actor Analysis – Beginners Tutoria

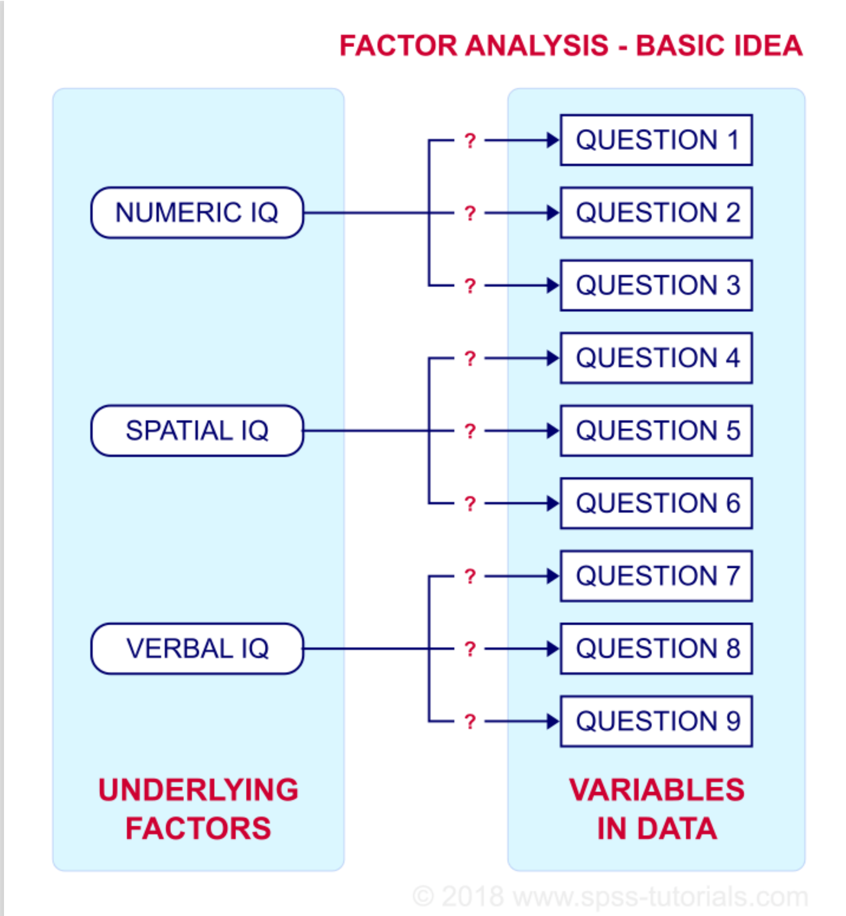
You are here: Home → Basics → SPSS - Popular Tutorials → SPSS Factor Analysis – Beginners Tutorial

What is Factor Analysis?

Factor analysis is a statistical technique for identifying which underlying factors are measured by a (much larger) number of observed variables.

Such "underlying factors" are often variables that are difficult to measure such as IQ, depression or extraversion. For measuring these, we often try to write multiple questions that -at least partially- reflect such factors. The basic idea is illustrated below.

Ne
Sn



Now, if questions 1, 2 and 3 all measure numeric IQ, then the Pearson correlations among these items should be substantial: respondents with high numeric IQ will typically score high on all 3 questions and reversely. The same reasoning goes for questions 4, 5 and 6: if they really measure “the same thing” they'll probably correlate highly.
However, questions 1 and 4 -measuring possibly unrelated traits- will not necessarily correlate. So if my factor model is correct, I could expect the correlations to follow a pattern as shown below.

Correlation Matrix

	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9
Question 1	**1.00**	**.75**	**.76**	.04	.11	.10	.04	-.06	.01
Question 2	**.75**	**1.00**	**.78**	-.01	.00	.02	.00	-.06	.02
Question 3	**.76**	**.78**	**1.00**	-.06	-.03	-.02	.08	-.05	.02
Question 4	.04	-.01	-.06	**1.00**	**.85**	**.82**	.10	.00	.05
Question 5	.11	.00	-.03	**.85**	**1.00**	**.86**	-.06	-.08	-.04
Question 6	.10	.02	-.02	**.82**	**.86**	**1.00**	.04	.04	.02
Question 7	.04	.00	.08	.10	-.06	.04	**1.00**	**.71**	**.78**
Question 8	-.06	-.06	-.05	.00	-.08	.04	**.71**	**1.00**	**.78**
Question 9	.01	.02	.02	.05	-.04	.02	**.78**	**.78**	**1.00**

Confirmatory Factor Analysis

Right, so after measuring questions 1 through 9 on a simple random sample of respondents, I computed this correlation matrix. Now I could ask my software if these correlations are likely, given my theoretical factor model. In this case, I'm trying to **confirm** a model by fitting it to my data. This is known as “**confirmatory factor analysis**”.
SPSS does not include confirmatory factor analysis but those who are interested could take a look at AMOS.

Exploratory Factor Analysis

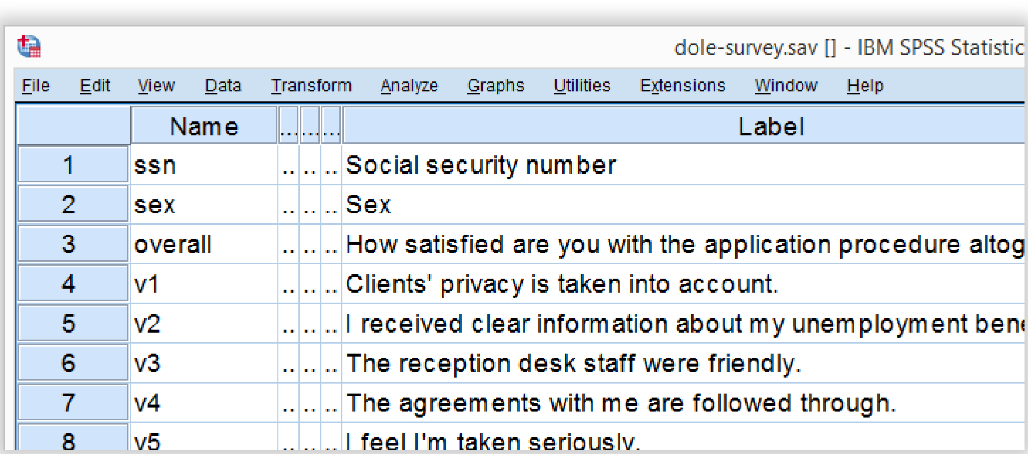
But what if I don't have a clue which -or even how many- factors are represented by my data? Well, in this case, I'll ask my software to suggest some model given my correlation matrix. That is, I'll **explore** the data. Hence, “exploratory factor analysis”. The simplest possible explanation of how it works is that

the software tries to find groups of variables that are highly intercorrelated.

Each such group probably represents an underlying common factor. There's different mathematical approaches to accomplishing this but the most common one is **principal components analysis** or PCA. We'll walk you through with an example.

Research Questions and Data

A survey was held among 388 applicants for unemployment benefits. The data thus collected are in dole-survey.sav, part of which is shown below.



	Name	...	Label
1	ssn	.. .. ..	Social security number
2	sex	.. .. ..	Sex
3	overall	.. .. ..	How satisfied are you with the application procedure altog
4	v1	.. .. ..	Clients' privacy is taken into account.
5	v2	.. .. ..	I received clear information about my unemployment bene
6	v3	.. .. ..	The reception desk staff were friendly.
7	v4	.. .. ..	The agreements with me are followed through.
8	v5	.. .. ..	I feel I'm taken seriously.

The survey included 16 questions on client satisfaction. We think these measure a smaller number of underlying satisfaction factors but we've no clue about a model. So our **research questions** for this analysis are:

- how many factors are measured by our 16 questions?
- which questions measure similar factors?
- which satisfaction aspects are represented by which factors?

Quick Data Check

Now let's first make sure we have an idea of what our data basically look like. We'll inspect the frequency distributions with corresponding bar charts for our 16 variables by running the syntax below.

```
*Show variable names, values and labels in output tab

set
tnumbers both /* show values and value labels in outp
tvars both /* show variable names but not labels in c
ovars names. /* show variable names but not labels in

*Basic frequency tables with bar charts.

frequencies v1 to v20
/barchart.
```

Result

Ne
Sn

v1 Clients' privacy is taken into account.

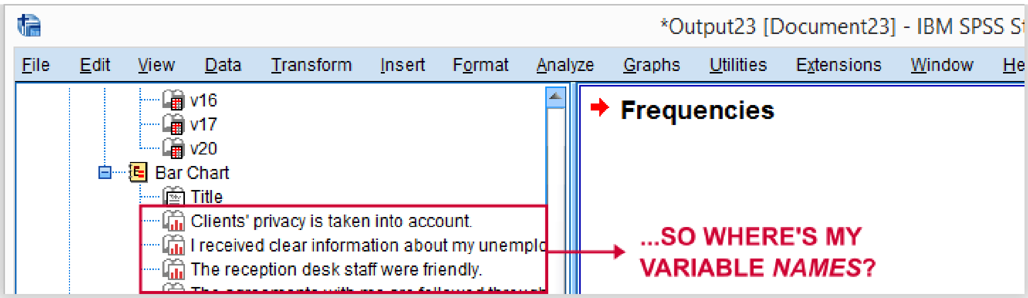
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Fully disagree	3	.8	.8	.8
	2	15	3.9	4.0	4.8
	3	42	10.8	11.2	16.0
	4	80	20.6	21.3	37.3
	5	81	20.9	21.6	58.9
	6	79	20.4	21.1	80.0
	7 Fully agree	57	14.7	15.2	95.2
	8 No answer	18	4.6	4.8	100.0
	Total	375	96.6	100.0	
Missing	System	13	3.4		
Total		388	100.0		

This very minimal data check gives us quite some **important insights** into our data:

- All frequency distributions look **plausible**. We don't see anything weird in our data.
- All variables are 1 **positively coded**: higher values always indicate more positive sentiments.
- All variables have 2 a value 8 (“**No answer**”) which we need to set as a user missing value.
- All variables have some 3 **system missing values** too but the extent of missingness isn't too bad.

A somewhat annoying flaw here is that we don't see variable *names* for our bar charts in the output outline.

Ne
Sn



If we see something unusual in a chart, we don't easily see which variable to address. But in this example -fortunately- our charts all look fine.

So let's now set our missing values and run some quick descriptive statistics with the syntax below.

```
*Set 8 ('No answer') as user missing value for all va

missing values v1 to v20 (8).

*Inspect valid N for each variable.

descriptives v1 to v20.
```

Result

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
v1 Clients' privacy is taken into account.	357	1	7	4.92	1.438
v2 I received clear information about my unemployment benefit.	361	1	7	3.82	1.477
v20 The letters I receive have an appropriate voice.					1.441
Valid N (listwise)	149				

ONLY 149 CASES HAVE NO MISSING VALUES ON ANY VARIABLES IN THIS TABLE.

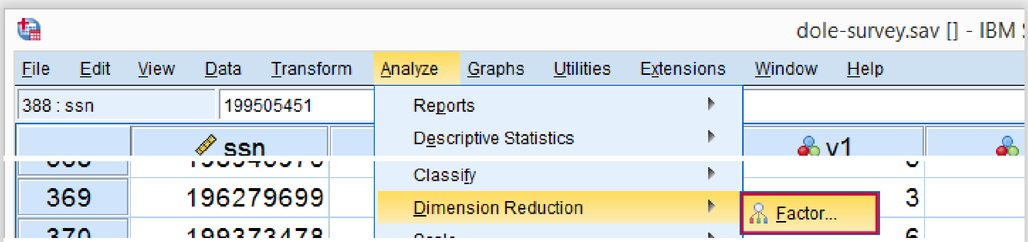
Note that none of our variables have many -more than some 10%- missing values. However,

only 149 of our 388 respondents have zero missing values

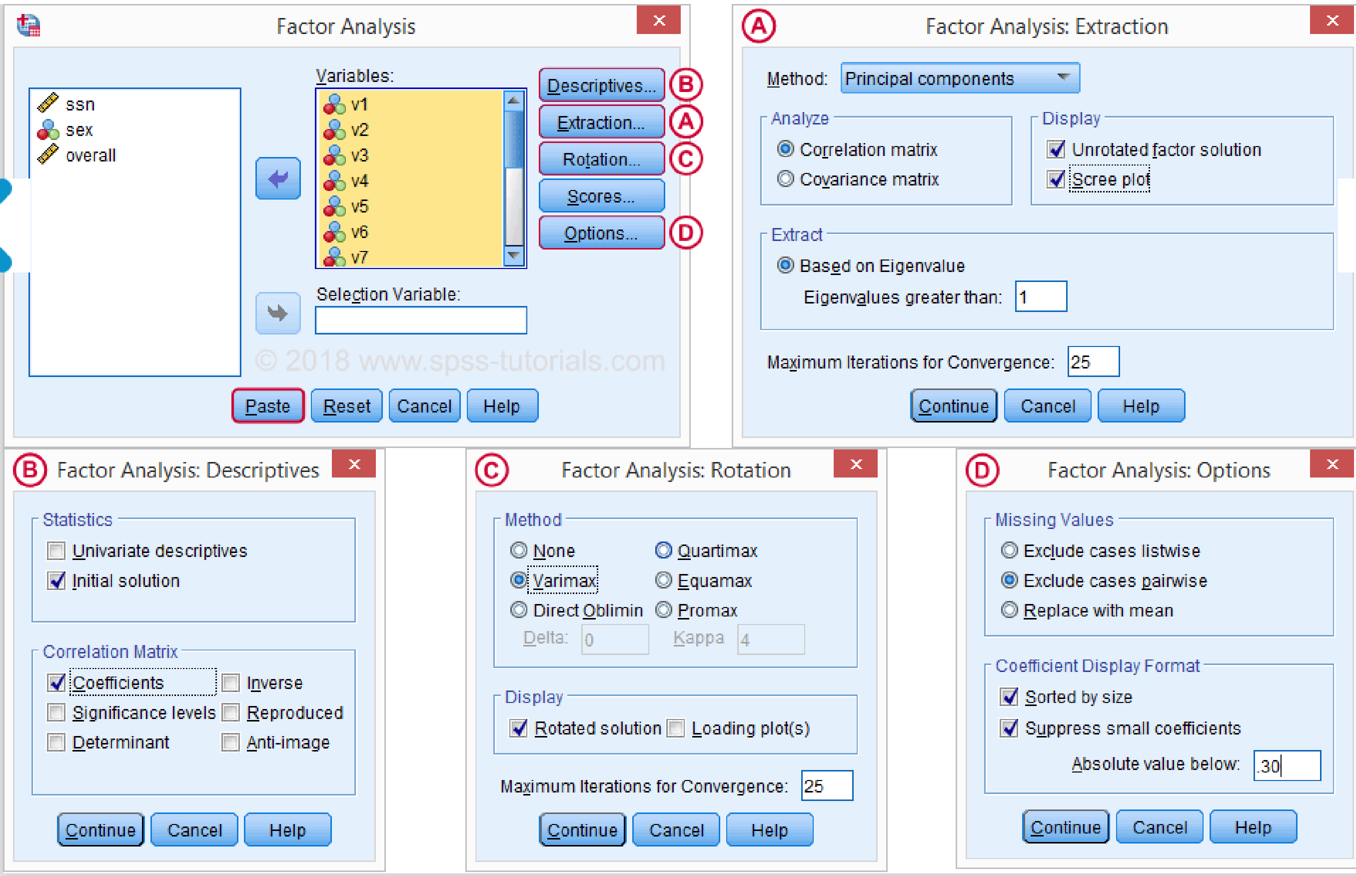
on the entire set of variables. This is very important to be aware of as we'll see in a minute.

Running Factor Analysis in SPSS

Let's now navigate to Analyze ▶ Dimension Reduction ▶ Factor as shown below.



In the dialog that opens, we have a ton of options. For a “standard analysis”, we'll select the ones shown below. If you don't want to go through all dialogs, you can also replicate our analysis from the syntax below.



D Avoid "Exclude cases listwise" here as it'll only include our 149 "complete" respondents in our factor analysis. Clicking Paste results in the syntax below.

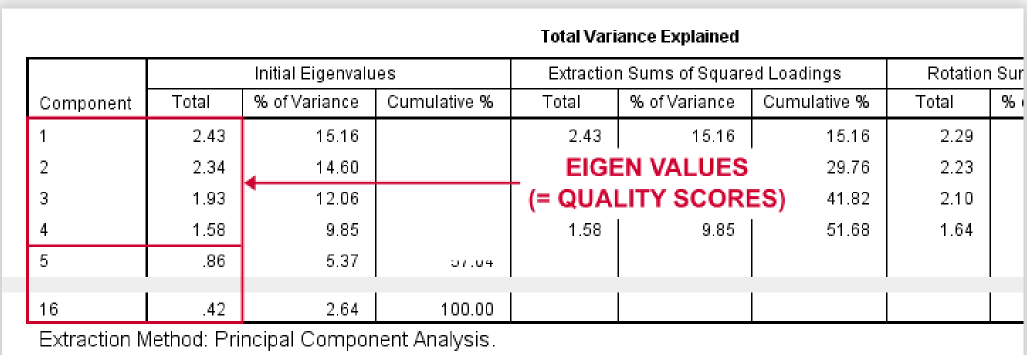
SPSS Factor Analysis Syntax

```
*Show both variable names and labels in output.

set tvars both.

*Initial factor analysis as pasted from menu.
```

```
FACTOR
/VARIABLES v1 v2 v3 v4 v5 v6 v7 v8 v9 v11 v12 v13 v14
/MISSING PAIRWISE /*IMPORTANT!*/
/PRINT INITIAL CORRELATION EXTRACTION ROTATION
/FORMAT SORT BLANK(.30)
/PLOT EIGEN
/CRITERIA MINEIGEN(1) ITERATE(25)
/EXTRACTION PC
/CRITERIA ITERATE(25)
/ROTATION VARIMAX
/METHOD=CORRELATION.
```

Factor Analysis Output I - Total Variance Explained

Right. Now, with 16 input variables, PCA initially extracts 16 factors (or "components"). Each component has a **quality score** called an **Eigenvalue**. Only components with high Eigenvalues are likely to represent a real underlying factor.



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sur	
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	%
1	2.43	15.16		2.43	15.16	15.16	2.29	
2	2.34	14.60				29.76	2.23	
3	1.93	12.06				41.82	2.10	
4	1.58	9.85		1.58	9.85	51.68	1.64	
5	.86	5.37	57.04					
16	.42	2.64	100.00					

Extraction Method: Principal Component Analysis.

So what's a high Eigenvalue? A common rule of thumb is to

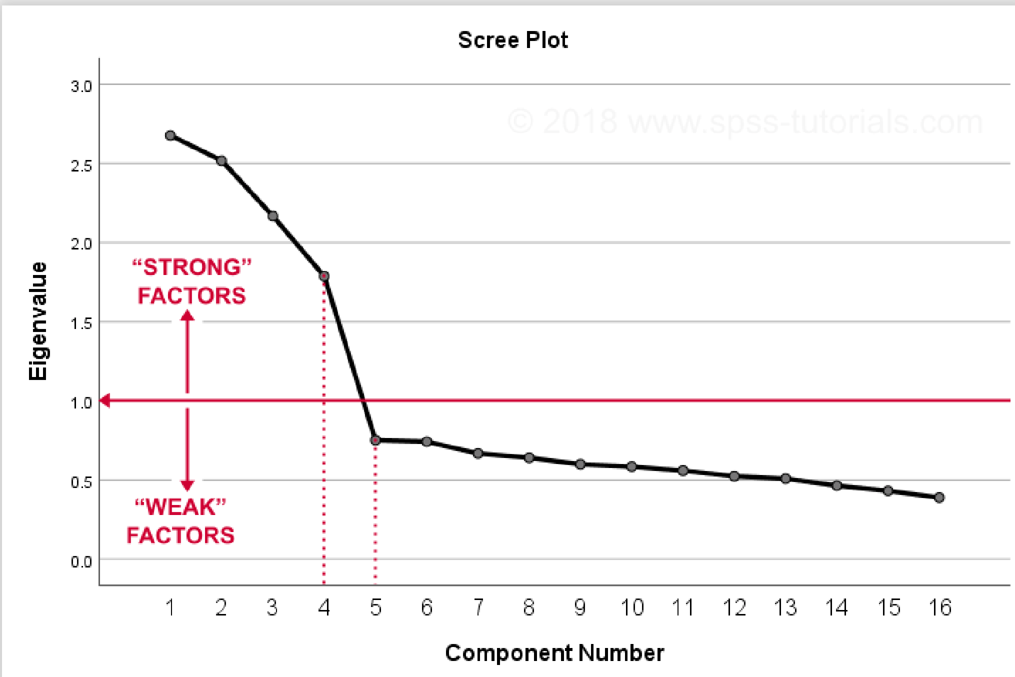
select components whose Eigenvalue is at least 1.

Applying this simple rule to the previous table answers our first research question:

our 16 variables seem to measure 4 underlying factors.

This is because only our first 4 components have an Eigenvalue of at least 1. The other components -having low quality scores- are not assumed to represent real traits underlying our 16 questions. Such components are considered “scree” as shown by the line chart below.

Factor Analysis Output II - Scree Plot



A scree plot visualizes the Eigenvalues (quality scores) we just saw. Again, we see that the first 4 components have Eigenvalues over 1. We consider these “strong factors”. After that -component 5 and onwards- the Eigenvalues **drop off dramatically**. The sharp drop between components 1-4 and components 5-16 strongly suggests that 4 factors underlie our questions.

Ne
Sn

Factor Analysis Output III - Communalities

So to what extent do our 4 underlying factors account for the variance of our 16 input variables? This is answered by the r square values which -for some really dumb reason- are called **communalities** in factor analysis.

Communalities

	Initial	Extraction
v1 Clients' privacy is taken into account.	1.000	.596
v2 I received clear information about my unemployment benefit.	1.000	.499
v3 The reception desk staff were friendly.	1.000	.618
v4 The agreements with me are followed through.	1.000	.648
v5 I feel I'm taken seriously.	1.000	.614
v6 My contact person succeeds in motivating me.	1.000	.532
v7 My contact	1.000	.623
v8 My contact	1.000	.497
v9 It's clear to	1.000	.505
v11 My contact person points out fitting job opportunities.	1.000	.669
v12 I have clear agreements about the remaining procedures.	1.000	.565
v13 It's easy to find information regarding my unemployment benefit.	1.000	.539
v14 My contact person always does what she/he promises.	1.000	.624
v16 I've been told clearly how my application process will continue.	1.000	.536
v17 I know who can answer my questions on my unemployment benefit	1.000	.508
v20 The letters I receive have an appropriate tone of voice.	1.000	.574

Extraction Method: Principal Component Analysis.

COMMUNALITIES: PROPORTIONS OF VARIANCE ACCOUNTED FOR BY SELECTED COMPONENTS

Right. So if we predict v1 from our 4 components by multiple regression, we'll find r square = 0.596 -which is v1' s communality. Variables having **low communalities** -say lower than 0.40- don't contribute much to measuring the underlying factors.

You could consider **removing** such variables from the analysis. But keep in mind that doing so changes *all* results. So you'll need to rerun the entire analysis with one variable omitted. And then perhaps rerun it again with another variable left out.

If the scree plot justifies it, you could also consider selecting an **additional component**. But don't do this if it renders the (rotated) factor loading matrix less interpretable.

Ne
Sn

Factor Analysis Output IV - Component Matrix

Thus far, we concluded that our 16 variables probably measure 4 underlying factors. But

which items measure which factors?

The component matrix shows the Pearson correlations between the items and the components. For some dumb reason, these correlations are called **factor loadings**.

Component Matrix[a]

	Component			
	1	2	3	4
v17 I know who can answer my questions on my unemployment benefit	.641			
v16 I've been told clearly how my application process will continue.	.640			
v13 It's easy to find information regarding my unemployment benefit.	.607			
v2 I received clear information about my unemployment benefit.	.564			
v9 It's clear to me what my rights are.	.540		.343	
v5 I feel I'm taken seriously.		.692		
v1 Clients' privacy is taken into account.		.601	-.396	
v3 The reception desk staff were friendly.	.302	.590	-.330	
v20 The letters I receive have an appropriate tone of voice.		.580	-.322	
v6 My contact person succeeds in motivating me.		.455	.433	
v7 My contact person takes her/his time with me.		.385	.558	
v11 My contact person points out fitting job opportunities.	-.321	.458	.554	
v8 My contact person carefully prepares her/his interviews with me.		.356	.539	
v4 The agreements with me are followed through.				.741
v14 My contact person always does what she/he promises.				.720
v12 I have clear agreements about the remaining procedures.				.577

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Ideally, we want each input variable to measure precisely one factor. Unfortunately, that's not the case here. For instance, v9 measures (correlates with) components 1 and 3. Worse even, v3 and v11 even measure components 1, 2 and 3 simultaneously. If a variable has more than 1 substantial factor loading, we call those **cross loadings**. And we don't like those. They complicate the interpretation of our factors.
The solution for this is **rotation**: we'll **redistribute the factor loadings** over the factors according to some mathematical rules that we'll leave to SPSS. This redefines what our factors represent. But that's ok. We hadn't looked into that yet anyway.
Now, there's different rotation methods but the most common one is the **varimax rotation**, short for "**vari**able **max**imization. It tries to redistribute the factor loadings such that each variable measures precisely one factor -which is the ideal scenario for understanding our factors. And as we're about to see, our varimax rotation works perfectly for our data.

Factor Analysis Output V - Rotated Component Matrix

Our rotated component matrix (below) answers our second research question: “**which variables measure which factors?**”

Rotated Component Matrix[a]

	Component			
	1	2	3	4
v16 I've been told clearly how my application process will continue.	.728			
v13 It's easy to find information regarding my unemployment benefit.	.726			
v17 I know who can answer my questions on my unemployment benefit	.708			
v2 I received clear information about my unemployment benefit.	.703			
v9 It's clear to me what my rights are.	.700			
v3 The reception desk staff were friendly.		.784		
v1 Clients' privacy is taken into account.		.769		
v5 I feel I'm taken seriously.		.763		
v20 The letters I receive have an appropriate tone of voice.		.757		
v11 My contact person points out fitting job opportunities.			.808	
v7 My contact person takes her/his time with me.			.780	
v6 My contact person succeeds in motivating me.			.724	
v8 My contact person carefully prepares her/his interviews with me.			.705	
v4 The agreements with me are followed through.				.803
v14 My contact person always does what she/he promises.				.787
v12 I have clear agreements about the remaining procedures.				.733

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 4 iterations.

Our last research question is: “**what do our factors represent?**” Technically, a factor (or component) represents whatever its variables have in common. Our rotated component matrix (above) shows that our first component is measured by

- v17 - I know who can answer my questions on my unemployment benefit.
- v16 - I've been told clearly how my application process will continue.
- v13 - It's easy to find information regarding my unemployment benefit.

- v2 - I received clear information about my unemployment benefit.
- v9 - It's clear to me what my rights are.

Note that these variables all relate to the respondent receiving clear information. Therefore, we interpret component 1 as “clarity of information”. This is the **underlying trait** measured by v17, v16, v13, v2 and v9.
After interpreting all components in a similar fashion, we arrived at the following descriptions:

- Component 1 - “**Clarity of information**”
- Component 2 - “**Decency and appropriateness**”
- Component 3 - “**Helpfulness contact person**”
- Component 4 - “**Reliability of agreements**”

We'll set these as variable labels after actually adding the factor scores to our data.

Adding Factor Scores to Our Data

It's pretty common to add the actual factor scores to your data. They are often used as predictors in regression analysis or drivers in cluster analysis. SPSS FACTOR can add factor scores to your data but this is often a bad idea for 2 reasons:

- Factor scores will only be added for cases without missing values on any of the input variables. We saw that this holds for only 149 of our 388 cases.
- Factor scores are z-scores: their mean is 0 and their standard deviation is 1. This complicates their interpretation.

Ne
Sn

In many cases, a better idea is to **compute factor scores as means** over variables measuring similar factors. Such means tend to correlate almost perfectly with "real" factor scores but they don't suffer from the aforementioned problems. Importantly, we should do so only if all input variables have **identical measurement scales**. Since this holds for our example, we'll add factor scores with the syntax below.

Computing and Labeling Factor Scores Syntax

```
*Create factors as means over variables per factor.

compute fac_1 = mean(v16,v13,v17,v2,v9).
compute fac_2 = mean(v3,v1,v5,v20).
compute fac_3 = mean(v11,v7,v6,v8).
compute fac_4 = mean(v4,v14,v12).

*Label factors.

variable labels
fac_1 'Clarity of information'
fac_2 'Decency and appropriateness'
fac_3 'Helpfulness contact person'
fac_4 'Reliability of agreements'.

*Quick check.

descriptives fac_1 to fac_4.
```

Result

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Clarity of information	388	1.25	6.25	3.91	1.09
Decency and appropriateness	388	2.00	7.00	4.98	1.11
Helpfulness contact person	388	2.00	7.00	4.43	1.13
Reliability of agreements	388	1.00	6.50	3.95	1.21
Valid N (listwise)	388				

Ne
Sn

This descriptives table shows how we interpreted our factors. Because we computed them as means, they have the same 1 - 7 scales as our input variables. This allows us to conclude that

- “Decency and appropriateness” is rated **best** (roughly 5.0 out of 7 points) and
- “Clarity of information” is rated **worst** (roughly 3.9 out of 7 points).

Thanks for reading.

Let me know what you think!

Your name*

Your email address*

Your website

Your comment*

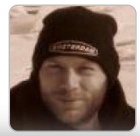
Done!

Required field. Your comment will show up after approval from a moderator.

Ne
Sn

This tutorial has 36 comments

By Ruben Geert van den Berg on November 17th, 2019



Good question!

First off, neither correlations nor fa

normally distributed variables

Expand comment | all comments

1 … 8

Ne
Sn

Get In Touch!

Ruben Geert van den Berg

LinkedIn

Facebook

SPSS Help (Netherlands)

Sigma Plus Statistiek

www.sigma-plus-statistiek.nl

info@sigma-plus-statistiek.nl

SPSS Help (International)

SPSS tutorials

www.spss-tutorials.com

info@spss-tutorials.com

Ne
Sn