the WINNOWER

SOCIAL SCIENCES

Examining the S factor in Mexican states

EMIL O. W. KIRKEGAARD

CORRESPONDENCE:
emil@emilkirkegaard.dk

DATE RECEIVED:
June 10, 2015

Abstract

Two datasets of socioeconomic data was obtained from different sources. Both were factor analyzed and revealed a general factor (S factor). These factors were highly correlated with each other (.79 to .95), HDI (.68 to .93) and with cognitive ability (PISA; .70 to .78). The federal district was a strong outlier and excluding it improved results.

Method of correlated vectors was strongly positive for all 4 analyses (r's .78 to .92 with reversing).

Introduction

In a number of recent articles (Kirkegaard 2015a,b,c,d,e), I have analyzed within-country regional data to examine the general socioeconomic factor, if it exists in the dataset (for the origin of the term, see e.g. Kirkegaard 2014). This work was inspired by Lynn (2010) whose datasets I have also reanalyzed. While doing work on another project (Fuerst and Kirkegaard, 2015*), I needed an S factor for Mexican states, if such exists. Since I was not aware of any prior analysis of this country in this fashion, I decided to do it myself.

The first problem was obtaining data for the analysis. For this, one needs a number of diverse indicators that measure important economic and social matters for each Mexican state. Mexico has 31 states and a federal district, so one can use a decent number of indicators to examine the S factor. Mexico is a Spanish speaking country and English comprehension is fairly poor. According to Wikipedia, only 13% of people speak English there. Compare with 86% for Denmark, 64% for Germany and 35% for Egypt.

S factor analysis 1 – Wikipedian data

Data source and treatment

Unlike for the previous countries, I could not easily find good data available in English. As a substitute, I used data from Wikipedia:

- en.wikipedia.org/wiki/List_of_Mexican_states_by_unemployment
- en.wikipedia.org/wiki/List_of_Mexican_states_by_fertility_rate
- en.wikipedia.org/wiki/List_of_Mexican_states_by_homicides
- en.wikipedia.org/wiki/List_of_Mexican_states_by_infant_mortality
- en.wikipedia.org/wiki/List_of_Mexican_states_by_life_expectancy
- en.wikipedia.org/wiki/List_of_Mexican_states_by_literacy_rate
- en.wikipedia.org/wiki/List_of_Mexican_states_by_GDP
- en.wikipedia.org/wiki/List_of_Mexican_states_by_poverty_rate

- en.wikipedia.org/wiki/List_of_Mexican_states_by_Human_Development_Index
- en.wikipedia.org/wiki/List_of_Mexican_states_by_population
- en.wikipedia.org/wiki/Ranked_list_of_Mexican_states

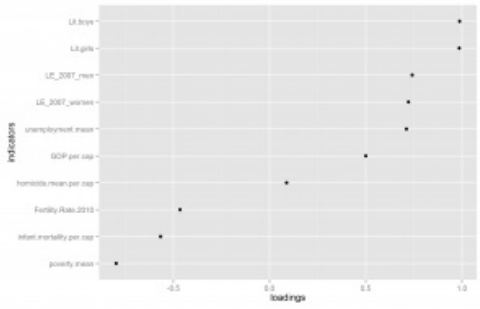
These come from various years, are sometimes not given per person, and often have no useful source given. So they are of unknown veracity, but they are probably fine for a first look. The HDI is best thought of as a proxy for the S factor, so we can use it to examine construct validity.

Some variables had data for multiple time-points and they were averaged.

Some data was given in raw numbers. I calculated per capita versions of them using the population data also given.

Results

The variables above minus HDI and population size were factor analyzed using minimum residuals to extract 1 factor. The loadings plot is shown below.



The literacy variables had a near perfect loading on S (.99). Unemployment unexpectedly loaded positively and so did homicides per capita altho only slightly. This could be because unemployment benefits are only in existence in the higher S states such that going unemployed would mean starvation. The homicide loading is possibly due to the drug war in the country.

Analysis 2 – Data obtained from INEG

Data source and treatment

Since the results based on Wikipedia data was dubious, I searched further for more data. I found it on the Spanish-language statistical database, *Instituto Nacional De Estadística Y Geografía*, which however had the option of showing poorly done English translations. This is not optimal as there are many translation errors which may result in choosing the wrong variable for further analysis. If any Spanish-speaker reads this, I would be happy if they would go over my chosen variables and confirm that they are correct. I ended up with the following variables:

1. Cost of crime against individuals and households
2. Cost of crime on economic units
3. Annual percentage change of GDP at 2008 prices
4. Crime prevalence rate per 10,000 economic units
5. Crime prevalence rate per hundred thousand inhabitants aged 18 years and over, by state
6. Dark figure of crime on economic units
7. Dark figure (crimes not reported and crimes reported that were not investigated)
8. Doctors per 100 000 inhabitants
9. Economic participation of population aged 12 to 14 years
10. Economic participation of population aged 65 and over
11. Economic units.
12. Economically active population. Age 15 and older
13. Economically active population. Unemployed persons. Age 15 and older

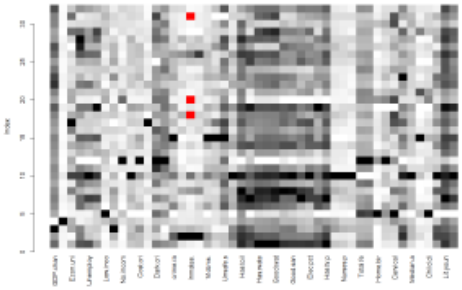
14. Electric energy users
15. Employed population by income level. Up to one minimum wage. Age 15 and older
16. Employed population by income level. More than 5 minimum wages. Age 15 and older
17. Employed population by income level. Do not receive income. Age 15 and older
18. Fertility rate of adolescents aged 15 to 19 years
19. Female mortality rate for cervical cancer
20. Global rate of fertility
21. Gross rate of women participation
22. Hospital beds per 100 thousand inhabitants
23. Inmates in state prisons at year end
24. Life expectancy at birth
25. Literacy rate of women 15 to 24 years
26. Literacy rate of men 15 to 24 years
27. Median age
28. Nurses per 100 000 inhabitants
29. Percentage of households victims of crime
30. Percentage of births at home
31. Percentage of population employed as professionals and technicians
32. Prisoners rate (per 10,000 inhabitants age 18 and over)
33. Rate of maternal mortality (deaths per 100 thousand live births)
34. Rate of inhabitants aged 18 years and over that consider their neighborhood or locality as unsafe, per hundred thousand inhabitants aged 18 years and over
35. Rate of inhabitants aged 18 years and over that consider their state as unsafe, per hundred thousand inhabitants aged 18 years and over
36. Rate sentenced to serve a sentence (per 1,000 population age 18 and over)
37. State Gross Domestic Product (GDP) at constant prices of 2008
38. Total population
39. Total mortality rate from respiratory diseases in children under 5 years
40. Total mortality rate from acute diarrheal diseases (ADD) in population under 5 years
41. Unemployment rate of men
42. Unemployment rate of women
43. Households
44. Inhabited housings with available computer
45. Inhabited housings that have toilet
46. Inhabited housings that have a refrigerator
47. Inhabited housings with available water from public net
48. Inhabited housings that have drainage
49. Inhabited housings with available electricity
50. Inhabited housings that have a washing machine
51. Inhabited housings with television
52. Percentage of housing with piped water
53. Percentage of housing with electricity
54. Proportion of population with access to improved sanitation, urban and rural
55. Proportion of population with sustainable access to improved sources of water supply, in urban and rural areas

There are were data for multiple years for most of them. I used all data from the last 10 years, approximately. For all data with multiple years, I calculated the mean value.

For data given in raw numbers, I calculated the appropriate per unit measures (per person, per economically active person (?), per household).

A matrix plot for all the S factor relevant data (e.g. not population size) is shown below. It shows missing data in red, as well as the relative difference between datapoints. Thus, cells that are

completely white or black are outliers compared to the other data.



One variable (inmates per person) has a few missing datapoints.

Multiple other variables had strong outliers. I examined these to determine if they were real or due to data error.

Inspection revealed that the GDP per person data was clearly incorrect for one state (Campeche) but I could not find the source of error. The data is the same as on the website and did not match the data on Wikipedia. I deleted it to be safe.

The GDP change outlier seems to be real (Campeche) which has negative growth. According tothis site, it is due to oil fields closing.

The rest of the outliers were hard to say something about due to the odd nature of the data ("dark crime"?), or were plausible. E.g. Mexico City (aka Federal District, the capital) was an outlier on nurses and doctors per capita, but this is presumably due to many large hospitals being located there.

Some data errors of my own were found and corrected but there is no guarantee there are not more. Compiling a large set of data like this frequently results in human errors.

Factor analysis

Since there were only 32 cases — 31 states + federal district — and 47 variables (excluding the bogus GDP per capita one), this gives problems for factor analysis. There are various recommendations, but almost none of them are met by this dataset (Zhao, 2009). To test limits, I decided to try factor analyzing all of the variables. This produced warnings:

The estimated weights for the factor scores are probably incorrect. Try a different factor extraction method.
In factor.scores, the correlation matrix is singular, an approximation is used
Warning messages:
1: In cor.smooth(R) : Matrix was not positive definite, smoothing was done
2: In cor.smooth(R) : Matrix was not positive definite, smoothing was done
3: In cor.smooth(r) : Matrix was not positive definite, smoothing was done
4: In cor.smooth(r) : Matrix was not positive definite, smoothing was done

Warnings such these do not always mean that the result is nonsense, but they often do. For that reason, I wanted to extract an S factor with a smaller number of variables. From the 47, I selected the following 21 variables as generally representative and interpretable:

1. GDP.change, #Economic
2. Unemploy.men.rate,
3. Unemploy.women.rate,
4. Low.income.peap,
5. High.income.peap,
6. Prof.tech.employ.pct,

7. crime.rate.per.adult, #crime
8. Inmates.per.pers,
9. Unsafe.neighborhood.percept.rate,
10. Has.water.net.per.hh, #material goods
11. Elec.pct,
12. Has.wash.mach.per.hh,
13. Doctors.per.pers, #Health
14. Nurses.per.pers,
15. Hospital.beds.per.pers,
16. Total.fertility,
17. Home.births.pct,
18. Maternal.death.rate,
19. Life.expect,
20. Women.participation, #Gender equality
21. Lit.young.women #education

Note that peap = per economically active person, hh = household.

The selection was made by my judgment call and others may choose different variables.

Automatic reduction of dataset

As a robustness check and evidence against a possible claim that I picked the variables such as to get an S factor that most suited my prior beliefs, I decided to find an automatic method of selecting a subset of variables for factor analysis. I noticed that in the original dataset, some variables overlapped near perfectly. This would mean that whatever they measure, it would get measured twice or more when extracting a factor. Highly correlated variables can also create nonsense solutions, especially when extracting more than 1 factor.

Another piece of insight comes from the fact that for cognitive data, general factors extracted from a less broad selection of subtests are worse measures of general cognitive ability than those from broader selections (Johnson et al, 2008).

Lastly, subtests from different domains tend to be less correlated than those from the same domain (hence the existence of group factors).

Combining all this, it seems a decent idea that to reduce a dataset by 1 variable, one should calculate all the intercorrelations and find the highest one. Then one should remove one of the variables responsible for it. One can do this repeatedly to remove more than 1 variable from a dataset. Concerning the question of which of the two variables to remove, I can think of three ways: always removing the first, always the second, choosing at random. I implemented all three settings and chose the second as the default. This is because in many datasets the first of a set of highly correlated variables is usually the 'primary one', E.g. unemployment, unemployment men, unemployment women. The algorithm also outputs step-by-step information concerning which variables was removed and what their correlation was.

Having written the R code for the algorithm, I ran it on the Mexican dataset. I wanted to obtain a solution using the largest possible number of variables without getting a warning from the factor extraction function. So I first removed 1 variable, and then ran the factor analysis. When I received an error, I removed another, and so on. After having removed 20 variables, I no longer received an error. This left the analysis with 27 variables, or 6 more than my chosen selection. The output from the reduction algorithm was:

```
> s3 = remove.redundant(s, 20)
[1] "Dropping variable number 1"
[1] "Most correlated vars are Good.water.prop and Piped.water.pct r=0.997"
```

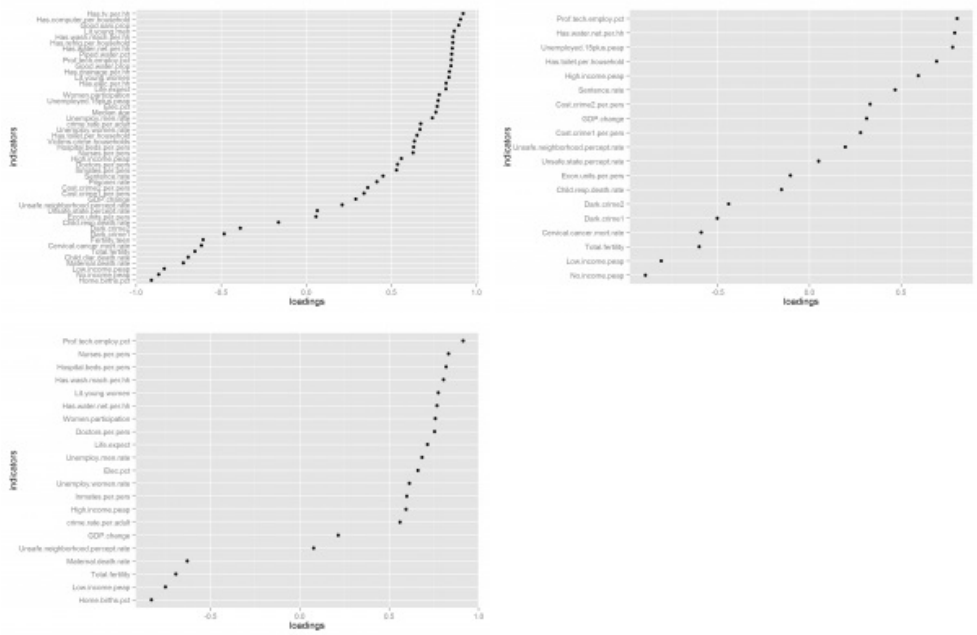
[1] "Dropping variable number 2"
[1] "Most correlated vars are Piped.water.pct and Has.water.net.per.hh r=0.996"
[1] "Dropping variable number 3"
[1] "Most correlated vars are Fertility.teen and Total.fertility r=0.99"
[1] "Dropping variable number 4"
[1] "Most correlated vars are Good.sani.prop and Has.drainage.per.hh r=0.984"
[1] "Dropping variable number 5"
[1] "Most correlated vars are Victims.crime.households and crime.rate.per.adult r=0.97"
[1] "Dropping variable number 6"
[1] "Most correlated vars are Nurses.per.pers and Doctors.per.pers r=0.962"
[1] "Dropping variable number 7"
[1] "Most correlated vars are Lit.young.men and Lit.young.women r=0.938"
[1] "Dropping variable number 8"
[1] "Most correlated vars are Elec.pct and Has.elec.per.hh r=0.938"
[1] "Dropping variable number 9"
[1] "Most correlated vars are Has.wash.mach.per.hh and Has.refrig.per.household r=0.926"
[1] "Dropping variable number 10"
[1] "Most correlated vars are Prisoner.rate and Inmates.per.pers r=0.901"
[1] "Dropping variable number 11"
[1] "Most correlated vars are Unemploy.women.rate and Unemploy.men.rate r=0.888"
[1] "Dropping variable number 12"
[1] "Most correlated vars are Women.participation and Has.computer.per.household r=0.877"
[1] "Dropping variable number 13"
[1] "Most correlated vars are Hospital.beds.per.pers and Doctors.per.pers r=0.87"
[1] "Dropping variable number 14"
[1] "Most correlated vars are Has.computer.per.household and Prof.tech.employ.pct r=0.868"
[1] "Dropping variable number 15"
[1] "Most correlated vars are Unemploy.men.rate and Unemployed.15plus.peap r=0.866"
[1] "Dropping variable number 16"
[1] "Most correlated vars are Has.tv.per.hh and Has.elec.per.hh r=0.864"
[1] "Dropping variable number 17"
[1] "Most correlated vars are Has.elec.per.hh and Has.drainage.per.hh r=0.851"
[1] "Dropping variable number 18"
[1] "Most correlated vars are Median.age and Prof.tech.employ.pct r=0.846"
[1] "Dropping variable number 19"
[1] "Most correlated vars are Home.births.pct and Low.income.peap r=0.806"
[1] "Dropping variable number 20"
[1] "Most correlated vars are Life.expect and Has.water.net.per.hh r=0.796

In my opinion the output shows that the function works. In most cases, the pair of variables found was either a (near-)double measure e.g. percent of population with electricity and percent of households with electricity, or closely related e.g. literacy in men and women. Sometimes however, the pair did not seem to be closely related, e.g. women's participation and percent of households with a computer.

Since this dataset selected the variable with missing data, I used the irmi() function from the VIM package to impute the missing data (Templ et al, 2014).

Factor loadings: stability

The factor loading plots are shown below.



Each analysis relied upon a unique but overlapping selection of variables. Thus, it is possible to correlate the loadings of the overlapping parts for each analysis. This is a measure of loading stability in different factor analytic environments, as also done by Ree and Earles (1993) for general cognitive ability factor (g factor). The correlations were .98, 1.00, .98 (n's 21, 27, 12), showing very high stability across datasets. Note that it was not possible to use the loadings from the Wikipedian data factor analysis because the variables were not strictly speaking overlapping.

Factor loadings: interpretation

Examining the factor loadings reveals some things of interest. Generally for all analyses, whatever that is generally considered good loads positively, and whatever considered bad loads negatively.

Unemployment (together, men, women) has positive loadings, whereas it 'should' have negative loadings. This is perhaps because the lower S factor states have more dysfunctional or no social security nets such that not working means starvation, and that this keeps people from not working. This is merely a conjecture because I don't know much about Mexico. Hopefully someone more knowledgeable than me will read this and have a better answer.

Crime variables (crime rate, victimization, inmates/prisoner per capita, sentencing rate) load positively whereas it should load negatively. This pattern has been found before, see Kirkegaard (2015e) for a review of S factor studies and crime variables.

Factor scores

Next I correlated the factor scores from all 4 analysis with each other as well as HDI and cognitive ability as measured by PISA tests (the cognitive data is from Fuerst and Kirkegaard, 2015*; the HDI data from Wikipedia). The correlation matrix is shown below.

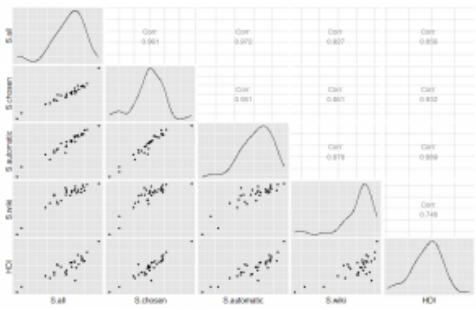
"regression" method	**S.all**	**S.chosen**	**S.automatic**	**S.wiki**	**HDI**	**Cognitive ability**
S.all	1.00	-0.08	-0.02	0.08	-0.17	-0.12
S.chosen	-0.08	1.00	0.93	0.84	0.93	0.65
S.automatic	-0.02	0.93	1.00	0.89	0.88	0.74
S.wiki	0.08	0.84	0.89	1.00	0.76	0.78
HDI	-0.17	0.93	0.88	0.76	1.00	0.53
Cognitive ability	-0.12	0.65	0.74	0.78	0.53	1.00

Strangely, despite the similar factor loadings, the factor scores from the factor extracted from all the variables had about no relation to the others. This probably indicates that the factor scoring method could not handle this type of odd case. The default scoring method for the factor analysis is "regression", but there are a few others. Bartlett's method yielded results for S.all that fit with the other factors, while none of the others did. See the psych package documentation for details (Revelle, 2015). I changed the extraction method for all the other analyses to Bartlett's to remove method specific variance. The new correlation table is shown below:

Bartlett's method	**S.all**	**S.chosen**	**S.automatic**	**S.wiki**	**HDI.mean**	**Cognitive ability**
S.all	1.00	0.79	0.88	0.88	0.68	0.74
S.chosen	0.79	1.00	0.95	0.87	0.93	0.70
S.automatic	0.88	0.95	1.00	0.88	0.89	0.74
S.wiki	0.88	0.87	0.88	1.00	0.75	0.78
HDI.mean	0.68	0.93	0.89	0.75	1.00	0.53
Cognitive ability	0.74	0.70	0.74	0.78	0.53	1.00

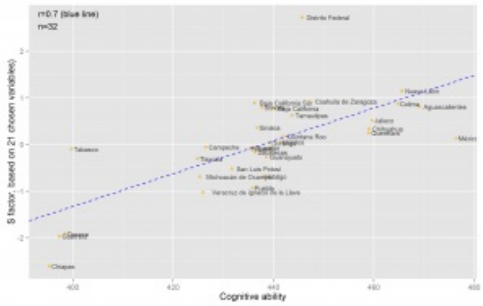
Intriguingly, now all the correlations are stronger. Perhaps Bartlett's method is better for handling this type of extraction involving general factors from datasets with low case to variable ratios. It certainly deserves empirical investigation, including reanalysis of prior datasets. I reran the earlier parts of this paper with the Bartlett method. It did not substantially change results. The correlations between loadings across analysis increased a bit (to .98, 1.00, .99).

One possibility however is that the stronger results is just due to Bartlett's method creating outliers that happen to lie on the regression line. This did not seem to be the case, see scatterplots below.



S factor scores and cognitive ability

The next question is to what degree the within country differences in Mexico can be explained by cognitive ability. The correlations are in the above table as well, they are in the region .70 to .78 for the various S factors. In other words, fairly high. One could plot all of them vs. cognitive ability, but that would give us 4 plots. Instead, I plot only the S factor from my chosen variables since this has the highest correlation with HDI and thus the best claim for construct validity. It is also the most conservative option because of the 4 S factors, it has the lowest correlation with cognitive ability. The plot is shown below:



We see that the federal district is a strong outlier, just like in the study with US states and Washington DC (Kirkegaard, 2015c). One should then remove it and rerun all the analyses. This includes the S factor extractions because the presence of a strong 'mixed case' (to be explained further in a future publication) affects the S factor extracted (see again, Kirkegaard, 2015c).

Analyses without Federal District

I reran all the analyses without the federal district. Generally, this did not change much with regards to loadings. Crime and unemployment still had positive loadings.

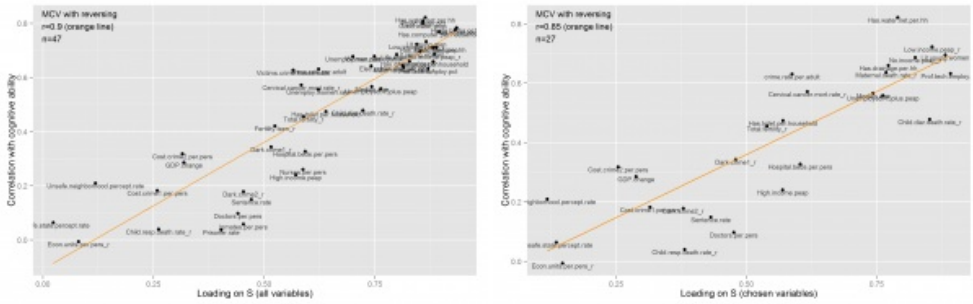
The loadings correlations across analyses increased to 1.00, 1.00, 1.00.

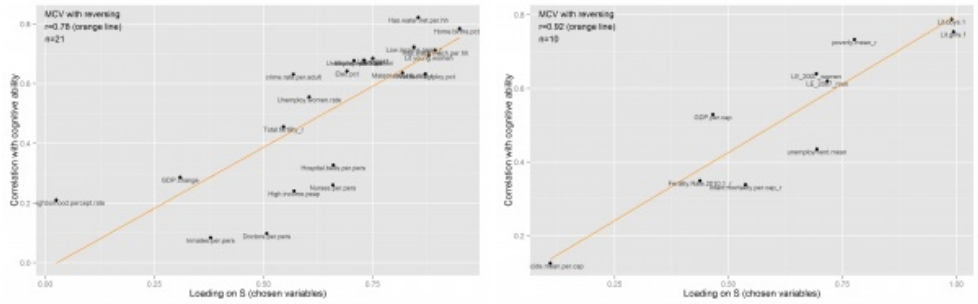
	S.all	S.chosen	S.automatic	S.wiki	HDI mean	Cognitive ability
S.all	1.00	0.99	0.98	0.93	0.85	0.78
S.chosen	0.99	1.00	0.98	0.94	0.88	0.80
S.automatic	0.98	0.98	1.00	0.90	0.90	0.75
S.wiki	0.93	0.94	0.90	1.00	0.75	0.77
HDI mean	0.85	0.88	0.90	0.75	1.00	0.56
Cognitive ability	0.78	0.80	0.75	0.77	0.56	1.00

The factor score correlations increased meaning that the federal district outlier was a source of discrepancy between the extraction methods. This can be seen in the scatterplots above in that there is noticeable variation in how far from the rest the federal district lies. After this is resolved, the S factors from the INEG dataset are in near-perfect agreement (.99, .98, .98) while the one from Wikipedia data is less so but still respectable (.93, .94, .90). Correlations with cognitive ability also improved a bit.

Method of correlated vectors

In line with earlier studies, I examine whether the measures that are better measures of the latent S factor are also correlated more highly with the criteria variable, cognitive ability.





The MCV results are strong: .90 .78 .85 and .92 for the analysis with all variables, chosen variables, automatically chosen variables and Wikipedian variables respectively. Note that these are for the analyses without the federal district, but they were similar with it too.

Discussion and conclusion

Generally, the present analysis reached similar findings to those before, especially with the one about US states. Cognitive ability was a very strong correlate of the S factors, especially once the federal district outlier was removed before the analysis. Further work is needed to find out why unemployment and crime variables sometimes load positively in S factor analyses with regions or states as the unit of analysis.

MCV analysis supported the idea that cognitive ability is related to the S factor, not just some non-S factor source of variance also present in the dataset.

Supplementary material

Data files, R code, figures are available at the Open Science Framework repository.

References

- Fuerst, J. and Kirkegaard, E. O. W. (2015*). Admixture in the Americas part 2: Regional and National admixture. (Publication venue undecided.)
- Johnson, W., Nijenhuis, J. T., & Bouchard Jr, T. J. (2008). Still just 1g: Consistent results from five test batteries. *Intelligence*, 36(1), 81-95.
- Kirkegaard, E. O. W. (2014). The international general socioeconomic factor: Factor analyzing international rankings. *Open Differential Psychology*.
- Kirkegaard, E. O. W. (2015a). S and G in Italian regions: Re-analysis of Lynn's data and new data. The Winnower.
- Kirkegaard, E. O. W. (2015b). Indian states: G and S factors. *The Winnower*.
- Kirkegaard, E. O. W. (2015c). Examining the S factor in US states. *The Winnower*.
- Kirkegaard, E. O. W. (2015d). The S factor in China. *The Winnower*.
- Kirkegaard, E. O. W. (2015e). The S factor in the British Isles: A reanalysis of Lynn (1979). *The Winnower*.
- Lynn, R. (2010). In Italy, north–south differences in IQ predict differences in income, education, infant mortality, stature, and literacy. Intelligence, 38(1), 93-100.
- Ree, M. J., & Earles, J. A. (1991). The stability of g across different methods of estimation. Intelligence, 15(3), 271-278.
- Revelle, W. (2015). psych: Procedures for Psychological, Psychometric, and Personality Research. CRAN
- Templ, M., Alfons A., Kowarik A., Prantner, B. (2014). VIM: Visualization and Imputation of Missing Values. CRAN
- Zhao, N. (2009). The Minimum Sample Size in Factor Analysis. Encorewiki.

* = not yet published, year is expected publication year.