Factor Analysis: Quit Being a Whiny Baby And Learn It Using SAS Enterprise Guide

AnnMaria DeMars, The Julia Group & 7 Generation Games, Santa Monica, CA

ABSTRACT

Why should YOU know about factor analysis? Imagine yourself in this scenario - someone, maybe you, has collected survey data at great expense. Maybe you paid subjects to answer questions about themselves, gave students credit to participate in a study, and now you have dozens, perhaps hundreds, of variables on each person. How on earth do you analyze these data? You could just go through and start putting questions together to form sub-scales, but that is pretty arbitrary. Enter factor analysis to help you make sense of your data.

Factor analysis is extremely useful. Conceptually, it is relatively easy to understand - mathematically, um, not so much so. Fortunately, SAS Enterprise Guide will have you analyzing with ease (really). You take a large number of questions and find what few, underlying traits they represent, such as supervision, collaborative decision-making and ambition. How do you know the number of factors? How do you decide which survey item goes with which factor? Why would you rotate and which rotation would you use?
This presentation uses examples from the 500 Family Survey to demonstrate how to: generate screen plots, factor patterns and commonality estimates with SAS Enterprise Guide. Decision-making rules for number of factors, rotation and commonality are discussed. The need for iteration is explained and the ease of iteration with SAS Enterprise Guide is illustrated. Don't be scared - if you even halfway understood correlation in your college statistics course, you can master factor analysis. You'll be glad you did. (Code will also be provided for the Enterprise-Guide-phobic.)

INTRODUCTION

The first step in programming is - THINK. Let's start our discussion of factor analysis by thinking about the problem that it solves.

Let's use the 500 Family Study data set as an example. It has 460 different items. Even if you select only the variables that are of particular interest to your research topic, let's say, relationships between parents and children, you have 42 individual questions. How are you planning to discuss these?
If you are like too many people, you'll give 42 statements like:

- 9% of adolescents reported parents never checked their homework
- 23% said their parents checked it rarely ….

blah blah blah for 42 variables.
There are three problems with doing it this way:

1. No one, unless they are on your dissertation committee, is ever going to read it,
2. God forbid you might want to actually look at RELATIONSHIPS among things, say, parental supervision, communication and positive views of parents. What are you going to do now? Look at how each of the 42 variables relates to the other 42 variables? That's 1,764 variables, if you are keeping track. And don't you DARE run a correlation matrix and just interpret the 88 that are significant.
3. Individual variables are notoriously unreliable.

Let's talk about reliability, variance and validity, which you already know quite a bit about just by living. Answer these questions:
You have studied for a final exam in Biology 101. There is one question, "What is the relationship between respiration and photosynthesis?" Is this a fair test?

Your child is in fifth grade. Her weekly spelling test consists of one word. Is this a fair test?

Okay, these aren't trick questions ... The plain fact is that people are complicated and whether it is their knowledge of biology, how well they can spell or their relationship with their mom, you aren't going to get as much information from one single question as from several put together. This is why spelling tests aren't just one word. People VARY. They vary a lot. If you give only one question then all of your students fall into one of two groups – they spelled it correctly, 100% or they spelled it incorrectly 0%. That is almost certainly not valid, surely some of your students are better spellers than others and they don't fall into just two groups. Same with any number of other characteristics – some people wouldn't cross the street to piss on their parents if they were on fire and others have practically never left the womb at age 14. Then there is everybody in between.

Every individual item has a "true variance" aspect, how much you know about biology or spelling, how much of a loving family you are, and the statistician's favorite whine – totally random, that is *error variance*. Maybe you know a

lot about biology but you just happened to miss that day, skip that chapter on photosynthesis. It's error in the sense that it doesn't really represent the underlying construct (idea) you are trying to measure. Perhaps your family doesn't ever have dinner together because Dad works the day shift and Mom works the night shift ….

So, other things being equal, the more questions you add together, the better. A spelling test with ten words is going to be more reliable, have more variance and be more accurate (valid) than a test with only one word. The same is true of a biology test or a measure of family functioning.

That other things being equal qualifier is important. If you add up the question on biology, the one on spelling and the one on your mom, you don't have a more reliable or valid test of anything, even if you do have more variance.

SO … you should have done this in the first place, but if you didn't, better late than never. Sort your items into groups that might be related. In my example, I have parental supervision, decision-making and discussion. That just happens to be 42 questions. I may revisit this decision later, but that's what I'm going to use for now.

NOW do you get to do a factor analysis? No.

NEXT you want to make sure your data are accurate. For reasons totally behind my comprehension, many people use codes for missing data, like 9. So, you have a bunch of people who have a score of 9 on a 1 to 5 scale and it completely throws off your results. Look at the descriptive statistics for your data and at a bare minimum see that you don't have any variables out of range like that. This is just a reminder to always check your data before doing any even barely sophisticated statistics.
You're set.

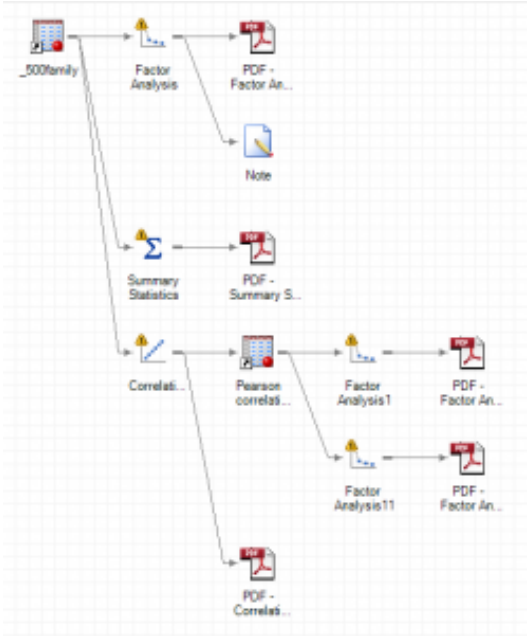
You understand WHY you are doing a factor analysis - because you want to combine these many, many questions into a few reliable, valid scales that have some decent variance.

You understand WHAT you are doing a factor analysis with – you have selected out the appropriate questions you want to use.
You have checked to see that your data at least don't totally suck.

Well, congratulations, you are now all ready for "Mama AnnMaria's Point-y Click-y Guide to Factor Analysis".

AN OVERVIEW OF A FACTOR ANALYSIS PROJECT USING SAS ENTERPRISE GUIDE

So, if you were paying attention, we just figured out WHY to do a factor analysis, today's post is about how. I'm using SAS Enterprise Guide because I had it open on my computer.
Here is what the completed project looks like:



Here is what I did, reading from the top:

1. I opened a data set.

2. I ran a factor analysis and looked at it. When I looked at it, I saw that over 120 of the records were missing out of less than 500 people. I made a note of this – literally.
Thing to know: the default for SAS is to delete a record if it is missing ANY of the variables.

3. I ran summary statistics to see if maybe there was one that 200 people were missing, say it was about how much input parents have into your job choices and most of the kids did not work. If that had been the case, I could have just dropped that one variable. It wasn't. So…

4. I ran correlations of all the variables and then

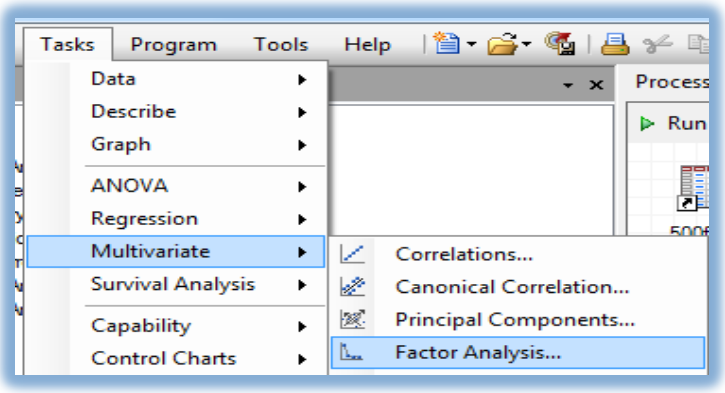
5. I factor analyzed the correlation matrix (WAY easier than it sounds!) After I took a look at the results from this analysis, I thought I could do better, so ...

6. I re-analyzed the data requesting only three factors.
With the overview out of the way, let's take a look at each part.

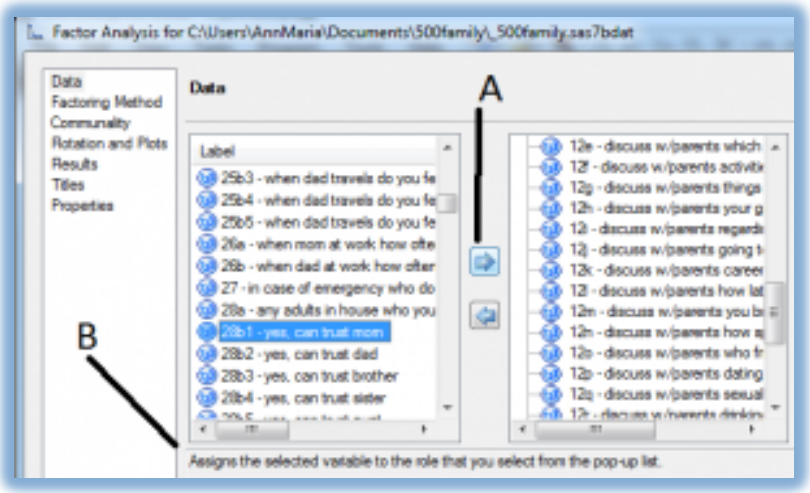
FACTOR ANALYSIS IN SIX EASY STEPS

Step 1. Open the data set is a piece of cake, go to File > Open > Data
Select the data set you want, just like you open a file in Microsoft Word or anything else.

Step 2. To do the Factor Analysis, click TASKS then MULTIVARIATE and then select FACTOR ANALYSIS

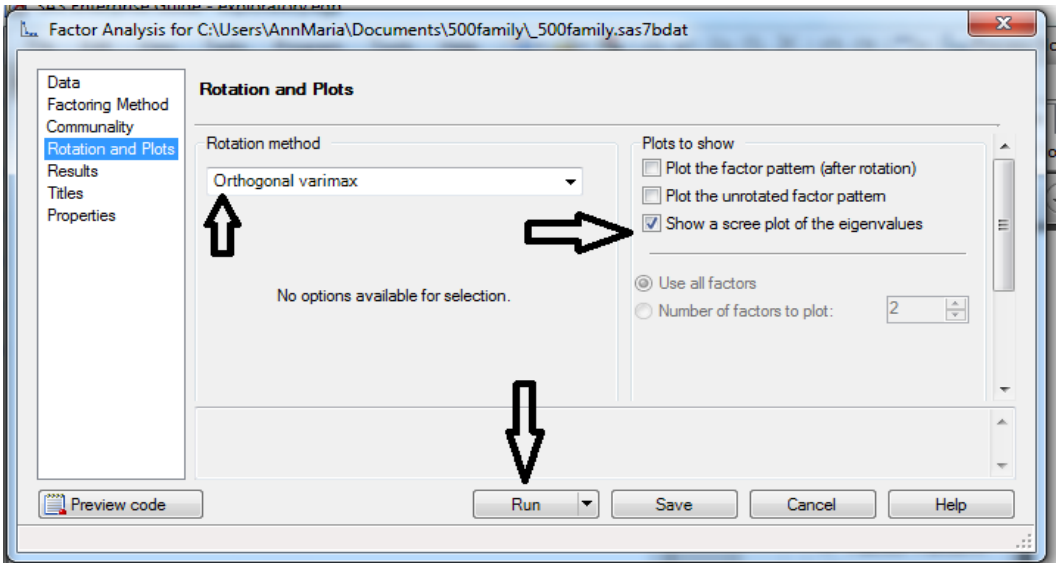


A window will pop up where you select the variables you want to use in the analysis



Click on a variable and then click the arrow, which I have so helpfully labeled as "A". Notice that SAS Enterprise Guide in the box I have equally helpfully labeled "B" often gives you tips on what you are supposed to do in a given

situation. You're welcome. You can hold down the shift key, and select a bunch of variables at once, too.



You can leave most of the defaults but I would strongly suggest that you change two of them under ROTATION AND PLOTS. Generally, you'll find a rotated factor pattern easier to interpret. I usually start with ORTHOGONAL VARIMAX rotation, which assumes that your factors are unrelated. I always want a scree plot, so I check that. Then, click RUN.

When you get your results, do NOT look at your results first. Be smarter than most people and look at your log. To do that you click on the tab that says LOG

When you do, you see this:

WARNING: 123 OF 465 OBSERVATIONS IN DATA SET WORK.SORTTEMTABLESORTED OMITTED DUE TO MISSING VALUES.
If we didn't have a lot of people missing data, we could skip the next few steps, but hey, that's life. One of my big gripes about many statistics courses and textbooks is they pretend that data is always just pristine and perfect. There are very few times in real life that your data are like that, and this is not one of them.

So …. before going any further, look at the descriptive statistics for the data. Normally I look at this before any other analyses to make sure the data are not out of range, there aren't people who show an age of 999 or who scored 99 on a scale of 1 to 10. There aren't variables that were skipped by 90% of the sample. I did that with these data but since now I am missing over one-fourth of the sample, I decide to look again.

Step 3. To get descriptive statistics using SAS Enterprise Guide, go to TASKS > DESCRIBE > SUMMARY STATISTICS

A window will pop up and just as you did above, select the variables you want to analyze. When I look at the results, I can see that the data are fine. The variables are on a 0 (=Never) to 3 (=often) scale and that all looks right. The sample size is 431, 428, 429, 415. In other words, for each question, a few people overlooked it or skipped it, but if you add all of those people who missed one here or there together it comes out to 123 people.

How (and why) to factor analyze the correlation matrix.

A factor analysis is a look at which items on a questionnaire are related. We hope to find a group of items that are related to each other and then put them into a scale of say, parental supervision. What else looks at whether a bunch of items are related? Why, a correlation matrix.

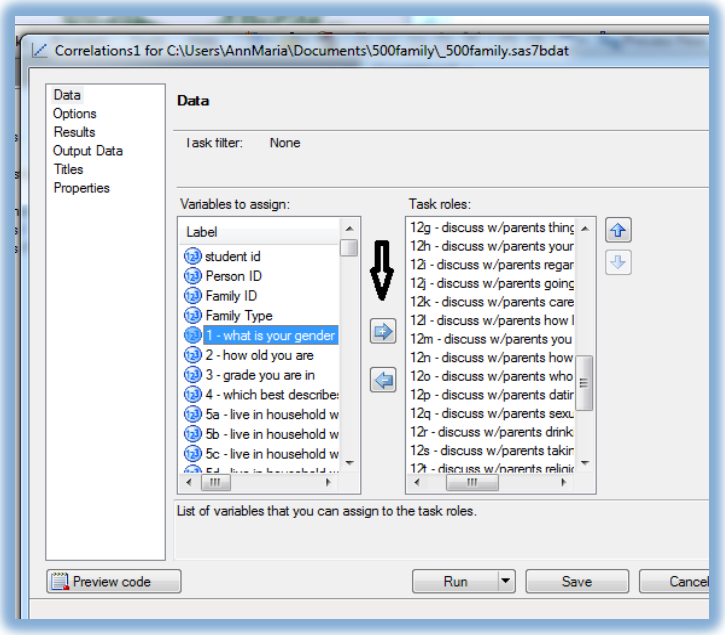
Factor analyzing a correlation matrix is SO easy (I am not making this up) When I first went to graduate school in the 1970s (yes, I'm old, what of it?), if one were to comment casually,
"And then I factor-analyzed the correlation matrix to solve that problem."

Everyone would say, "Ooooh."

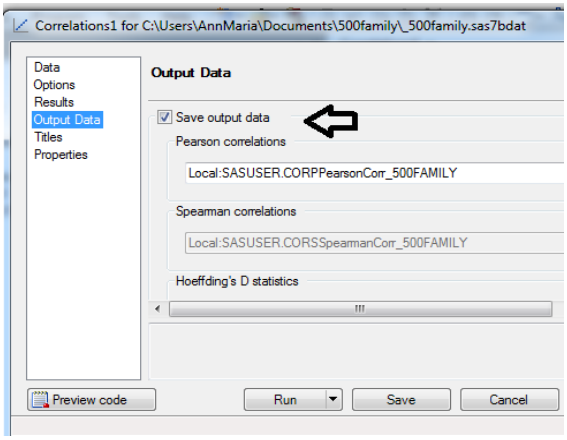
Back in those days before statistical calculators, when computers only came in big (really big) blue boxes and SAS Enterprise Guide probably was still compiled on punched cards and your department had to pay for computer time, you did it **by hand** if you were a broke graduate student. Yes, I mean literally with a pencil in your hand. First, you computed the correlations between each of the items and then you applied some equations you can find in any detailed book on factor analysis, which I had forgotten and then looked up again in the documentation (SAS Institute, Inc, 2011). It doesn't matter anyway because no one does it that way any more.

Step 4: From the TASKS menu, pull down to MULTIVARIATE and then CORRELATIONS.

Next, click on the variables you want to use in your analysis and click the blue arrow in between panes of the window to select them as my analysis variables. You can also shift-click to select a bunch at once. Don't click RUN yet!



You need to output the correlation matrix as a SAS Enterprise Guide data set of type= CORR. Fortunately, that is super-duper easy. You just click in the far left pane on the option that says OUTPUT DATA. Then click in the box next to SAVE OUTPUT DATA. Now you can click RUN.



Now that you have a correlation matrix, you can go ahead factor analyze it like you did in Step 2.

Note that "The data set created by the CORR procedure is automatically given the TYPE=CORR data set option, so *you do not have to specify TYPE=CORR.*"

So, now that you have your data set that was just created by the CORR procedure to use as the input data set, you just click on it, to select it, then

Step 5. From the top menu, select TASKS > MULTIVARIATE > FACTOR ANALYSIS like before and there you have it.

But …. What exactly do you have?

WHAT EXACTLY DOES ALL OF THIS FACTOR ANALYSIS CRAP MEAN, ANYWAY?

My doctoral advisor, the late, great Dr. Eyman, used to tell me that my psychometric theory lectures were, "A light treatment of a very serious subject. " Well, with all due respect to a truly wonderful mentor, I still have to state unequivocally that the majority of students when looking at a factor analysis for the first (or second or third) time are thinking more like me.

The questions to answer are:

- What exactly is a factor anyway?
- How many factors are present in these data?
- What does each factor that you extracted represent?

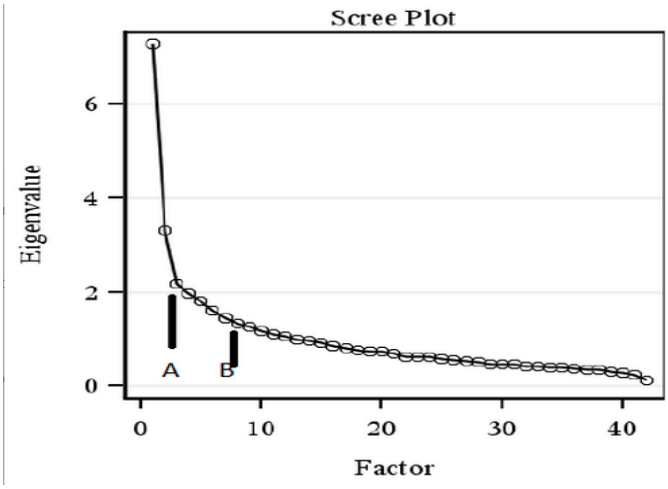
Conceptually, a factor is some underlying trait that is measured indirectly by the items you measured directly. For example, I want to measure a factor of "mathematical aptitude". So, I ask a bunch of questions like, "What is 7 x 6?" and "If two trains left the station at the same time, going 100 miles an hour in opposite directions, how far apart would they be 45 minutes later?" I'm really not that interested in your ability to answer that specific question about trains.

Factor analysis is also referred to as a 'dimension reduction technique'. It's much simpler to understand a relationship between, say college GPA and two factors of quantitative aptitude and verbal aptitude than to explain the correlations among 120 separate questions and college GPA. The measures could be anything – test scores, individual items on a test, measurements of various dimensions like height or weight, agricultural measures like yield of a rice field or economic ones like family income. You're factor analyzing a correlation matrix of these measures (if your input data set was not a correlation matrix, it's going to be transformed into one before it's analyzed). Correlations are standardized to have a variance of 1.

One thing you want to look at is the eigenvalues. An eigenvalue is the amount of variance in the individual measures explained by the factor. (If you don't believe me, square the loadings in the factor pattern and add them up. The total is the eigenvalue. Prediction: At least one person who reads this will do exactly that and be surprised that I am right. Contrary to appearances, I do not make this all up.) So if the eigenvalue is 1.0 it has explained exactly as much variance as a single item. What good is that? It would take you 42 factors with an eigenvalue of 1.0 to explain all of the variance in a set of 42 measures. You're not reducing the dimensions any. For that reason, a common criterion for deciding the number of factors is "Minimum eigenvalue greater than 1."

The problem is, and it has been documented many times over, this criterion, although it is the default for many software packages, tends to give you too many factors. I prefer two other methods. My favorite is the parallel analysis criterion, which does many iterations of analysis of a data set of random numbers. The idea is you should get factors that explain more than if you analyzed random data. There is a useful SAS macro for doing that.

Or … you can just look at a scree plot, which, although not quite as accurate involves no more effort than staring. Here is my scree plot from the 42 variables I analyzed from the 500 family study. As every good statistician (and Merriam-Webster) knows, scree is " an accumulation of loose stones or rocky debris lying on a slope or at the base of a hill or cliff". The challenge is to distinguish which factors should be retained and which are just showing small random relationships among variables, like the bits of rubble.



Clearly, we want to keep our first factor, with an eigenvalue of 7.3. Our second, with an eigenvalue of 3.3 looks like a keeper as well. So, do we take the third factor with an eigenvalue of 2.2 or do we say that is just part of the scree-type random correlations? I'm saying we keep it. Were you hoping for something more scientific? Well, I guess you're disappointed, then.

By the way, if we used the minimum eigenvalue of 1 criterion that would give us 12 factors which is just ridiculous. Liau et al. (2011) in a very serious paper for SAS Global Forum suggest not having less than 50% of the variance explained. That would mean your eigenvalues you keep add up to 21 at least, and not the 12.8 we have here (7.3 + 3.3 +2.2). To do that, instead of cutting the factors at our plot at 3, which I have so helpfully labeled Point A, we would instead cut it at Point B.
What we are doing now is an exploratory factor analysis so I am going to do this:

Step 6: (Remember Steps 1 through 5?) Based on my scree plot request a 3-factor solution.

- Inspect the factor pattern and see if that makes sense to me based on expertise in the content area that I am going to pretend to have. (Actually, if you're familiar with Baumrind's work, it is looking a bit like the control / warmth factors that she postulated so I am not completely pulling this out of my — um, head.)
- Run the parallel analysis macro and see the number of factors recommended by that (not covered in this paper).

THE FACTS OF FACTOR PATTERNS

Previously, I discussed how to go about rotating the factors. Now we're going to interpret the rotated factor pattern, shown on the following page. Let me recap, briefly. Agresti and Finlay (p.532) put it way better than me when they said: "Factor analysis is a multivariate statistical technique used for …
1. Revealing patterns of interrelationships among variables

2. Detecting clusters of variables, each of which contains variables that are strongly intercorrelated.

3. Reducing a large number of variables to a smaller number of statistically uncorrelated variables, the factors of factor analysis."

All of which is well and good but once you have your factors, what do they mean? How do you interpret them?
Important point one: The correlation of a variable with a factor is called the loading.
Important point two: To ease interpretation we'd really like to have "simple structure", that is, where variables load close to 1.0 on one factor and close to zero on the others. I mean, really, if you think about it, if your items load equally on all factors it's going to be pretty hard to interpret.

Let's take a look at my example from the 500 Family Study, which you have probably forgotten already. To make it easier to interpret, I copied the factor pattern output into a spreadsheet and sorted by the loadings on the first, second and third factor. You can see that almost all of the items relating to discussion loaded on the first factor. So, I could say that factor 1 is "Communication with parents". The second factor seems to be mostly about rules, punishment and placing limits, such as punishments or reward for grades, curfew and time out with friends. The discussion questions that load more on this factor than the first are on discussion of breaking rules and discussion of curfew. The third factor is all of the items related to decision-making, with the exception of family purchases, which didn't really load on any of the three factors.

Item	Factor 1	Factor 2	Factor 3
discuss going to college	0.63305	-0.26221	0.172
discuss career plans	0.61471	-0.07078	-0.06
discuss becoming independent	0.6042	0.00678	-0.20583
discuss standing up for oneself	0.5164	0.13633	-0.3052
discuss loving people	0.565	0.08553	-0.26476
discuss regarding ACT, SAT	0.54084	0.0266	0.05
discuss drinking alcohol	0.54002	0.134	-0.082
discuss sexual relations	0.5376	0.376	0.02207
discuss your free time	0.53685	0.184	0.0713
discuss who friends are	0.53642	0.1388	-0.0072
discuss dating scene	0.5306	0.257	0.073
discuss which courses take at school	0.56	0.7	-0.0773
discuss taking drugs	0.50413	0.1841	-0.22
discuss your grades	0.46401	0.274	-0.472
discuss activities which interest you	0.45524	-0.001	-0.07346
discuss how spend money	0.4261	0.42005	-0.247
discuss things studied in class	0.4034	0.0185	-0.063
discuss where you are at nights	0.3418	0.24716	0.2723
discuss religion/faith/spirituality	0.3815	0.06626	-0.2854
discuss how late you stay out	0.37413	0.43231	0.22845
discuss your location in afternoons	0.36016	0.3358	0.1381
discuss you breaking rules	0.374	0.54808	-0.06003
discuss time watching tv	0.27282	0.35265	-0.15651
how often parents check your homework	0.2218	0.41444	-0.22286
who decides to staying out late	0.17882	-0.3734	0.22467
what is the weekend curfew	0.626	-0.6004	0.216
how often parents ask you to call when with frien	0.55	0.42204	-0.01468
how often parents help you with homework	0.3	0.35416	-0.241
who decides on family purchases	0.106	-0.01501	0.2142
how often parents check up on you with friends	0.476	0.568	-0.0001
how often parents punish/reward for grades	0.0285	0.4442	-0.07242
how often parents limit tv, video/computer game	0.08665	0.4824	-0.222
how often parents limit time out with friends	0.06884	0.61601	0.053
how often parents require you to do chores	0.03451	0.3687	0.03703
what is school night curfew	0.02033	-0.53548	0.22426
who decides on your classes at school	-0.031	-0.07205	0.6314
who decides regarding your friends	-0.04366	-0.28464	0.3627
who decides regarding dating	-0.05336	-0.2645	0.55408
who decides whether you can have job	-0.0555	-0.0417	0.6074
who decides if you go to college	-0.077	-0.00341	0.5614
who decides regarding money spent	-0.14205	-0.275	0.35588
who decides which college	-0.1702	0.0178	0.51488

Notice a few things— Just like correlations, loadings can be positive or negative. How late your curfew is loads negatively on the Rules Factor. That is, families that have stricter rules have an earlier curfew. How often parents limit time out with your friends loads positively on the Rules Factor. Although it's not ideal, variables can load on more than one factor. As noted, the discussion of breaking rules item loads both on the Communication Factor and the Rules Factor. Variables cannot load on any factor at all, like the decision on family purchases. My guess is that most parents decide most purchases without consulting their adolescent children.

CONCLUSION

The really useful result of factor analysis is that it allows you to take your 42 items, discard one as not really fitting and distill the others down into three factors. Instead of using 41 individual items to predict your outcome of interest, say delinquent behavior, you can use three. It's almost certain that those three factors will be far more reliable than any individual item, and your results will be far easier to explain as well, say, "Students who have more communication with their parents, moderate rules and moderate input on decision-making have the lowest rate of delinquent behavior and highest academic achievement."

Not sure if that is true or not but with these factors we are now in a good position to test that. I just need a couple more measures, of delinquent behavior and academic achievement, and I can test my hypotheses. I expect there will be a linear relationship with communication (negative for delinquency and positive for academics) and a curvilinear relationship with the other two measures (inverse for delinquency). I guess that will be my next thing to do when I have some spare time. Or, you can wander on over to ICPSR.org and download the 500 Family Study data (Schneider & Waite, 1998-2000) yourself.

REFERENCES

- Agresti, A. & Finlay, B. (2009). Statistical methods for the social sciences (4th Ed.) Upper Saddle River, NJ: Pearson.
- CPSR

- Liau, A., Tan, T. K. & Khoo, A. (2011). Scale Measurement: Comparing Factor Analysis and Variable Clustering. http://support.sas.com/resources/papers/proceedings11/352-2011.pdf

- Merriam-Webster (2013). Merriam-Webster Dictionary on-line. http://www.merriam-webster.com/dictionary/scree

- SAS Institute Inc. (2011). SAS/STAT 9.3 User's Guide, Cary, NC: SAS Institute Inc.

- Schneider, Barbara, and Linda J Waite. The 500 Family Study [1998-2000: United States]. ICPSR04549-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor] http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/4549

ACKNOWLEDGEMENTS

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

AnnMaria De Mars
The Julia Group/ 7 Generation Games
2111 7th St. #8
Santa Monica, CA 90405
(310) 717-9089
annmaria@thejuliagroup.com
http://www.thejuliagroup.com www.7generationgames.com