the WINNOWER

SOCIAL SCIENCES

A replication of the S factor among US states using a new and larger dataset

EMIL O. W. KIRKEGAARD

CORRESPONDENCE:
emil@emilkirkegaard.dk

DATE RECEIVED:
June 25, 2015

Abstract

A dataset of 127 variables concerning socioeconomic outcomes for US states was analyzed. Of these, 81 were used in a factor analysis. The analysis revealed a general socioeconomic factor. This factor correlated .961 with one from a previous analysis of socioeconomic data for US states.

Introduction

It has repeatedly been found that desirable outcomes tend to be associated with other desirable outcomes and likewise for undesirable outcomes. When this is the case, one can extract a general factor — the general socioeconomic factor (S factor) — such that the desirable outcomes load positively and the undesirable outcomes negatively. This pattern has been found at the country level (1), within country divisions of many countries (2–10), at the city district level (11), at the level of first names (12) and at the level of country of origin groups in two countries (13,14).

A previous study have found that the pattern holds for US states too (7). However, a new and larger dataset has been found, so it is worth examining whether the pattern holds in it, and if so, how strongly correlated the extracted factor scores are between the datasets. This would function as a kind of test-retest reliability.

Data sources

The previous study (7) of the S factor among US states used a dataset of 25 variables compiled from various official statistics found at *The 2012 Statistical Abstract* website. The current study relies upon a dataset compiled by *Measure of America*, a website that visualizes social inequality. It is possible to download the datasets their maps rely upon here.

As done with earlier studies, I excluded the capital district. I also excluded the data for US as a whole since it was not a state like the other cases.

The dataset contains a total of 127 variables. However, not all of these are useful for examining the S factor:

- 4 variables are the composite indexes calculated by Measure of America. These are fairly similar to the Human Development Index scores, except that they are scaled differently.
- 6 variables concern the population sizes in percent of 6 sociological race categories: Non-Hispanic White, Latino, African American, Asian, Amerindian (Native American) and other.
- 1 variable contains the total population size for each state.
- A number of variables were not given in a form adjusted for population size e.g. per capita, percent or rate per 100k persons. These variables were excluded: Rape (total number), Homeless Population (total number), Medicare Recipients (thousands), Medicaid Recipients (thousands), Army Recruits (total), Total Military Casualties in Operations Enduring Freedom and Iraqi Freedom to April 2010, Prisoners State or Federal Jurisdiction (total number), Women in Congressional Delegation (total), Men in Congressional Delegation (total), Carcinogen Releases (pounds), Lead Releases (pounds), Dioxin Releases (grams), Superfund Sites (total), Protected Forest (acres), and Protected Farm and Ranch Land (acres).
- 1 variable was excluded due to being heavily reliant on local natural environment (presence of water and forests): Farming fishing and forestry occupations (%).
- 1 variable was excluded because most of its data was missing: State Earned Income Tax Credit (% of federal Earned Income Tax Credit).

The variables that were not given in per population format almost always had a sibling variable that was given in a suitable format and which was included in the analysis. After these exclusions, 101 variables remained for analysis.

Missing data

An analysis of missing data showed that some variables still had missing data. Because the dataset had more variables than cases, it was not possible to impute the missing data using multiple regression as commonly done in these analyses. For this reason, these variables were excluded. After this, 93 variables remained for analysis.

Duplicated, reverse-coded and highly redundant variables

An analysis of correlations among variables showed that 2 of them had duplicates (r = 1): Diabetes (% age 18 and older) and Low-Birth-Weight Infants (% of all infants). I'm not sure why this is the case.

Furthermore, 4 variables had a reverse-coded sibling (r = -1):

1. Less Than High School (%) + At Least High School Diploma (%)
2. 4th Graders Reading Below Proficiency (%) + 4th Grade National Assessment of Educational Progress in Reading (% at or above proficient)
3. Urban Population (%) + Rural Population (%)
4. Public High School Graduation Rate (%) + High School Freshmen Not Graduating After 4 Years (%).

Finally, some variables were so strongly related to other variables that keeping both would perhaps result in factor analytic errors or headily influence the resulting factor. I decided to use a threshold of |.9| as the limit. If any pair of variables correlated at this level or above, one of them was excluded. There were 6 pairs of variables like this and the first of the pair was excluded:

1. Poverty Rate (% below federal poverty threshold) + Child Poverty (% living in families below the poverty line), r = .985.
2. Poverty Rate (% below federal poverty threshold) + Children Under 6 Living in Poverty (%), r = .968.
3. Management professional and related occupations (%) + At Least Bachelor's Degree (%), r = .925.
4. Preschool Enrollment (% enrolled ages 3 and 4) + 3- and 4-year-olds Not Enrolled in Preschool (%), r = -.925.
5. Army Recruits (per 1000 youth) + Army Recruits (per 1000 youth), r = .914.
6. Graduate Degree (%) + At Least Bachelor's Degree (%), r= .910.

The army recruit variable seems to be a duplicate, but the numbers are not identical for all cases. The two preschool enrollment variables seem to be meant to be a reverse-coding of each other, but they don't correlate perfectly negatively.

After exclusion of these variables, there were 81 remaining.

Factor analysis

Next I extracted a general factor from the data. Since one previous study had found instability across extraction methods when extracting factors from datasets with more variables than cases (2), I examined the stability across all possible extraction and scoring methods, 30 in total (6 extraction methods, 5 scoring methods). 11 of these 30 methods did not result in an error tho they gave warnings. There was no loading instability or scoring instability across methods: all correlations >.996.[1] I saved the results from the minres+regression combination.

Inspection of the loadings revealed no important variables with the 'wrong loading' i.e., either a desirable outcome but with a negative loading or an undesirable outcome with a positive loading. Some variables are debatable. E.g. binge drinking in adults has a loading of .566, but this could be seen as a good thing (sufficient free time and money to spend it drinking large quantities of alcohol), or a bad thing (binge drinking is bad for one's health). Figure 1 shows the loadings plot.

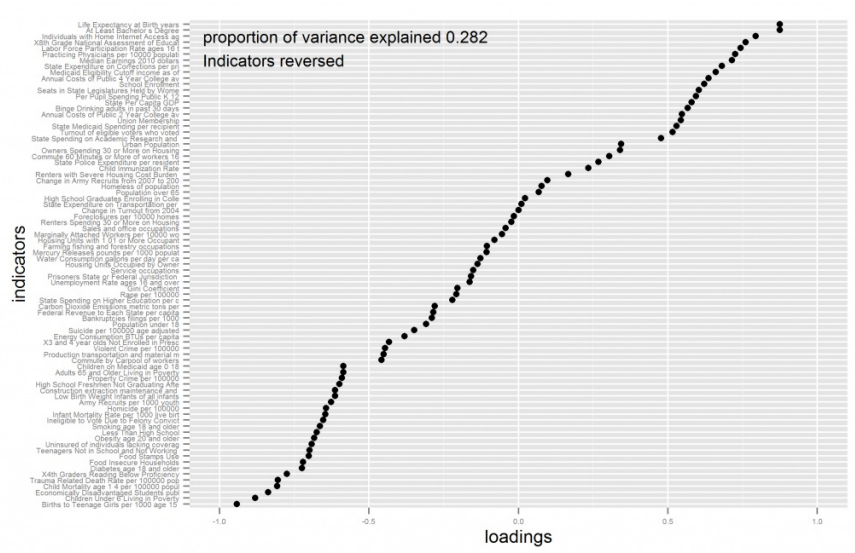

Figure 1: Loadings on the S factor. Some variable names were too long and were cut at the 40th character. Consult the main data file to see the full name.

Factor scores

The extracted factor scores were compared with previously obtained similar measures:

- HDI2010 scores calculated from HDI2002 scores found in (16).
- Measure of America's own American Human Development Index found in the dataset.
- The S factor scores from the previous study of US states (7).

The correlation matrix is shown in Table 1.

	HDI2010	S_previous	S_current	AHDI
HDI2010		0.868	0.843	0.750
S_previous	0.852		0.961	0.922
S_current	0.826	0.961		0.941
AHDI	0.724	0.913	0.945	

Table 1: Correlation matrix of S and HDI scores. Weighted correlations below the diagonal (sqrt of population).

The correlation between the previously obtained S factor and the new one was very strong at .961. The two different HDI measures had the lowest correlation. This is the expected result if they are the worst approximations of the S factor. Note however that the HDI2010 is rescaled from 2002 data, whereas the AHDI and current S factor are based on 2010 data. The previous S factor is based on data from approximately the last 10 years that were averaged.

Mixedness

Finally, factorial mixedness was examined using two methods detailed in a previous paper (17). In short, mixedness is when cases are incongruent with the overall factor structure found for the data. The methods showed convergent results (r = .65). Figure 2 shows the results.

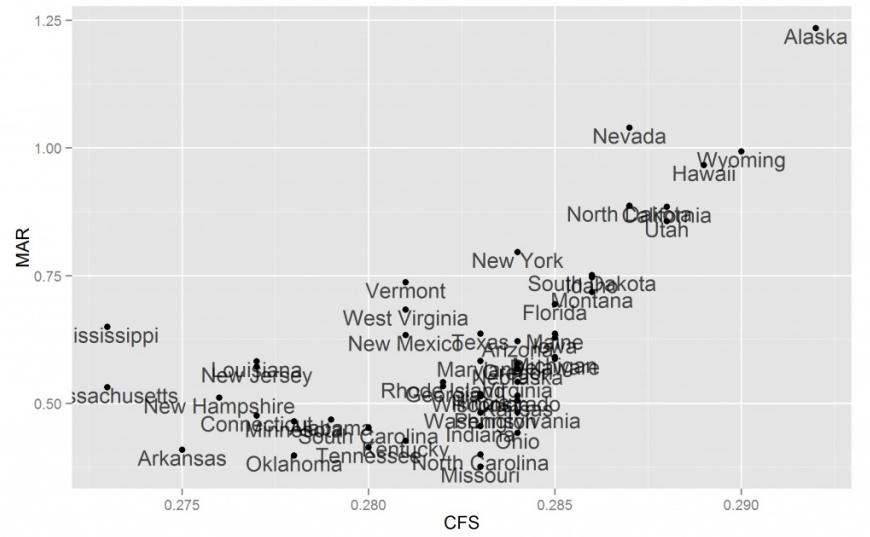

Figure 2: Factorial mixedness in cases.

If one was doing a more detailed study, one could examine the residuals at the case level and see if one can find the reasons for why an outlier state is an outlier. In the case of Alaska, the residuals for each variable are shown in Table 2.

Variable	**Residual**
Population.over.65….	-3.34
Renters.with.Severe.Housing.Cost.Burden..gross.rent…50..of.household.income.	-2.64
School.Enrollment….	-2.44
High.School.Graduates.Enrolling.in.College….	-2.08
Infant.Mortality.Rate..per.1000.live.births.	-2.07
Low.Birth.Weight.Infants….of.all.infants.	-1.99
Change.in.Turnout.from.2004	-1.89
Adults.65.and.Older.Living.in.Poverty….	-1.83
Gini.Coefficient	-1.74
Bankruptcies..filings.per.1000.	-1.67
Less.Than.High.School….	-1.34
Children.Under.6.Living.in.Poverty….	-1.32
Diabetes….age.18.and.older.	-1.29
Individuals.with.Home.Internet.Access….ages.3.and.older.	-1.14
Child.Immunization.Rate….	-1.12
Economically.Disadvantaged.Students….public.K.12.	-1.01
Renters.Spending.30..or.More.on.Housing….	-0.85
Housing.Units.Occupied.by.Owner….	-0.80
Production.transportation.and.material.moving.occupations….	-0.80
Children.on.Medicaid….age.0..18.	-0.75
Commute.60.Minutes.or.More….of.workers.16.and.over.	-0.73
Prisoners.State.or.Federal.Jurisdiction..total.number.	-0.68
Change.in.Army.Recruits.from.2007.to.2008….	-0.53

Food.Stamps.Use….	-0.52
Service.occupations….	-0.50
Foreclosures..per.10000.homes.	-0.47
Urban.Population….	-0.44
Owners.Spending.30..or.More.on.Housing….	-0.39
Annual.Costs.of.Public.4.Year.College..average…	-0.38
Property.Crime..per.100000.	-0.28
Unemployment.Rate….ages.16.and.over.	-0.28
Obesity….age.20.and.older.	-0.25
Homicide..per.100000.	-0.23
Practicing.Physicians..per.10000.population.	-0.22
Water.Consumption..gallons.per.day.per.capita.	-0.22
Seats.in.State.Legislatures.Held.by.Women….	-0.21
Sales.and.office.occupations….	-0.06
Turnout….of.eligible.voters.who.voted.	-0.04
Life.Expectancy.at.Birth..years.	-0.02
State.Spending.on.Academic.Research.and.Development…..per.capita.	0.03
Farming.fishing.and.forestry.occupations….	0.17
Mercury.Releases..pounds.per.1000.population.	0.18
Births.to.Teenage.Girls..per.1000.age.15.19.	0.20
Medicaid.Eligibility.Cutoff..income.as…of.poverty.line.	0.21
8th.Grade.National.Assessment.of.Educational.Progress.in.Math….at.or.above.proficient.	0.25
At.Least.Bachelor.s.Degree….	0.30
Smoking….age.18.and.older.	0.31
High.School.Freshmen.Not.Graduating.After.4.Years….	0.33
Child.Mortality..age.1.4.per.100000.population.	0.46
Food.Insecure.Households….	0.47
Uninsured….of.individuals.lacking.coverage.	0.66
Labor.Force.Participation.Rate….ages.16.to.64.	0.74
Army.Recruits..per.1000.youth.	0.87
Binge.Drinking….adults.in.past.30.days.	0.96
Annual.Costs.of.Public.2.Year.College..average…	0.97
Homeless….of.population.	0.98
State.Expenditure.on.Corrections….per.prisoner.	1.02
4th.Graders.Reading.Below.Proficiency….	1.05
Marginally.Attached.Workers..per.10000.working.age.Adults.	1.16
Median.Earnings..2010.dollars.	1.25
Population.under.18….	1.25
Teenagers.Not.in.School.and.Not.Working….ages.16.19.	1.36
Trauma.Related.Death.Rate..per.100000.population.	1.45

3..and.4.year.olds.Not.Enrolled.in.Preschool….	1.50
Ineligible.to.Vote.Due.to.Felony.Convictions..per.100000.voting.age.population.	1.53
Construction.extraction.maintenance.and.repair.occupations….	1.57
State.Medicaid.Spending..per.recipient.	1.70
State.Spending.on.Higher.Education….per.capita.	1.77
Violent.Crime…per.100000.	1.82
Union.Membership….	2.02
Commute.by.Carpool….of.workers.	2.02
Per.Pupil.Spending.Public.K.12….	2.06
Housing.Units.with.1.01.or.More.Occupants.per.Room….	2.10
Carbon.Dioxide.Emissions..metric.tons.per.capita.	2.16
Federal.Revenue.to.Each.State….per.capita.	2.22
Suicide..per.100000.age.adjusted.	2.35
State.Police.Expenditure….per.resident.	2.39
State.Per.Capita.GDP….	2.67
Rape..per.100000.	4.06
Energy.Consumption..BTUs.per.capita.	4.41
State.Expenditure.on.Transportation….per.person.	6.41

Table 2: Residuals per variable for Alaska.

The meaning of the numbers is this: It is the number of standard deviations that Alaska is above or below on each variable given its score on the S factor (-.24); How much it deviates from the expected level. We see that the Alaskan state spends a much more on transportation per person than expected (more than 6 standard deviations). This is presumably due to it being located very far north compared to the other states and has the lowest population density. It also spends more energy per citizen, again presumably related to the climate. I'm not sure why rape is so common, however.

One could examine the other outlier states in a similar fashion, but this is left as an exercise to the reader.

Discussion and conclusion

The present analysis used a much larger dataset of 81 very diverse variables than the previous study of the S factor in US states which used 25, yet the findings were almost identical (r = .961). This should probably be interpreted as being because the S factor can be very reliably measured when an appropriate number of and diversity of socioeconomic variables are used. It should be noted however that many of the variables between the datasets overlapped in content, e.g. expected life span at birth.

Supplementary material

Data files and source code is available on OSF.

References

1. Kirkegaard EOW. The international general socioeconomic factor: Factor analyzing international rankings. Open Differ Psychol [Internet]. 2014 Sep 8 [cited 2014 Oct 13]; Available from: openpsych.net/ODP/2014/09/the-international-general-socioeconomic-factor-factor-analyz\ning-international-rankings/

2. Kirkegaard EOW. Examining the S factor in Mexican states. The Winnower [Internet]. 2015 Apr 19 [cited 2015 Apr 23]; Available from: thewinnower.com/papers/examining-the-s-factor-in-mexican-states

3. Kirkegaard EOW. S and G in Italian regions: Re-analysis of Lynn's data and new data. The Winnower [Internet]. 2015 Apr 23 [cited 2015 Apr 23]; Available from: thewinnower.com/papers/s-and-g-in-italian-regions-re-analysis-of-lynn-s-data-and-new-d\nata

4. Kirkegaard EOW. The S factor in the British Isles: A reanalysis of Lynn (1979). The Winnower

[Internet]. 2015 Mar 28 [cited 2015 Apr 23]; Available from: thewinnower.com/papers/the-s-factor-in-the-british-isles-a-reanalysis-of-lynn-1979

5. Kirkegaard EOW. Indian states: G and S factors. The Winnower [Internet]. 2015 Apr 23 [cited 2015 Apr 23]; Available from: thewinnower.com/papers/indian-states-g-and-s-factors

6. Kirkegaard EOW. The S factor in China. The Winnower [Internet]. 2015 Apr 23 [cited 2015 Apr 23]; Available from: thewinnower.com/papers/the-s-factor-in-china

7. Kirkegaard EOW. Examining the S factor in US states. The Winnower [Internet]. 2015 Apr 23 [cited 2015 Apr 23]; Available from: thewinnower.com/papers/examining-the-s-factor-in-us-states

8. Kirkegaard EOW. The S factor in Brazilian states. The Winnower [Internet]. 2015 Apr 30 [cited 2015 May 1]; Available from: thewinnower.com/papers/the-s-factor-in-brazilian-states

9. Kirkegaard EOW. The general socioeconomic factor among Colombian departments. The Winnower [Internet]. 2015 Jun 16 [cited 2015 Jun 16]; Available from: thewinnower.com/papers/1390-the-general-socioeconomic-factor-among-colombian-departmen\nts

10. Carl N. IQ AND SOCIOECONOMIC DEVELOPMENT ACROSS REGIONS OF THE UK. J Biosoc Sci. 2015 Jun;FirstView:1–12.

11. Kirkegaard EOW. An S factor among census tracts of Boston. The Winnower [Internet]. 2015 Jun 2 [cited 2015 Jun 2]; Available from: thewinnower.com/papers/an-s-factor-among-census-tracts-of-boston

12. Kirkegaard EOW, Tranberg B. What is a good name? The S factor in Denmark at the name-level. The Winnower [Internet]. 2015 Jun 4 [cited 2015 Jun 6]; Available from: thewinnower.com/papers/what-is-a-good-name-the-s-factor-in-denmark-at-the-name-level

13. Kirkegaard EOW. Crime, income, educational attainment and employment among immigrant groups in Norway and Finland. Open Differ Psychol [Internet]. 2014 Oct 9 [cited 2014 Oct 13]; Available from: openpsych.net/ODP/2014/10/crime-income-educational-attainment-and-employment-among-imm\nigrant-groups-in-norway-and-finland/

14. Kirkegaard EOW, Fuerst J. Educational attainment, income, use of social benefits, crime rate and the general socioeconomic factor among 71 immigrant groups in Denmark. Open Differ Psychol [Internet]. 2014 May 12 [cited 2014 Oct 13]; Available from: openpsych.net/ODP/2014/05/educational-attainment-income-use-of-social-benefits-crime-r\nate-and-the-general-socioeconomic-factor-among-71-immmigrant-groups-in-denmark/

15. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research [Internet]. 2015 [cited 2015 Apr 29]. Available from: cran.r-project.org/web/packages/psych/index.html

16. Stanton EA. Inequality and the Human Development Index [Internet]. ProQuest; 2007 [cited 2015 Jun 25]. Available from: www.google.com/books?hl=en&lr=&id=87oZIFPLCykC&oi=fnd&pg=PR5&dq=INEQUALITY+AND+THE+HUM\nAN+DEVELOPMENT+INDEX+&ots=I1FCqCH_fZ&si

17. Kirkegaard EOW. Finding mixed cases in exploratory factor analysis. The Winnower [Internet]. 2015 Apr 28 [cited 2015 May 1]; Available from: thewinnower.com/papers/finding-mixed-cases-in-exploratory-factor-analysis

Footnotes

1 The factor analysis was done with the fa() function from the psych package (15). The cross-method check was done with a home-made function, see the supplementary material.