# Instructions

The following exercise builds on the previous exploratory factor analysis exercise. In the data folder you have exploratory and confirmatory SPSS data files. We'll start with the exploratory data file. But for final reporting purposes you would run all the retained models using the confirmatory data file.

The aim of the following analyses is to assess the relative model fit of various models of this personality test and to develop a good model.

We will fit the models using all 25 items, but note that potentially you could exclude one or more of the items from the set of analyses. Nonetheless, to make model comparisons meaningful it is generally best to use a common set of items. Thus, if you delete a exclude a few items, exclude them from all models.

There are many possible models that you could fit.

1. We will fit:

(a) a one factor model

(b) a five factor model with uncorrelated factors

(c) a five factor model with a single global factor

(d) a five factor model with all factors inter-correlated

(e) a more data driven model where we take one of the existing models and add some reasonable modifications to improve the fit (add a justification based on theory and modification indices for this model)

2. Create a table with the model fits for each model (i.e., rows are models, columns are fit statistics): include at least chi-square, df, RMSEA, SRMR, and CFI.

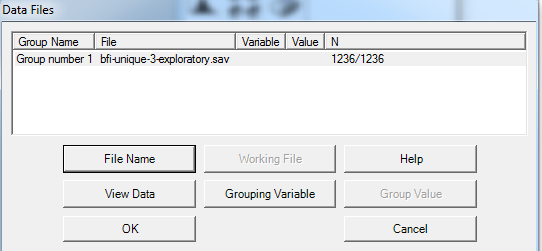
# Answers

## Getting Started

Start "Amos Graphics".

We first need to connect the SPSS data file to Amos.

Go to "File - Data File" and click "file name" and select "bfi-unique-3-exploratory.sav"

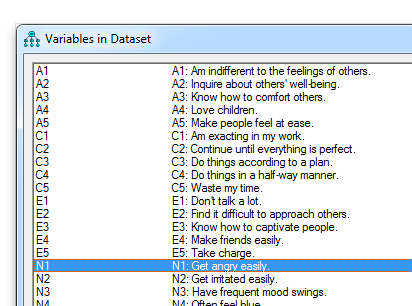


This will hopefully confirm the sample size that you were expecting.

In Amos you can click on the following icon:



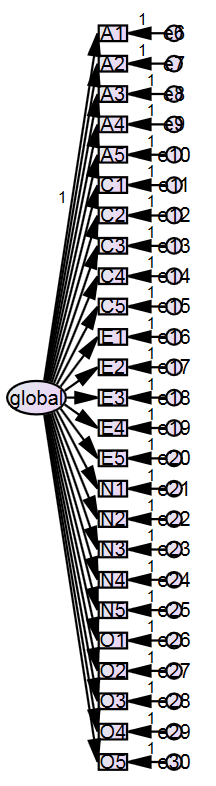
to bring up a list of the variables in the data file. These variables can then be drag and dropped.



Part of the skill of working with Amos is learning to use the range of tools available to draw and manipulate path diagrams. I'll explain in detail what the tools do in relation to the first model.

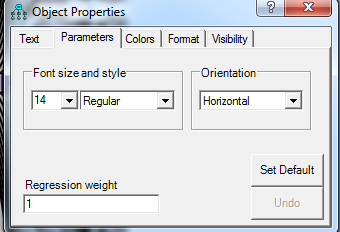
## (a) a one factor model

The first model is going to look something like the following. There is one latent factor which I've labelled "global", which is causing the individual items. In addition each item has an individual latent residual variable that is predicting what is not predicted by the latent factor.



To create this diagram, use the "draw indicator variable" tool . Once the tool has been activated, hold the left mouse button down on the path diagram page, and drag down. This will create an ellipse that will represent the latent factor. Then click on this ellipse to create indicator variables. You will need 25 indicators, one for each item. You can use the rotation tool  to change the orientation of the indicators relative to the latent factor. Also, be mindful of how large you draw the initial latent factor as this will influence the default size of subsequent indicators.

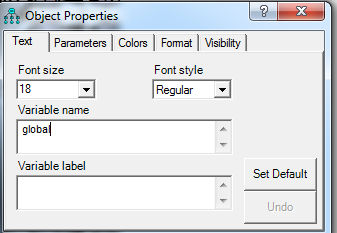
You should also note that one of the arrows from the latent factor to the indicator has a value of "1". The draw indicator tool does this by default.



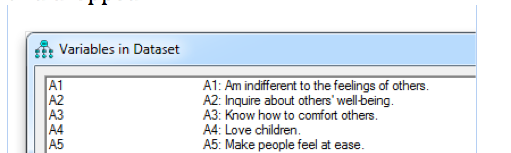
This is one way of ensuring that the latent global factor has an identifiable metric. Another option is to remove the regression weight and give the global factor a variance of 1.

The next step is to name (a) the latent factor, (b) the indicators, and (c) the residuals.

(a) **latent factor:** To name the latent factor double click on it to bring up the object properties and enter a descriptive name in "variable name". This would typically be something like "extraversion", "neuroticism", or whatever you are calling the latent factor. In this case, for a one factor model I've named it "global" as in global personality factor.



(b) **indicators:** To name the indicators (i.e., the rectangular boxes), these need to be the same as the variable names in the linked SPSS data file. This is because indicators represent observed variables. There are two ways to do this. One, you can type the variable name directly into object properties for the particular box. Or two, you can drag and drop the variable name from the "Variables in dataset" box into the indicator rectangle. This second approach is often more reliable than typing the name yourself. It also means that the variable label will be copied as well.

****

Also, note that by default, variable labels are displayed instead of variable names in the path diagram. This can get far too messy if you have long variable labels. If you want to just display the names, you can go to "view - interface properties" and "misc" tab and then unselect "display variable labels".

(c) **residuals:** The residuals also need a unique name. Often they are just labelled e1, e2, e3, and so on. Alternatively, you could label them something like "A1\_e", "A2\_e" and so on to make it extra clear that they are the residual for a particular variable. If you are going to be explicitly adding paths between residuals then you may want this second scheme. If you are just wanting to ensure they have a unique name, then you can use "plugins - name unobserved variables" to quickly add names using the e1, e2, e3 convention.

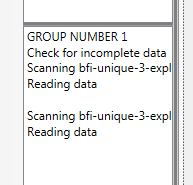
We have now created the model. The next steps are to configure, run and then interpret the analysis.

**Save the model:** In Amos there are many different files that are created. The files that store the data files have an "amw" extension. Given that you are going to be creating a few of these files, it is important to adopt a useful naming convention. I like to number my models and then have a descriptive name (like "01-one-factor.amw"), and the next model could be "02-five-factor-uncorrelated.amw".

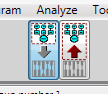
**Configure analysis:** To configure the analysis click on the abacus symbol with some coloured squares . This brings up a range of options. A few points

* Estimation tab: This gives you a few options for estimating parameters in the model and defining fit.
* Numerical: This defines various convergence criteria for determining when to stop iterative fitting algorithms.
* Output: This is one of the major tabs that you are likely to want to alter. By default, you get fairly minimal output. In particular, we almost always want "standardised estimates" (i.e., the standardised loadings). That's probably all we need at the moment.
  + Squared multiple correlations: these are particularly useful in SEM where you have a variable that is being predicted. It's like R-squared in a regression.
  + sample moments: Moments include means, variances, and covariances (as well as standardised covariances, i.e., correlations). Sample moments are thus, these statistics in your sample data.
  + implied moments: These are the moments implied by your model. Thus, if the implied moments differ a lot from the sample moments, then you have a bad model.
  + residual moments: These represent the difference between sample and implied moments. These are useful for flagging aspects of the data not captured by the model.
  + modification indices: These are useful when you are looking for ways to improve your model fit.
  + Indirect, direct, and total effects: This is useful in SEM where you have mediational pathways.
* Bootstrap: This is an alternative way of getting parameter estimates and standard errors particularly suited when you have non-normal data.

**Run Analysis:** Click on the abacus to run the analysis (the abacus metaphor makes sense if you consider that the abacus is an old tool for performing mathematical calculations). If estimation was successful you should see some output in the second from bottom box:

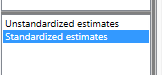


and you also see a change to the input - output arrows at the top left of the screen. Specifically when the output button lights up red, then you can click to view output of the model.

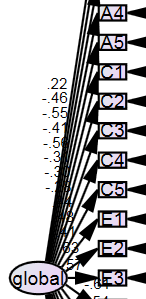


**Interpret Output:** There are a few ways of viewing output.

1. You can click the red arrow shown above. This will display coefficients directly on the path diagram. You can also click standard or unstandardised estimates to change whether what is displayed on the path diagram is standardised or unstandardised.



Here's a subset of the output of the standardised output on the path diagram:

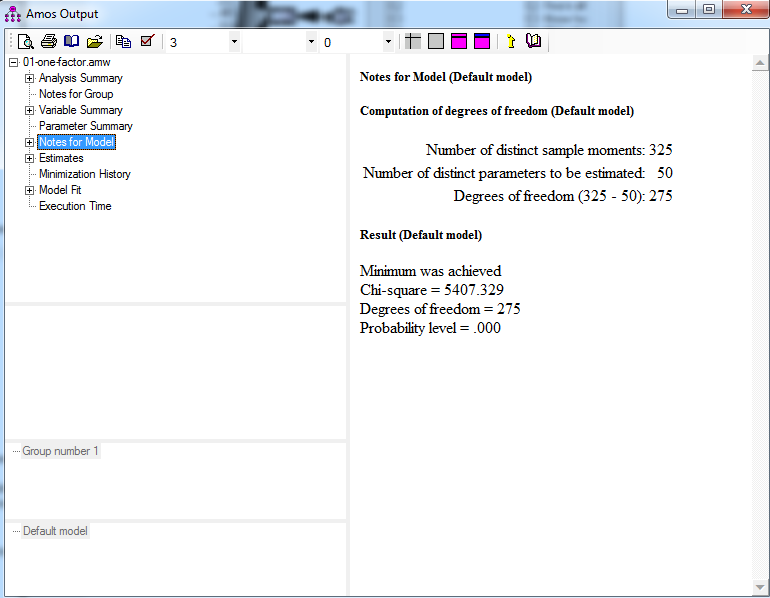


These numbers are equivalent to loadings in a factor analysis. In this case they represent the correlation between the factor and the item.

However, as you can see, it is very difficult to read such output when you have a lot of items. Thus, for both cleaner output and more output, there is the Amos output viewer.

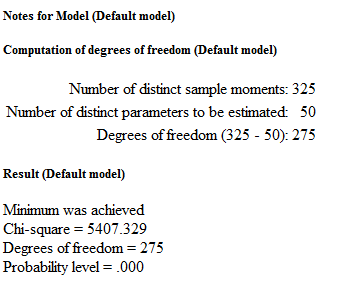
2. Amos output: To view the amos output viewer, click the spreadsheet icon .

A screenshot is shown below. There is a navigation bar on the left and output on the right.



If you asked for more output in the model configuration dialog box, then you that output would be displayed here. There are many things you can examine, here is some relevant information to look at:

**Notes for model**

****

**sample moments**: There are 325 sample moments I.e., if you think about a correlation 25 \* 25 covariance matrix, there 25 variances and (25\*25 - 25) / 2 covariances = 300, which equals 300 + 25 sample moments.

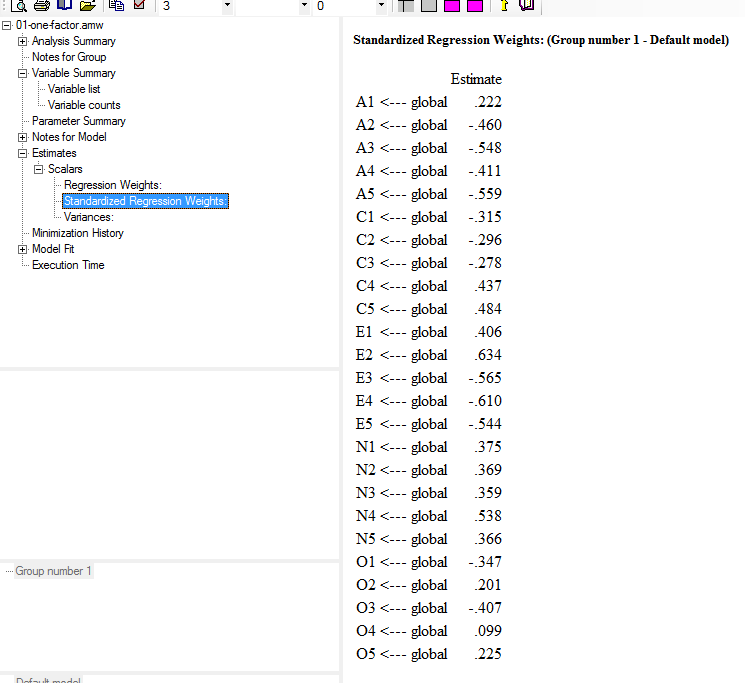
**Estimated parameters**: you have 25 residuals, 24 regression coefficients (i.e., one was fixed to 1), and 1 latent factor variance (i.e., 25 + 24 + 1 = 50).

**Degrees of freedom:** I.e., 325 - 50=275 degrees of freedom.

**Model convergence:** This overall summary shows that the model converged ("minimum was achieved")

**chi-square:** This is an overall measure of discrepancy between the model and the data. It is influenced both by the sample size and the size of the discrepancy. Higher values mean poorer fit. However, it should rarely be interpreted in isolation. Rather, it is useful when comparing different models, and it is also useful as a basis for creating other fit indices.

Standardised regression weights:

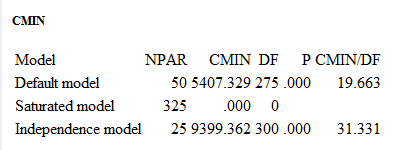


Above shows the standardised regression weights, which we could not easily see in the path diagram view. For example, it shows that certain items load strongly on the global factor such as E2, E4, A3, A5, etc. and some items load weakly (e.g., O4, O2, etc.). By looking at the direction of the loadings and the size of the loadings we can see that high scores seem to reflect more negative emotions, fewer positive emotions, and perhaps lower self-esteem. In general, high scores seem to relate to a social undesirability factor, whether this be real or the result of a response bias.

Regardless, of the specific issues here, you should observe how the meaning of latent variables can be understood in the same way as you would interpret a factor analysis factor. You should also note that regardless of model fit, it is still important to try to understand the meaning of the model in terms of what the coefficients mean.

**Model fit**

Amos provides a wide range of model fit information. I'll focus on chi-square, df, RMSEA, SRMR, and CFI. In general it is a good idea to have a separate table where you record this information for each of your models.



CMIN means "Chi-square". Amos provides fit information for three models.

* **Default model:** The default model is the model we actually drew.
* **Saturated model:** The saturated model assumes every variable is correlated with every other variable. Thus, there is no discrepancy between model and data, because the model is the data.
* **Independent model:** This assumes that the correlation or covariance between the 25 variables is zero. Thus, the only parameters estimated are the variances for the 25 variables. As we know, there are strong correlations between the items, so this model has a huge chi-square.

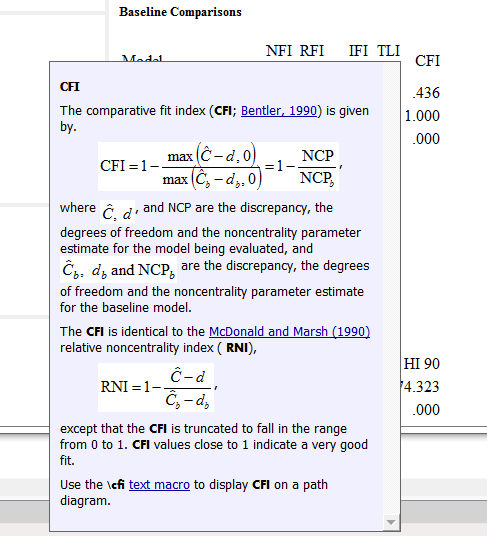
NPAR is the number of parameters estimated by the model

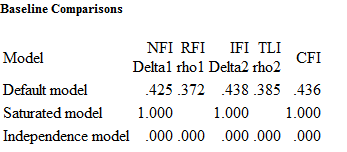
DF is the number of degrees of freedom

CMIN is the chi-square value for the model where larger values mean greater discrepancy between model and data

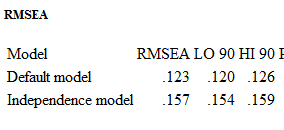
CMIN/DF is a measure of chi-square per degree of freedom. In general, lower chi-squares are good and more degrees of freedom are good. However, there is a trade-off. With fewer degrees of freedom we can get better fit. CMIN/DF is a simple index of parsimony adjusted fit. That said, there are probably other parsimony adjusted measures that are more suitable.

In general you can click on these names and get further information about the formula used to calculate them. E.g., here's the CFI info:



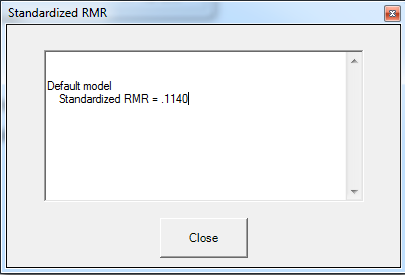


CFI is a commonly reported fit measure. It ranges from 0 to 1. Values above .9 or .95 are often considered indicators of good model fit. But once again, it is often best to interpret it in a comparative way. Nonetheless, we weren't expecting a one factor model to be very good, and this is coming across in the CFI.

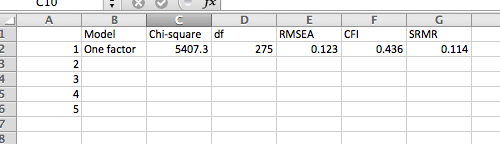


RMSEA is a measure of model discrepancy. A rule of thumb is that RMSEA < .05 represents close fit and <.08 represents reasonable fit. Clearly, here we have poor fit.

Standardized RMR: For some reason, getting the standardized root mean square residual takes a bit more effort. Click "plugins - standardised RMR" which will open an empty box, then re-run the model. This will then populate the box with the SRMR. A rough idea of SRMR can be to think of it as the typical absolute difference between model and data correlations, although this is a simplification. The point is that values closer to zero are good. Some have suggested ( http://davidakenny.net/cm/fit.htm ) that values below .08 are good.



Thus, after getting this fit statistics, we can start to build our table of model fits:

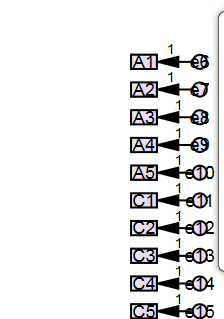


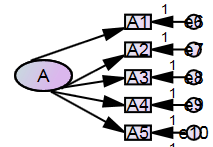
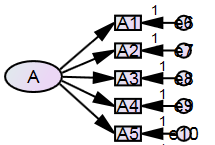
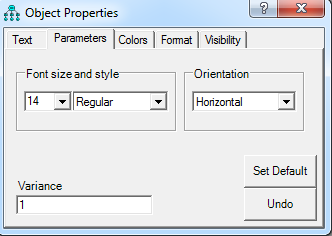
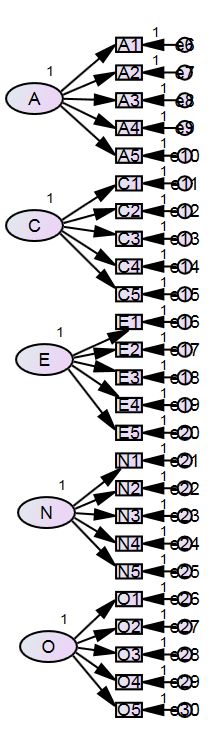
## (b) a five factor model with uncorrelated factors

After saving the previous model, then click save as and save the next model (e.g., "02-five-factor-uncorrelated.amw". We now want to modify the diagram so that we have five factors. One option is to start again. The other is modify the existing diagram. You may find often that you run into trouble when drawing your path diagrams and that it is often easier to start the diagram again. For example, you could use the latent variable tool  to create five latent factors each with five indicators.

**Specify model:** I'll modify the existing diagram instead.

1. I'll delete the global factor using the cross tool  and then click the global factor. This will also delete all the loading arrows:

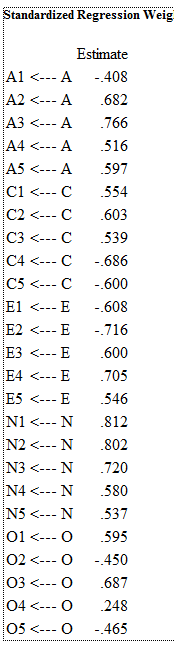


1. Draw five latent factors, one for each of the five personality factors using the draw unobserved tool . Hold the mouse down and drag to draw. By convention you should draw this to have an ellipse shape (ellipses for latent factors) rather than a circle (circles for residuals). If you want to replicate the exact shape of the first ellipse you can use the photocopy tool to create a copy .
2. Give each of these latent variables names. E.g., I've labelled mine E, A, N, O and C for the five factors respectively.
3. Add directed arrows from each factor to their corresponding items using the arrow tool . So you might get something like this in the case of the agreeableness factor:   
     
   It looks a little messy, so you can click on the magic wand tool and select the factor to tidy up and get something like this:  
   
4. For each of these factors, you need to constrain the variance. You can either add a 1 to the parameter of one of the loadings or define the variance of the factor as one. I'll constrain the factor variance to 1. Double click on the factor to bring up object properties and go to the "parameters tab and set the variance to 1:  
     
   At the end of this process, you should have a path diagram that looks something like this:  
   

**Running the model:** We can then run the model as usual first selecting the SRMR option and clicking run. You will be notified that you have specified uncorrelated factors. This is likely to result in poorer fit, but that's the whole point of this model. It will allow us to compare the fit to another model with correlated factors to see what the benefit is of allowing correlated factors.

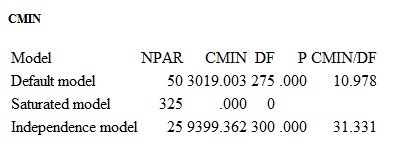
So click "proceed with the analysis".

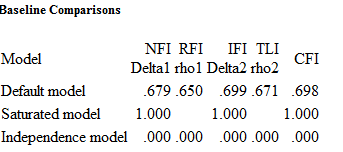
**Interpret output:** You could look at the standardised regression weights to understand how much each item loads on each factor.

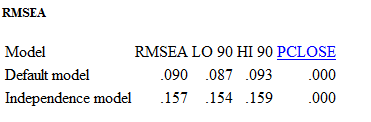


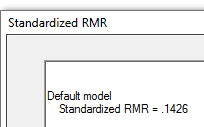
Notice how there are many fewer numbers that in a standard factor loading matrix in exploratory factor analysis. This is because all the other loadings are fixed to zero.

We can also extract the model fit:

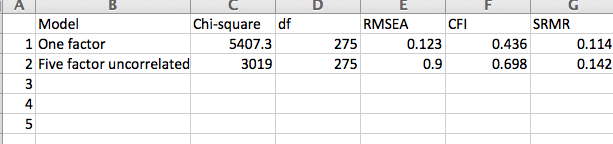








We can then add this data to the table of model fits:



## (c) a five factor model with a single global factor

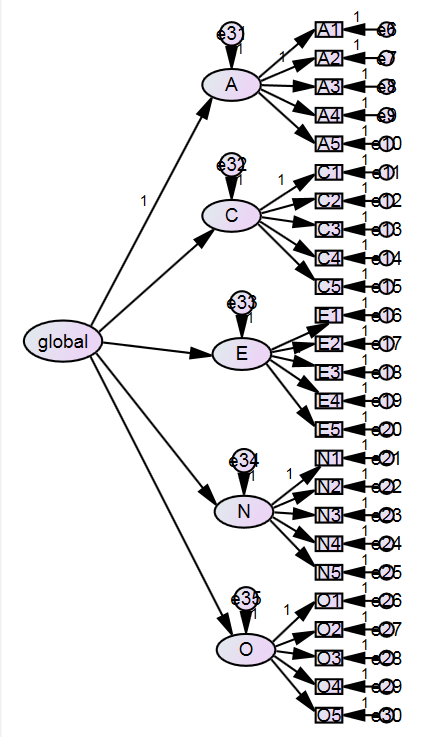
This model proposes a higher order factor that each of the big 5 might load on.

**Specifying model:** This requires drawing another elilpse and putting directional arrows to that factor. Once again this higher order latent factor also has to be given a defined metric and one way to do this is to constrain the variance of the factor to 1.

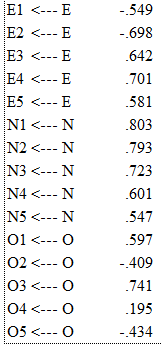
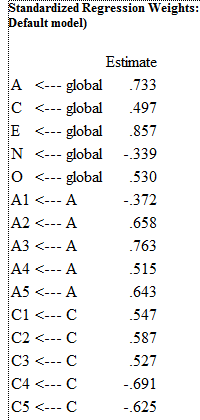
To do this

1. I needed to move my entire diagram a little to the right in order to make space for the higher level latent factor, you can do this by combining the select all tool (i.e., all the fingers are up) with the move tool  (i.e., a picture of a truck; trucks are for moving things). After doing that it's good to unselect all  (i.e., all fingers are down).
2. Add a latent variable and name it global
3. Draw directional arrows to each of the five latent factors from the newly created global factor. This should remove the variance constraint to 1 for the five factors.
4. Add residuals to each of the five latent factors  and name these residuals using the "plugins - name unobserved residuals tool".
5. Constrain one of the loadings between the global factor and the five latent factors to 1. I chose the arrow from global to agreeableness.
6. Because we can't use the variance equals 1 strategy to constrain variances of the factors, we should constrain one of the loadings from factor to item for each of the five factors to 1. For example, you could click on the arrow from A to A1 and constrain the parameter to 1 (but see below that A to A2 is a better choice). Then repeat for the other four factors. However, it will be best to choose an item that is positively worded for the factor. Otherwise, the meaning of the factor will be the opposite to what you might expect (i.e., extraversion will be introversion). So, for example, I chose A2, C1, E3, N1 and O1 for the constraints.

The end result could be a model that looks like this:



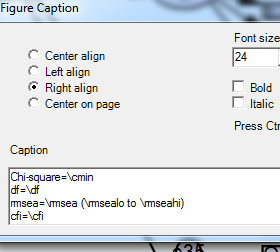
This then results in a number of interesting standardised regression weights.



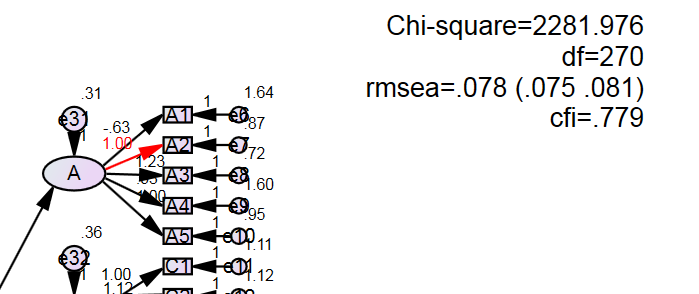
* The coefficients from the global factor to each personality factor indicate the direction and degree of relationship with the global factor. Interestingly, agreeableness, and extraversion appear to be the most strongly related factors to the global factors. Also, not surprisingly neuroticism is negatively related to the global factor.
* A quick look at the item loadings on factors suggest that they have not changed much from the uncorrelated five factor model.

A trick for quickly seeing relevant fit statistics is as follows. Each fit statistics as an abbreviation. See the help file on text macros for information. The basic idea is to add a title using the title tool , and then add a caption using the text macro codes. For example, \cmin is chi-square and \df is degrees of freedom. You can look up the codes under text macros in the help file.

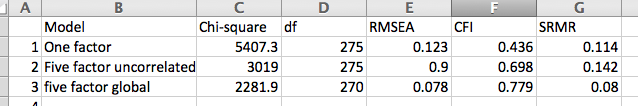
So for example, I specified this



which in turn displays in the model output like this:



This makes it a lot quicker to see your chosen fits at a glance. However, you still need to manually get SRMR from the plugin.



Recording these fits, it becomes very clear that this third model was a substantial improvement over the previous models. for only five additional degrees of freedom, we have dramatically reduced chi-square, and RMSEA and SRMR are approaching reasonable levels. That said, there are still presumably more improvements we can make.

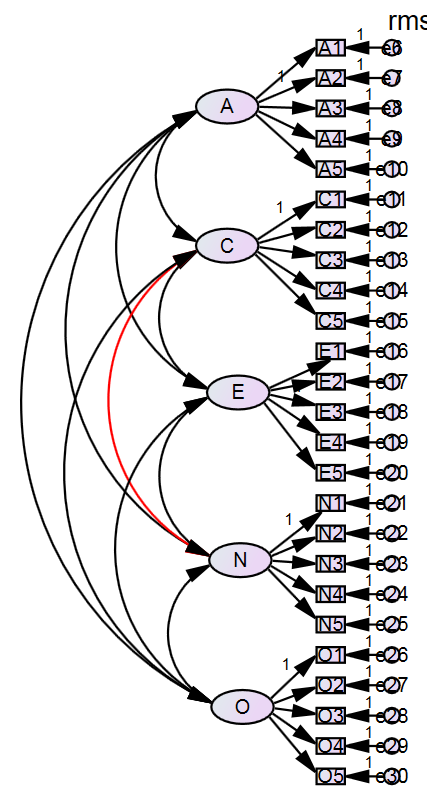
## (d) a five factor model with all factors inter-correlated

This model assumes that all of the big 5 factors are intercorrelated to potentially varying degrees.

To create this model

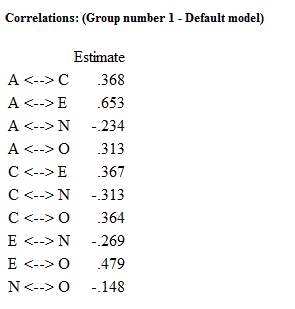
1. Delete the global factor and delete all the residuals from the previous model
2. There are two ways to create correlations. First, practice drawing double headed arrows between each of the latent factors . Note how if you go bottom-up the curve goes one way and how if you go top-down it goes the other way. In theory you could do this to draw 4 + 3 + 2 + 1 = 10 intercorrelations. However, there is a trick to speed up this process. Select all five factors using the single item selection tool  (i.e., a hand with one finger up). Then go to plugins - draw covariances. This will automatically draw double headed arrows between all specified covariances. It will also do so probably more neatly that you would manually.

It should look something like this:

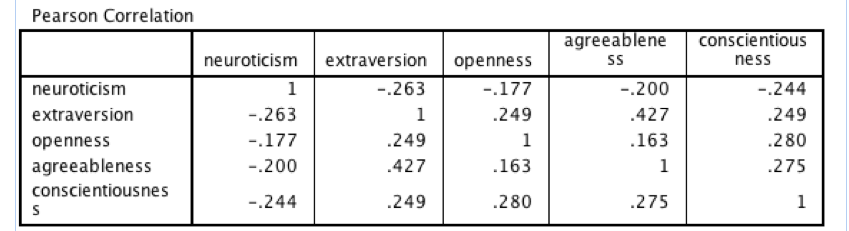


**Output interpretation:**

This will give you some additional output such as the correlations between latent factors:



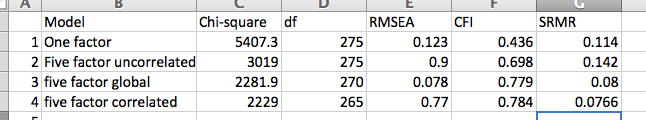
Note that these are the correlations between latent variables. Compare this to the correlations between scale scores from the previous exercise (albeit a few items are different):



In general, such correlations in CFA should be larger because they are in a sense adjusting for measurement error and estimating correlations between the latent variables assumed to underlie the observed items.

We can also see that certain correlations are larger than others. For example A and E correlate quite highly which may explain their high loading on the global factor.

When we add the fit information we get the following:



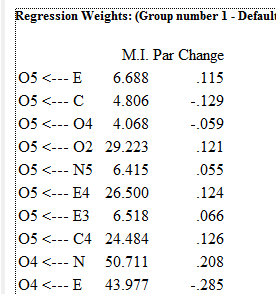
This model is an improvement over the previous model, but the improvement is quite small in comparison to the improvement that resulted from going from uncorrelated to either model 3 or this model.

## (e) a more data driven model where we take one of the existing models and add some reasonable modifications to improve the fit (add a justification based on theory and modification indices for this model)

There are many more models that we could explore. At this point I will use the opportunity to highlight the role of modification indices to see if it suggests any important cross-loadings or possibly correlated residuals.

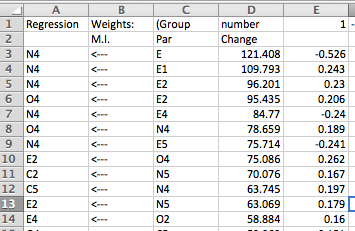
Given the five factor correlated model is the best so far, I'll use that as the starting point.

To get modification indices, go to "analysis properties" and "output" tab and select "modification indices" and then re-run the model. In the model output, you should have a modification indices tab. Let's start by looking at the regression weights. The top of it looks like this:



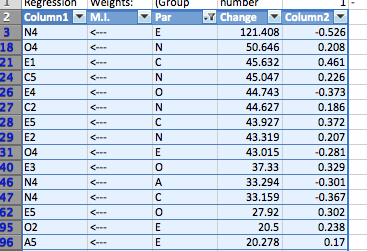
The M.I. is the modification index. It is a lower bound estimate of the reduction in chi-square that would result from adding that particular path. So for example, adding a cross-loading from E to item O5 should reduce chi-square by at least 6.688. Let's test this proposition. At present the basic correlated factors model has chi-square of 2229.047, now add an arrow from E to O5, the chi-square becomes 2217.347. So actually, it reduced chi-square by around 12, which is quite a bit more.

Thus, we can use modification indices to identify paths that we could add to improve our model. You can right click the modification index table to copy into Excel and then sort descending by modification index to flag the modifications that are likely to lead to the greatest improvement in fit.



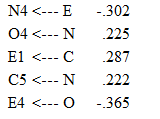
That said, it is important to incorporate theoretical judgement in selection the modifications. Typically in a CFA context, items are assumed to be driven by latent factors, so cross-loadings are reasonably to include. Otherwise, it tends to be residuals between items tend to be permitted if there is something compelling to say that two items share a unique feature such as a common wording or some particular common element.

So I converted the modification indices in Excel to a "table" and then filterdd the "par" column to only include the factors, which yielded the following:



In general, you would make a single change and then re-examine the modification indices. To speed up the process and just to highlight how it works here for this exercise, let's make the first five changes recommended.

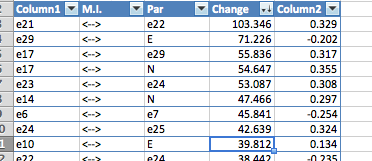
We do this by adding directional arrows (e.g., from E to N4). When we do just that our chi-square has gone down 1818.7 for just five additional degrees of freedom. You could also see these coefficients in the standardised regression weights:



Examination of item content should help to make this clearer. For example, N4 "often feel blue" loads negatively on extraversion. This might be because extraversion includes a lot of self-esteem and positive emotion related items, whereas most of the neuroticism factor focuses on anger, irritation, whereas N4 is one of the few purely emotion focused items.

Let's also highlight how you might add residuals between items if it makes sense.

Re-run the model and look at modification indices - covariances. Then, copy the table into Excel and sort by MI.

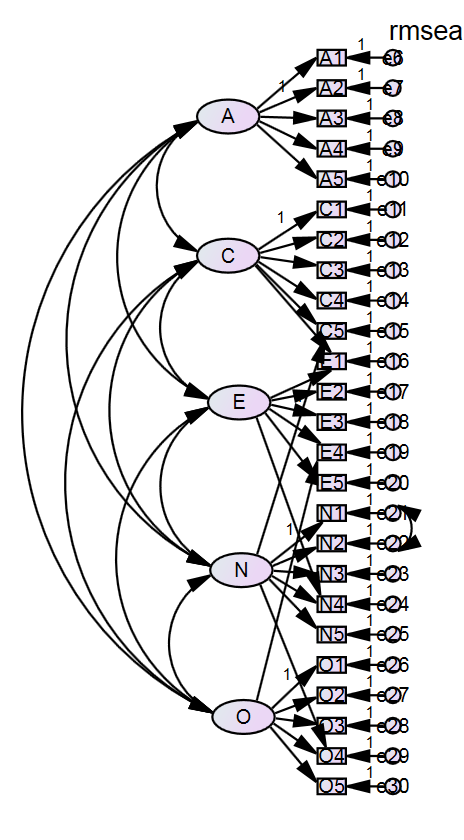


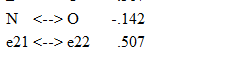
In this case, it is really just the covariances between item residuals which we might consider adding if it makes sense. When you start doing this it can be nicer if the residuals actually have names that relate to their items, but we wont bother with that for now. In my model e21 and e22 correspond to the residuals for N1 and N2 respectively.



The wording of these two items are clearly very similar. First irritation and anger are quite similar emotions, albeit on a spectrum. Secondly the actual text is the same but for the middle word. So we might choose to add a double headed arrow between the residuals for these two items.

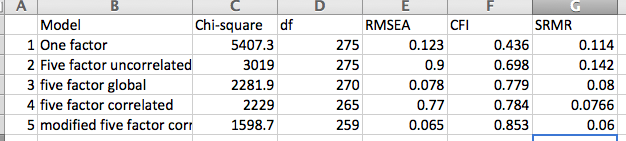
You could keep going with this process if you were serious about developing a good nuanced model based on modification indices. But we'll stop here. The finished model should look like this:

  
We can examine that new correlation under estimates - scalars -correlations:



We see that this is quite a large correlation between the unique aspects of the two items (i.e., that which is not captured by the neuroticism factor).

At the end of this process we are starting to get reasonable fits:



That said, a bit more work with modification indices could presumably improve this further.

Furthermore, the modification indices approach is quite data driven. This is why you ultimately want to report the fits that you obtain in your confirmatory sample. This is particularly important for the model driven approach.

## Summary thoughts

Hopefully through this process, you are getting a sense of how a model comparison approach works to confirmatory factor analysis.

* It is best to have several models. Start with simple naive models and gradually add complexity. Interpret fit statistics relative to other models. Justify additional complexity (i.e., fewer df) by the theory and improvement in fit.
* Models have fit statistics which guide us in deciding whether the model is an acceptable representation of the data. They also have regression weights, variances, and covariances (and their standardised versions) which help to explain what the model means.