# Instructions

Throughout the exercise, maintain an annotated syntax file that will enable reproduction of your analyses.

1. Read through "bfi-data-description.rtf" and learn about the data file. Examine "bfi-personality-meta-data.xlsx" to get a sense of how the variables were coded
2. (a) remove problematic cases; (b) deal with any missing data or outliers
3. Split the remaining cleaned data file into two datasets (one for exploration and one for confirmatory analyses)

Complete the following analyses on the exploratory dataset. The aim of the following analyses is to use exploratory factor analysis (EFA) and related tools to examine the structure of the test.

1. Examine means and frequencies of each response option to see whether any items lack variability (i.e., means that are very high or very low)
2. Perform an EFA to look at possible numbers of factors, and the degree to which items load appropriately on specified factors: (a) Should any items be dropped? (b) How many factors should be retained? (c) describe the structure of the test as suggested by EFA noting interesting features such as (i) relevant cross-loadings, (ii) what would be captured if fewer or more factors than finally specified were retained; (iii) whether some items load substantially more or less; etc.

Then move on to describing the test

1. Report means and standard deviations for each item of the test along with item text, and so on
2. Provide a description of how to score the test (e.g., which items belong to which scale) consistent with your preferred model of the data

**Optional exercises**

1. Score the tests
2. Report means, standard deviations, cronbach's alpha and intercorrelations for each scale
3. For any available demographic data, report the relationship with the scale scores

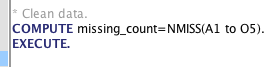
# Answers

## 2. (a) remove problematic cases;

This first requires some consideration of what are problematic cases. From my experience problematic cases for questionnaires include (a) participants who skip a lot of items; (b) participants who don't take the questionnaire seriously by answering in a random or grossly dishonest way; (c) participants with data entry errors. So let's deal with each of these possible errors in turn:

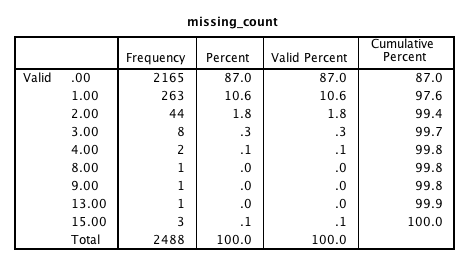
(a) participants who skip a lot of items:

Using "transform - compute" there is an "nmiss" function, which we can give all the items. This will indicate how many missing responses that each participant has. Don't forget to paste this syntax for reproducibility purposes.



* "compute" is a keyword indicating that a new variable is being created
* "missing\_count" is the name I chose for the new variable
* "nmiss" is the name of the function
* "A1 to O5" is a shorthand way of typing "A1, A2, A3, A4" and so on to "A5". It essentially tells SPSS to use all variables between A1 and O5 in the data frame
* "execute." is required for many commands in SPSS that are designed to create new variables.

We can then inspect the frequencies of missing data counts (using analyze - descriptives - frequencies)



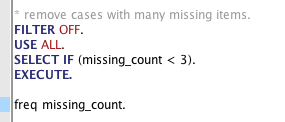
Given that there are 25 questions, there are only a few cases who have a large number of missing responses (e.g., perhaps the participants with 8 or more or some such). While it would be possible to retain such cases, I'm inclined to remove cases in an online setting who have skipped many items on the grounds that they probably did not take the exercise seriously. They also create a nuisance when it comes to modelling item-level data.

The simplest option would be just to remove any cases with any missing items. The sample size is large enough that it wont matter much from the perspective of statistical precision.

Ultimately, it comes down to a judgement call. It would not be difficult to impute some values for the cases with a few missing items. But if you had strong prior belief that any skipping was indicative of not taking the test seriously, then you might be more inclined to drop cases.

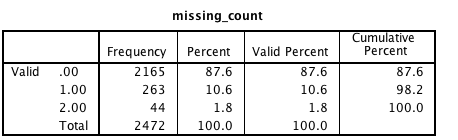
I'm going to drop cases with three or more missing items.

This can be achieved by using "data - select cases" then "if condition is satisfied" and then "delete unselected cases".



We'll then save this as a new data file. Like "bfi-unique-2.sav".

Running frequencies on this new data file:



(b) participants who don't take the study seriously

I often rely on an examination of item or study completion time to detect this. But in this case all we have if the set of responses on the items. We've done an initial removal of cases based on excessive missing data. The other option is to look at the pattern of responses to see whether it indicates some randomness in responding.

If we think about some of the possible response strategies that might indicate randomness, it might involve giving the same response to every question. In particular, where there are positive and negative worded items, then lack of response flipping can indicate randomness. It's also worth thinking in any real context about how likely such random data might be. I find in Internet research such random responding can occur more frequently. In particular, it's a problem where participants have an incentive to keep going (i.e., not just exit the survey half way through).

There are many ways to look at this, but a simple approach is to get Mahalanobis distance for each case based on the items, and then further inspect the cases with extreme values to see whether they are legitimate cases.

Go "Analyze - regression - linear" ;

DV ="id"

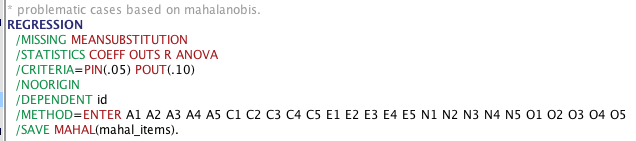
IV = all 25 personality test items;

Options: missing data - replace with mean

Save: mahalanobis

This will save a new variable to the data file. If you paste the syntax, you can control the name of the variable by following the MAHAL keyword with the desired name in parentheses (see <http://stackoverflow.com/q/18567939/180892> ). This can be useful if you want to make sure that your syntax is robust.

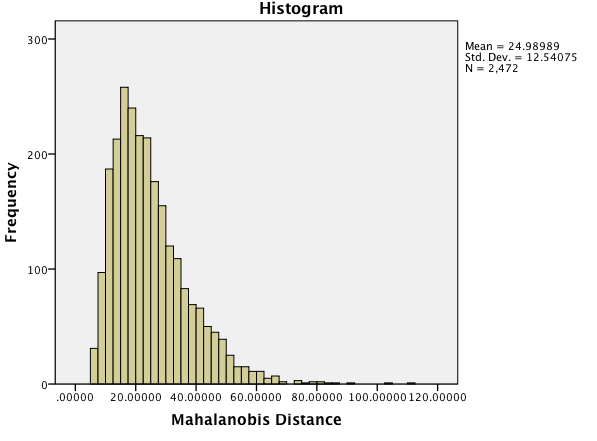
So the syntax might look like this where I have changed "MAHAL" to "MAHAL(mahal\_items)".



You can generally get context sensitive help about syntax by pressing F1 when you have the command active (e.g., like "REGRESSION").

In order to identify problematic cases, a histogram of Mahalanobis values can be useful. In particular, often problematic cases can appear as outliers or a second hump in the distribution.

I.e., go to "analyze - descriptives - frequencies" add "mahal\_items" and specify histogram:



This is a common shape for Mahalanobis Distance where there is a little bit of a skew. But there also appears to be a bit of a break in the distribution perhaps around 70.

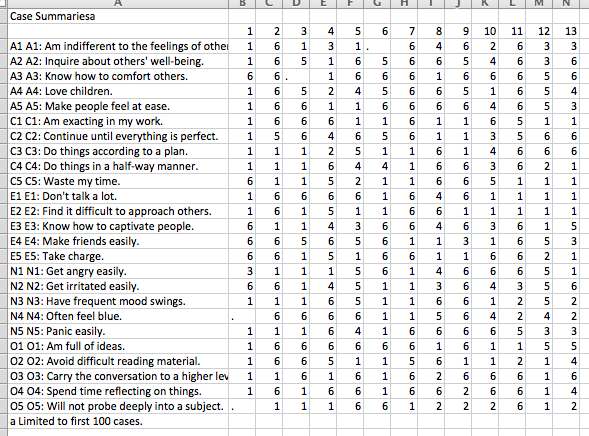
It can then be helpful to examine a few of these cases in terms of key properties.

You can do this by sorting column and looking at data view.

Another way is to select cases and then run summary cases.

* Use "data - select cases" select if "mahal\_items > 70"
* "Analyze - report - case summaries" to show particular variables (e.g., the 25 items)

After transposing rows and columsn and doing a tidy up in Excel, we have the following where each column is one of the high Mahal cases.



The judgement that needs to be made about these cases is whether these people reflect real but unusual respondents or whether they are indicative of random responding.

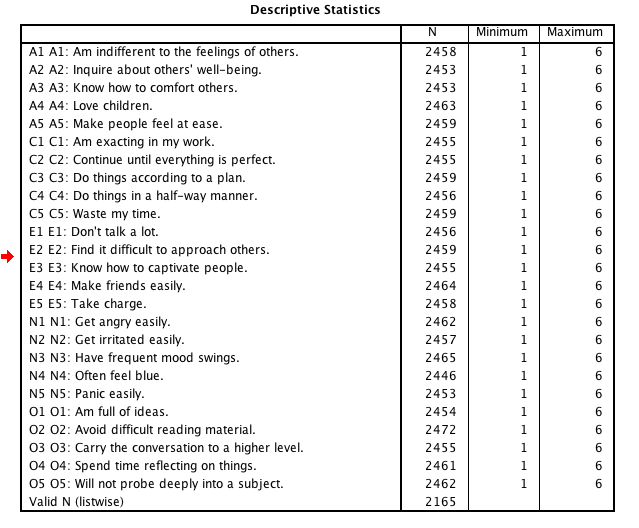
For example, take the most extreme mahal case shown as column 1. For the most part this person is just extremely low A, C, and O. However, there are also several possible contradictions such as "make friends easily = 6" but "find it difficult to approach others = 1" (that said I suppose you could make friends by letting them approach you). The personal also gets irritated easily (6) but does not have frequent mood swings (1). At the very least the person has almost entirely relied on using the extremes of the scale. A1 and A2 are also fairly inconsistent.

Given this analysis of the most extreme case, I am inclined to just leave the cases as they are. They are a small percentage of the whole. While some of these cases do contain possibly contradictory responses, presumably this might be typical of the way that some respondents think. Likewise, if a person occasionally misread a question or had the occasional rushed response, this is presumably not fatal for getting a reasonable response. Also, there are limits to the extent to which cases can actually be outliers when each item is inherently constrained by the 1 to 6 response scale.

(c) data entry errors

The data description file indicates that responses were on a 1 to 6 scale. So we should check that there are no values for items other than 1 to 6.

A basic check of this involves "analyze - descriptive - descriptives" (put in all items, and request min and max).

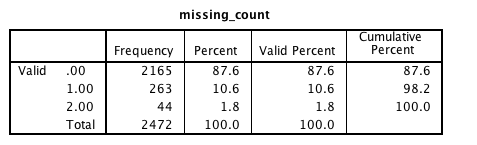


So that looks all good.

## (b) deal with any missing data or outliers

So we dealt with the major issues of missing data in the previous step.

However, there still quite a few cases with at least some missing data:



It would various item-level analyses easier if we could get some reasonable estimates of responses for these occasional missing items. Given that the missing data is pretty minimal, the exact strategy should not make much of a difference to the results, although if we were really keen we could check the robustness of any solution.

I'll do an EM imputation to replace the missing data for items.

Analyze - missing value analysis

case labels: id

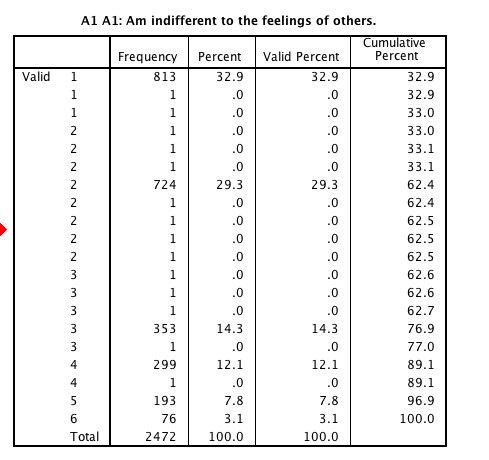
click use all variables

Estimation: click EM checkbox

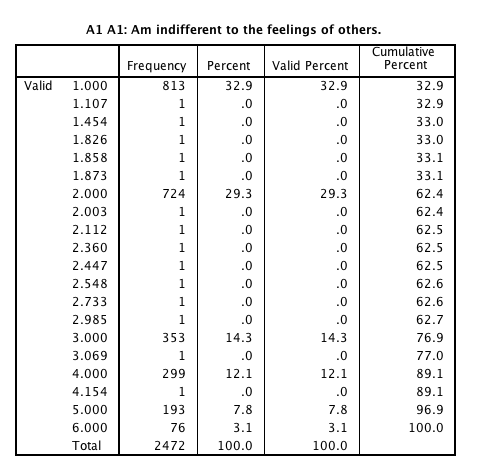
Click EM and choose a file location for the impute file

variables: A1 to O5 into both predicted variables and predictor variables (that means that the other variables like gender will be passed through into the new dataset but not used in the estimation)

Open the newly created imputed file. If you run frequencies, you'll see that the imputed values are a little different:



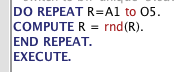
On closer inspection if you change the variable view to width = 4, decimal=3, you'll see that actually the imputed values are a decimal approximations:



This makes sense given that SPSS does not realise that the variables are discrete.

The data can be left as is, but alternatively, you could consider rounding the values to the nearest integer, just to make the imputed data look like actual data.

The following syntax will do that:

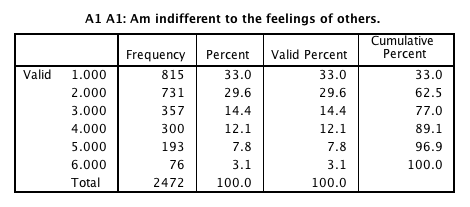


Several points can be made about this:

* "DO REPEAT" --- "END REPEAT" is a useful syntax command structure when you need to do a transformation across a wide range of variables. For example, you can use it to reverse a set of items or score a multiple choice test. In this case we are going to round all 25 personality test items.
* The first line of the command indicates that "R" will act as a variable name for the 25 variables. A1 to 05 is a short hand way of writing all the variable names.
* "COMPUTE R = rnd(R)." On each loop, R is replaced with one of the variables. Thus on the first loop this is the same as "compute A1 = rnd(A1)." "rnd" is a function that rounds numbers to the nearest integer. If you didn't know this, you could use the "transform - compute" menu to find out about the different functions available.

Admittedly this is fairly dangerous code and it is typically better to copy transformed variables into new variables so that you can check the old against the new version, but this is a little quicker.

A frequency check on A1 shows that it looks as it should

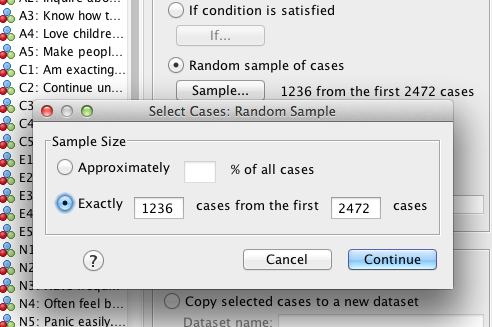


So now we have successfully imputed and rounded missing data. Once again, there are many choices and ways of doing this, some of which are quite sophisticated, but the above approach will suffice for our purposes.

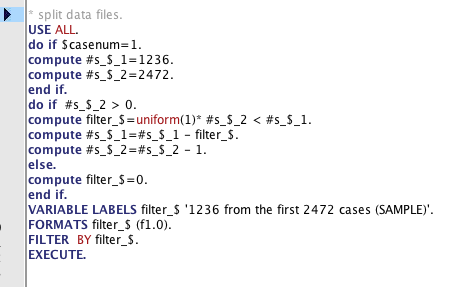
## 3. Split the remaining cleaned data file into two datasets (one for exploration and one for confirmatory analyses)

A good strategy when you have a big sample size and you are evaluating the psychometric structure of a test is to divide the dataset into an exploratory and a confirmatory dataset. There are various ways to do this in SPSS. A basic approach is to generate a uniform random variable, sort on that variable, and then assign the first half of cases to one dataset, and the other half to the other dataset.

A simpler way of performing those steps, is to use "data - select cases" then "random sample of cases", and then specify half goes to one sample.



If you press paste, you should get syntax like this:



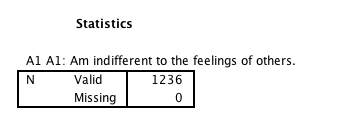
It's all pretty elaborate, but basically, you should end up with your dataset filtered by a variable called "filter\_$".

After that you can rename this variable to something like "dataset" and change the value labels to something like "exploratory" and "confirmatory".

It's then easiest create two copies of the dataset and label them exploratory and confirmatory and delete the data from these data files that does not correspond (i.e., delete the confirmatory data from the exploratory dataset and vice versa).

You can save copies of the data using "save as", and then use select cases with the "delete unselected cases" option to delete the unwanted cases.

Here's the new sample size for the exploratory dataset:

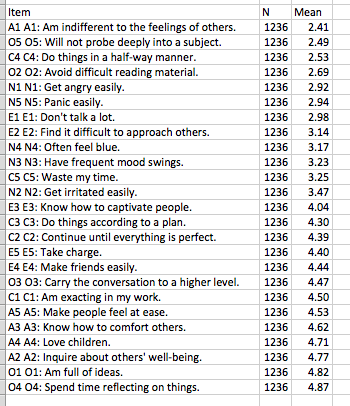


## 4. Examine means and frequencies of each response option to see whether any items lack variability (i.e., means that are very high or very low)

This is a common initial strategy for checking whether an item should be dropped. If it has a very high or very low mean relative to the response scale, then it may not adequately discriminate between respondents.

A quick check is "analyze - descriptives - descriptives" and put in all items.

Then a bit of Excel sorting by mean can produce the following table:



The items are on a 1 to 6 scale. The table itself is quite informative in terms of seeing what kinds of items get endorsed more or less than others. There certainly is a tendency for people to agree more with positive items than negative items. There is also something like an "everyday" factor guiding item endorsement. For example, knowing how to captivate people is positive, but it also represents something more difficult to agree with than merely spending time reflecting on things.

Regardless of the substantively interesting aspects of the data, there is the subjective question of whether any items suffer from floor or ceiling effects. I'd consider cut-offs of around less than 1.8 or greater than 5.2. But ultimately, this decision often interacts with the aim of the test development process, the number of items you have in the initial pool, and so on. If you have many items, then low or high means can be a flag that an item is less useful, which along with other indicators might be grounds for removal. In this case, I would not remove any of the items purely based on their means.

## 5. Perform an EFA to look at possible numbers of factors, and the degree to which items load appropriately on specified factors: (a) Should any items be dropped? (b) How many factors should be retained? (c) describe the structure of the test as suggested by EFA noting interesting features such as (i) relevant cross-loadings, (ii) what would be captured if fewer or more factors than finally specified were retained; (iii) whether some items load substantially more or less; etc.

The process of performing EFA is a bit of an art. First, there is a question of which items should be retained. Then there is the question of how the test should be used in terms of scoring scales, and interpreting scales, this generally relates to the question of how many factors represents the data well. And finally, there are broader questions around understanding the test itself in terms of the nuanced features such as factor correlations, cross loadings, and so on.

Let's start by looking at both the question of how many factors are appropriate and whether any items should be dropped.

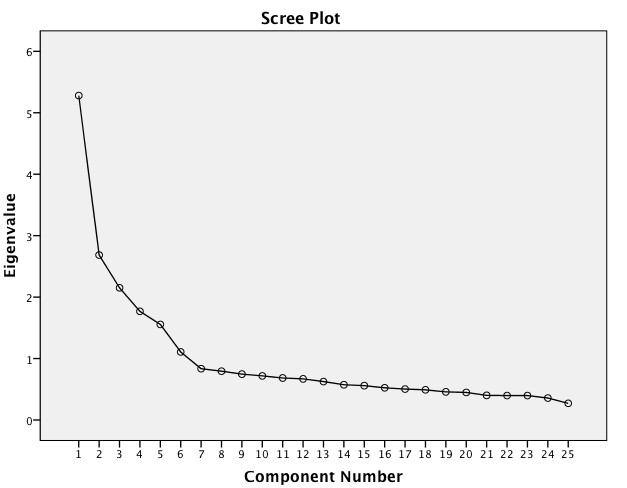
There are many decisions that have to be made: method of extraction, how many factors to extract, whether to rotate and what type of rotation. My starting point is commonly maximum likelihood with promax rotation (i.e., allowing correlated factors). In terms of numbers of factors I like to examine the scree plot. I also typically examine the theorised number of factors as well as a few more or a few less factors. I then look at how meaningful is each factor and look for signs of under or over extraction of factors.

So let's start by looking at the scree plot:

"analyze - dimension reduction - factor analysis"

variables: A1 to O5

extraction: scree plot



Several interesting properties can be observed from the scree plot:

* The first component is relatively large compared to the second and subsequent factors. This is consistent with some personality literature which suggests that a global factor underlies a portion of personality variance.
* Then there is a clear chain from components 2 to 5 before there is a noticeable drop to component 6 and another slight drop again at component 7 at which point the scree has clearly begun. Thus, the scree plot probably suggests 5 or maybe 6 factors would be good starting points.

Given that the test was designed to test 5 components, this seems like a good starting point for exploring the factor structure of the test.

So let's run our first proper factor analysis with 5 components:

"analyze - dimension reduction - factor analysis"

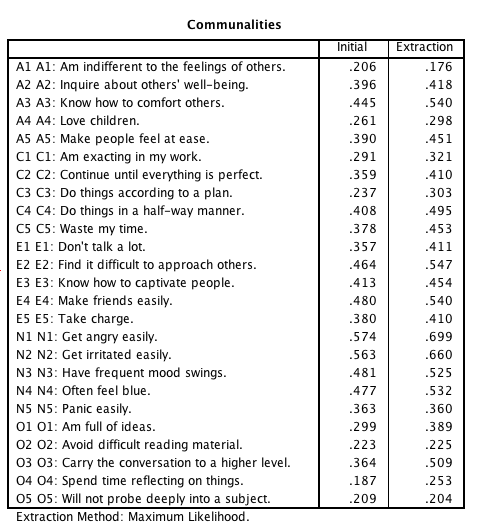
variables: A1 to O5

extraction: maximum likelihood; factors to extract = 5

rotation: promax

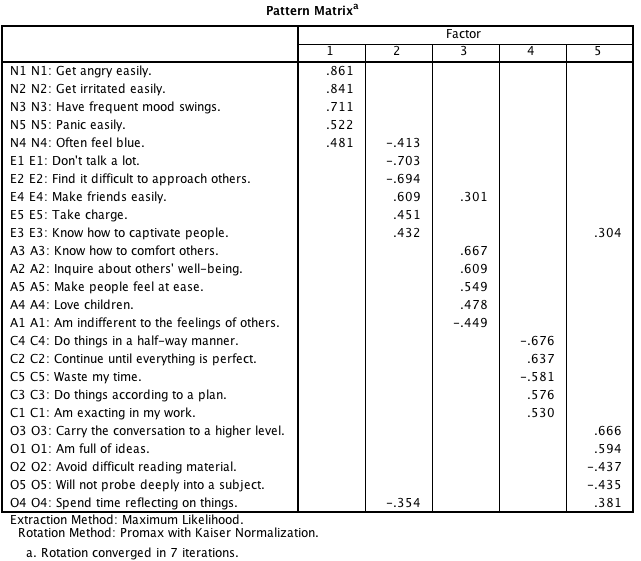
options: sorted by size; suppress small coefficients < .25

We can have a quick look at the communalities to see whether any items are poorly captured by the factor structure.



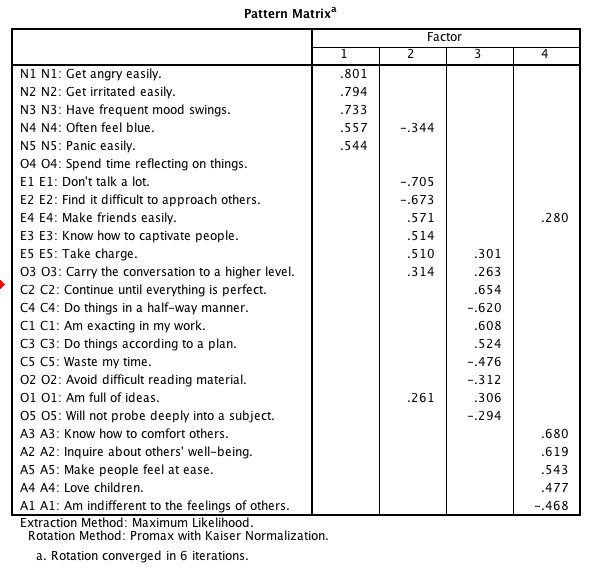
The initial column is the proportion of variance explained by other items. The extraction column is the proportion of variance explained in the item by other factors. This might flag A1, O1, O4 and O5 as being a bit low, but perhaps not so low that we'd discard the items.

The pattern matrix represents the loadings of the items on the factors. This is a very clean factor solution. Every item loads maximally on its designated factor, and there are only four cross-loadings larger than .25. We can talk more about the details of interpretation, but for now, the main point is that five factors looks pretty good.



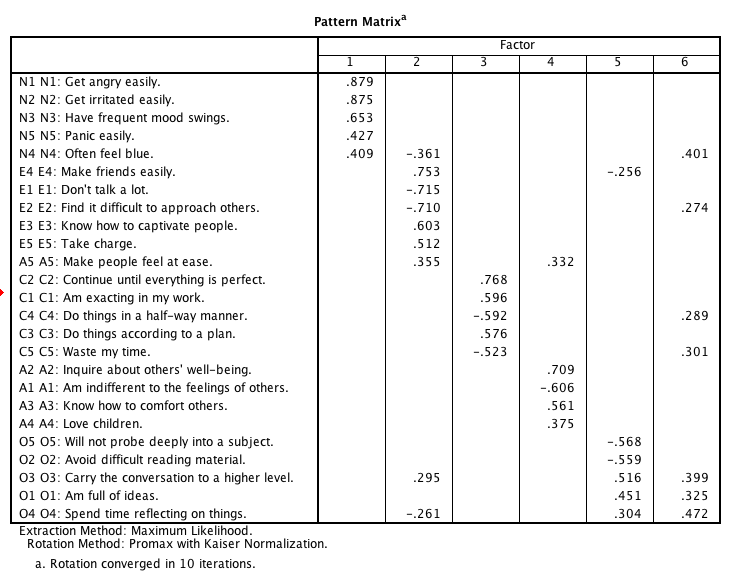
Let's contrast this with the pattern matrices for a few more or a few less factors.

Here's the four factor solution:



Mostly this has just collapsed Openness and conscientiousness together. More generally, the openess items appear to have lost a home. Item O4 has no major loading, item O3 loads on both extraversion and conscientiousness. O5 , O1, and O2 both have poor loadings. So combining theory and the pattern of loadings, this clearly looks like not enough factors.

Here's the six factor solution:



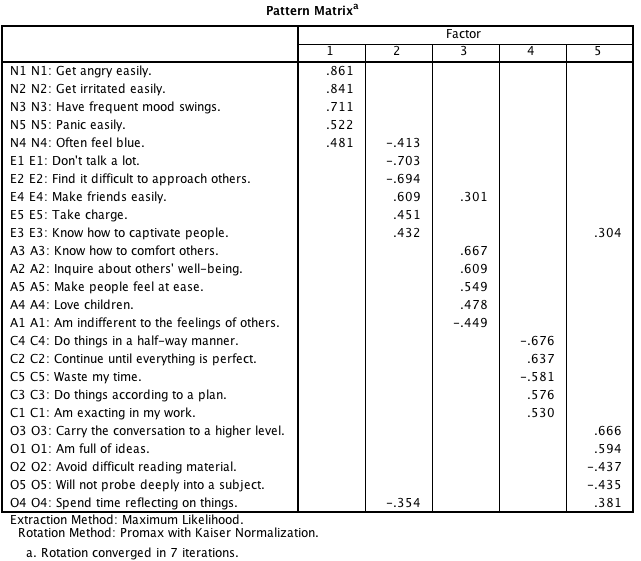
The first five factors here look fairly similar to the five factor solution. The sixth factor is a bit mysterious. If we take it seriously it seems to capture aspects of existing items: "feel blue", "difficulty approach others", "do things halfway", "waste my time", "conversation to higher level" "full of ideas" and "spend time reflecting". If I were to synthesise these items into a label, it sounds a bit like a stereotype for a philosophical disorganised introvert. That said, none of the loadings are very strong. And almost all the items load on the first five factors primarily.

Presumably even on good psychological tests there is at least some systematic variance beyond the core designated factors. We could try to analyse and understand these aspects of the data. That said, five factors seems to capture the most important aspects of the data. The sixth factor may have some meaning, but it does not seem to be particularly important.

Thus, I would conclude from this analysis that five factors is a good option for the test.

**Should any items be dropped?**

Returning back to the five factor pattern matrix, we can see whether any items should be dropped:

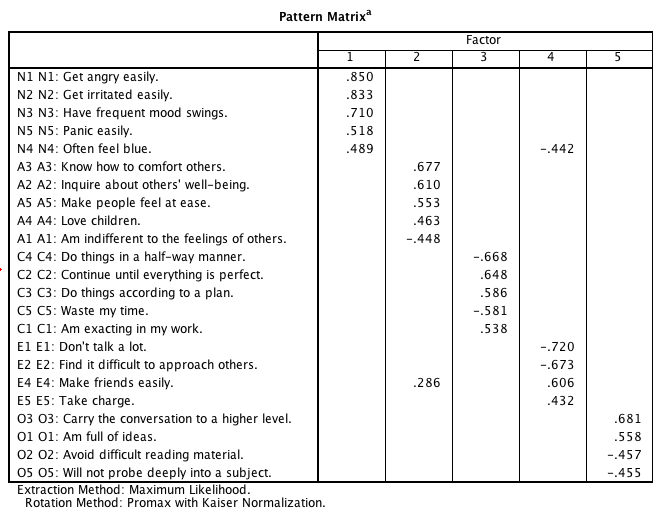


Problematic items are those with large cross-loadings, or items that don't load on the theorised factor, or items that lack any large loadings. In this case, the worst performing items are:

* N4: often feel blue; this is the lowest loading on the neuroticism factor and also seems to capture negative extraversion. However, from a theoretical perspective, this is a classic neuroticism item in that it capture the tendency to experience negative emotions.
* E3: know how to captivate people: This loads only moderately on extraversion and has a small loading on openness. This might be due to the concept of captivation involving both a social (i.e., extraversion) component and an interesting ideas component (i.e., openness).
* O4: spend time reflecting on things: This loads weakly on openness and negatively on extraversion. Presumably this is a positive aspect of introversion (i.e., quietly reflecting) as opposed to the fear of social interactions.

So, if I was going to remove any items, it would be E3 and O4.

We can see quickly what the test would look like without these two items.



The factor solution does look a bit cleaner.

In general when deciding to delete items, the decision is often influenced by a wide range of factors. Having only five items means that even removal of even a single item can risk reducing the reliability of the scale to an unacceptable level. This is a general reason why when developing a new scale it is wise to start with many more items that you ultimately hope to have. There is also the importance of content validity. In particular, it is important to consider what the meaning of the factor will be as you take out certain items.

Ultimately, I think the removal of the two items is something that would be open to debate. For the sake of the exercise, I'll exclude these two items.

**(c) describe the structure of the test as suggested by EFA noting interesting features such as (i) relevant cross-loadings, (ii) what would be captured if fewer or more factors than finally specified were retained; (iii) whether some items load substantially more or less; etc.**

This has largely be discussed already. In terms of the retained items, what remains is largely representative of what is known as simple structure. The exception is item N4 and to a lesser extent E4.

Interestingly, the centre of the neuroticism factor seems to be more anger and irritation than the more common tendency to experience negative emotions. This is probably because there is only one item measuring that aspect.

Similarly the extraversion factor seems to be centred more on ability to interact socially. General activity levels is a little less related. There also does not seem to be anything capturing desire for social interactions, as opposed to perceived ability. This emphasis on ability to form social interactions may explain the loading of N4 "feeling blue". I.e., people who feel they are unable to form social connections are more likely to experience negative emotions. In contrast, some people just prefer to engage in fewer social interactions.

## 6. Report means and standard deviations for each item of the test along with item text, and so on

Analyze - descriptives - descriptives

add all retained variables

options: just mean and sd

We also need to tidy up the table to remove various information such as the sample size that is repeated in every row. It's often good to copy and paste the table into Excel for such cleaning, although SPSS does provide options for table editing.

|  |  |  |
| --- | --- | --- |
|  | Mean | SD |
| A1 A1: Am indifferent to the feelings of others. | 2.41 | 1.38 |
| A2 A2: Inquire about others' well-being. | 4.77 | 1.24 |
| A3 A3: Know how to comfort others. | 4.62 | 1.31 |
| A4 A4: Love children. | 4.71 | 1.48 |
| A5 A5: Make people feel at ease. | 4.53 | 1.27 |
| C1 C1: Am exacting in my work. | 4.50 | 1.26 |
| C2 C2: Continue until everything is perfect. | 4.39 | 1.31 |
| C3 C3: Do things according to a plan. | 4.30 | 1.31 |
| C4 C4: Do things in a half-way manner. | 2.53 | 1.38 |
| C5 C5: Waste my time. | 3.25 | 1.64 |
| E1 E1: Don't talk a lot. | 2.98 | 1.64 |
| E2 E2: Find it difficult to approach others. | 3.14 | 1.62 |
| E4 E4: Make friends easily. | 4.44 | 1.45 |
| E5 E5: Take charge. | 4.40 | 1.37 |
| N1 N1: Get angry easily. | 2.92 | 1.57 |
| N2 N2: Get irritated easily. | 3.47 | 1.53 |
| N3 N3: Have frequent mood swings. | 3.23 | 1.61 |
| N4 N4: Often feel blue. | 3.17 | 1.59 |
| N5 N5: Panic easily. | 2.94 | 1.62 |
| O1 O1: Am full of ideas. | 4.82 | 1.15 |
| O2 O2: Avoid difficult reading material. | 2.69 | 1.56 |
| O3 O3: Carry the conversation to a higher level. | 4.47 | 1.21 |
| O5 O5: Will not probe deeply into a subject. | 2.49 | 1.32 |

## 7. Provide a description of how to score the test (e.g., which items belong to which scale) consistent with your preferred model of the data

There are a few options here.

I have chosen to score the tests as

N: mean of N1, N2, N3, N4, N5

E: mean of E1, E2, E4, E5

O: mean of O1, O2, O3, O5

A: mean of A1, A2, A3, A4, A5

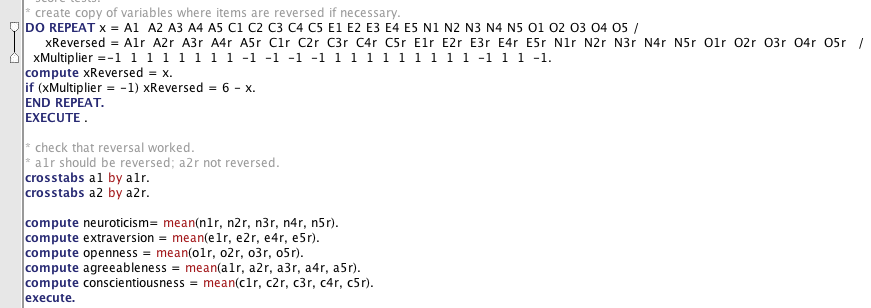
C: mean of C1, C2, C3, C4, C5

## 8. Score the tests

To score the tests in SPSS, we first have to reverse relevant items. We can then use a function like mean to create scale scores.

This post provides some examples:

<http://jeromyanglim.blogspot.com.au/2009/10/calculating-scale-scores-for.html>



Here is the copy of the syntax that I used.

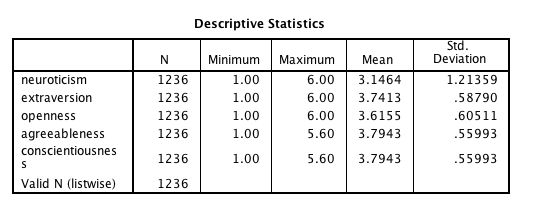
* Do repeat as mentioned previously can be used to loop over a set of items. In this case it reverses the item if necessary and copies it if it is not required. This requirement is stored in the vector of -1 and 1s.
* Crosstabs is used to check for a couple of examples whether the reversal worked correctly, i.e., items needing reversal were reversed.
* Scale scores can be obtained using "transform - compute", but pretty quickly it becomes easy just to write out the syntax for such things.
* Note that I've used the mean function, if we had missing data, I'd probably use something like mean.3 to ensure that at least 3 responses are required for a participant to obtain a score.

## 10. Report means, standard deviations, cronbach's alpha and intercorrelations for each scale

**Means and standard deviations**

Analyze - descriptives - descriptives

put in scale variables



Interestingly the level of endorsement is fairly similar for all but neuroticism.

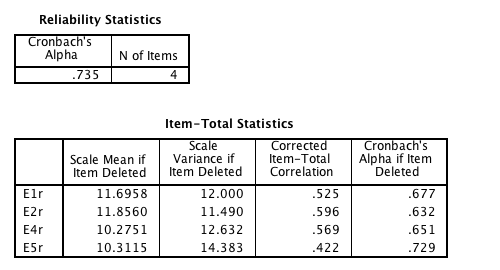
**Cronbach's alpha**

For each of the five scales go to "Analyze - scale - reliability"

Use the reversed version of the items.

You may also want to request "scale if item deleted" in order to check that items were properly reversed.

This for example shows the reliability of the extraversion scale (i.e., .735) and that all items have positive corrected-item total correlations which is consistent both with them being good items and correct item reversal.

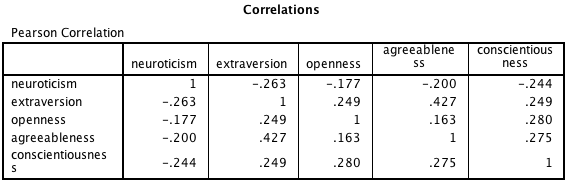


So the alphas were:

* Neuroticism: .823
* Extraversion: .735
* Agreeableness: .721
* Conscientiousness: .731
* Openness: .624

Broadly, this is less than ideal reliabilities, particularly for openness. This is presumably due to a few factors: (a) the number of items used is small; (b) the items within a factor are reasonable diverse, which is a good thing for content validity, but it does require additional items to get good internal consistency reliability.

**Correlations**Go to "Analyze - correlate - bivariate"

****

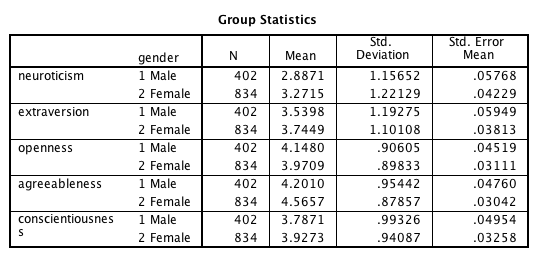
There are some modest correlations particularly between agreeableness and extraversion. The pattern of correlations is consistent with the notion of a global personality factor explaining some of modest amount of common variance.

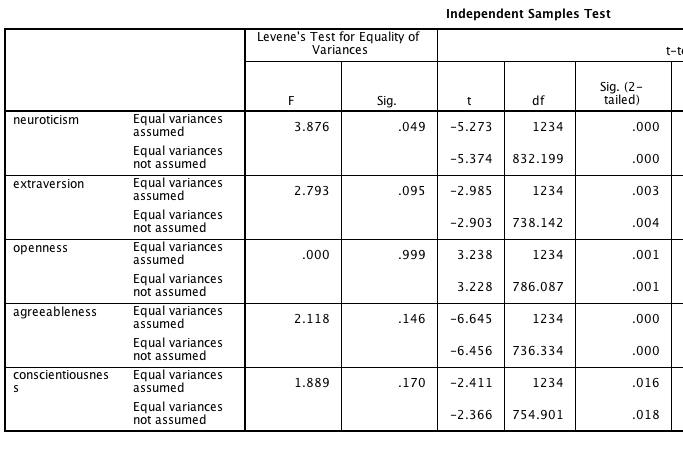
## 11. For any available demographic data, report the relationship with the scale scores

Gender:

We can do t-tests to compare means

Go to "analyze - compare means - independent groups t-test"





It can be helpful to calculate cohen's d to understand the size of all these significant differences. I do this by pasting the means and sds into excel and using a simple formula:

|  |  |
| --- | --- |
|  | F-M Cohen's d |
| neuroticism | 0.33 |
| extraversion | 0.17 |
| openness | -0.20 |
| agreeableness | 0.38 |
| conscientiousness | 0.14 |

This table shows the amount to which females are higher than male means in terms of the male standard deviation. The differences are in the small to medium range. In particular females appear more agreeable and neurotic.

**Education**

When looking at educational category, it is an ordinal scale.

Thus, one option for examining effect of education is using ANOVA with a linear and quadratic contrast. This allows for the examination of group differences, and to see whether there is any linear or qudratic effect (i.e., more education then more or less of a personality factor, and this might be u-shaped to some extent).

One way of doing this is:

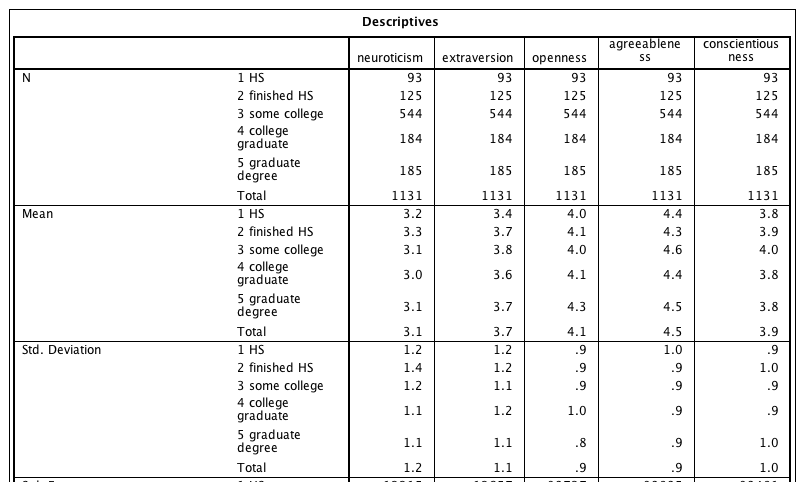
Analyze - commpare means - one way anova

put in scales as dv and education as factor

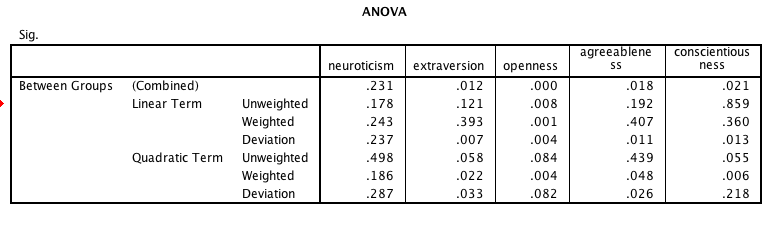
options: descriptives

contrasts: polynomial quadratic

The table of means and sds suggests some small differences (note that pivot trays was used to re-arrange the output).



This table of p-values for the combined, linear and quadratic effects shows the significance tests.



**Age**

Age is a variable which is well known to often have quadratic relationships with other variables.

A scatter plot can be a useful graphical way of examining this.

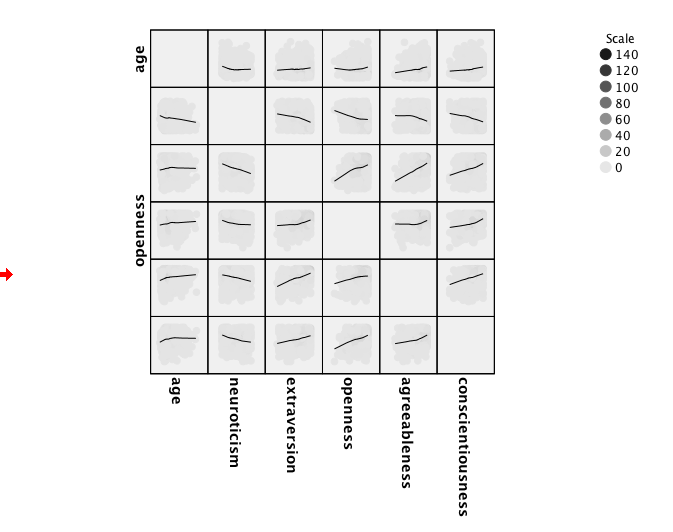
The plot below was obtained by going to:

graphs - legacy - scatter (matrix scatter)

putting in age and the five factors as variables

double clicking on the graph and choosing fit line at subtotal (loess) and bin elements by colour

The loess line is a localised regression fit which can highlight any non-linearity in the relationship. In this case, there is some slight kinking in the relationships between age and personality variables but the relationships look at least monotonic.



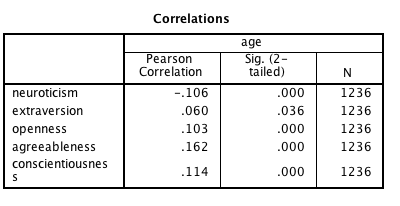
Given that a linear relationship seems like a reasonable approximation, let's look at the correlations between age and personality.

Analyze - correlate - bivariate will do.

A clever trick you can do with syntax is just show the combinations of interest. See the syntax help for correlations to learn more.

Thus, this syntax produces the following more focussed table.





This shows some modest correlations between personality and age. As people get older they get a little more open, agreeable and conscientious, and a little less neurotic.

## Comments on syntax file

I encourage you to have a look at the syntax file in the output folder. This will give you a feel for how you might structure your syntax file to help make your analyses more reproducible.