tendency		.514		
Statistics makes me cry		.496		
I don't understand statistics		.429		
I have never been good at mathematics			.833	
I slip into a coma whenever I see an equation			.747	
I did badly at mathematics at school			.747	
My friends are better at statistics than me				.648
My friends are better at SPSS than I am				.645
If I'm good at statistics my friends will think I'm a nerd				.556
My friends will think I'm stupid for not being able to cope with SPSS				.543
Everybody looks at me when I use SPSS				.427

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 9 iterations.

Oblique Rotation

Pattern Matrix[a]

	Component			
	1	2	3	4
I can't sleep for thoughts of eigen vectors	.706			
I wake up under my duvet thinking that I am trapped under a normal distribution	.691			
Standard deviations excite me	-.611			
I dream that Pearson is attacking me with correlation coefficients	.405			
I weep openly at the mention of central tendency	.400			
Statistics makes me cry				
I don't understand statistics				
My friends are better at SPSS than I am		.643		
My friends are better at statistics than me		.621		
If I'm good at statistics my friends will think I'm a nerd		.615		
My friends will think I'm stupid for not being able to cope with SPSS		.507		
Everybody looks at me when I use SPSS				
I have little experience of computers			.888	
SPSS always crashes when I try to use it			.713	
All computers hate me			.683	
I worry that I will cause irreparable damage because of my incompetence with computers			.680	
Computers have minds of their own and deliberately go wrong whenever I use them			.588	
Computers are useful only for playing games			.585	
People try to tell you that SPSS makes statistics easier to understand but it doesn't	.412		.462	
Computers are out to get me			.411	
I have never been good at mathematics				-.902
I slip into a coma whenever I see an equation				-.774
I did badly at mathematics at school				-.774

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.
a. Rotation converged in 29 iterations.

What Do the Factors Represent?

- We assume that algebraic factors represent psychological constructs
  - Factor 1: Fear of statistics (e.g., "I can't sleep for thoughts of eigenvectors")
  - Factor 2: Fear of peer evaluation (e.g., "My friends are better at statistics than me")
  - Factor 3: Fear of computers(e.g., "All computers hate me")
  - Factor 4: Fear of mathematics (e.g., "I have never been good at mathematics")
- The nature of these psychological dimensions is 'guessed at' by looking at the loadings for a factor
- There is no way to "know" what the factor represents…rather we have to decide ourselves
  - The same set of items may be referred to as "social dominance" by one researcher but "aggression" by another

Reliability

- Test-Retest Method
  - Complete the same measure on two occasions and calculates the correlation
- Alternate Form Method
  - Complete two slightly different forms of the same measure and calculates the correlation
- Split-Half Method
  - Splits the questionnaire into two random halves and calculates the correlation
- Cronbach's Alpha
  - Basically splits the questionnaire into all possible halves, calculates the scores, correlates them, and averages the correlation for all splits
  - Ranges from 0 (no reliability) to 1 (complete reliability)

Interpreting Cronbach's Alpha

- Kline (1999)
  - Reliable if $\alpha > .7$
- Depends on the number of items
  - More questions = bigger α
- Treat Subscales separately
- Remember to reverse score reverse phrased items!
  - If not, α is reduced and can even be negative

Reliability for Fear of Statistics Subscale

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Statiscs makes me cry	21.76	21.442	.536	.343	.802
Standard deviations excite me	20.72	19.825	.549	.309	.800
I dream that Pearson is attacking me with correlation coefficients	21.35	20.410	.575	.355	.796
I don't understand statistics	21.41	20.942	.494	.272	.807
People try to tell you that SPSS makes statistics easier to understand but it doesn't	20.97	20.639	.572	.337	.796
I weep openly at the mention of central tendency	21.25	20.451	.597	.389	.793
I can't sleep for thoughts of eigen vectors	20.51	21.176	.419	.244	.818
I wake up under my duvet thinking that I am trapped under a normal distribtion	20.96	19.939	.606	.399	.791

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.821	.823	8

Reliability for the Fear of Peer Evaluation Subscale

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
My friends will think I'm stupid for not being able to cope with SPSS	11.46	8.119	.339	.134	.515
My friends are better at statistics than me	10.24	6.395	.391	.167	.476
Everybody looks at me when I use SPSS	10.79	7.381	.316	.106	.522
My friends are better at SPSS than I am	10.20	7.282	.378	.144	.487
If I'm good at statistics my friends will think I'm a nerd	9.65	7.988	.239	.069	.563

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.570	.572	5

Reliability for Fear of Computers Subscale

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
I have little experience of computers	15.87	17.614	.619	.398	.791
All computers hate me	15.17	17.737	.619	.395	.790
Computers are useful only for playing games	15.81	20.736	.400	.167	.824
I worry that I will cause irreparable damage because of my incompetence with computers	15.64	18.809	.607	.384	.794
Computers have minds of their own and deliberately go wrong whenever I use them	15.22	18.719	.577	.350	.798
Computers are out to get me	15.33	19.322	.491	.250	.812
SPSS always crashes when I try to use it	15.52	17.832	.647	.447	.786

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.823	.821	7

Reliability for Fear of Mathematics Subscale

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
I have never been good at mathematics	4.72	2.470	.684	.470	.740
I did badly at mathematics at school	4.70	2.453	.682	.467	.742
I slip into a coma whenever I see an equation	4.49	2.504	.652	.425	.772

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.819	.819	3