the
WINNOWER

SOCIAL SCIENCES

# The S factor in Brazilian states

EMIL O. W. KIRKEGAARD

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:
emil@emilkirkegaard.dk

DATE RECEIVED:
June 10, 2015

## Abstract

Sizeable S factors were found across 3 different datasets (from years 1991, 2000 and 2010), which explained 56 to 71% of the variance. Correlations of extracted S factors with cognitive ability were strong ranging from .69 to .81 depending on which year, analysis and dataset is chosen. Method of correlated vectors supported the interpretation that the latent S factor was primarily responsible for the association (r's .71 to .81).

## Introduction

Many recent studies have examined within-country regional correlates of (general) cognitive ability (also known as (general) intelligence, general mental ability, *g*),. This has been done for the British Isles (Lynn, 1979; Kirkegaard, 2015g), France (Lynn, 1980), Italy (Lynn, 2010; Kirkegaard, 2015e), Spain (Lynn, 2012), Portugal (Almeida, Lemos, & Lynn, 2011), India (Kirkegaard, 2015d; Lynn & Yadav, 2015), China (Kirkegaard, 2015f; Lynn & Cheng, 2013), Japan (Kura, 2013), the US (Kirkegaard, 2015b; McDaniel, 2006; Templer & Rushton, 2011), Mexico (Kirkegaard, 2015a) and Turkey (Lynn, Sakar, & Cheng, 2015). This paper examines data for Brazil.

## Data

### *Cognitive data*

Data from PISA was used as a substitute for IQ test data. PISA and IQ correlate very strongly (>.9; (Rindermann, 2007)) across nations and presumably also across regions altho this hasn't been thoroly investigated to my knowledge.

### *Socioeconomic data*

As opposed to some of my prior analyses, there was no dataset to build on top of. For this reason, I tried to find an English-language database for Brazil with a comprehensive selection of variables. Altho I found some resources, they did not allow for easy download and compilation of state-level data, which I needed. Instead, I relied upon the Portugeese-language site, *Atlasbrasil.org*.*br*, which has a comprehensive data explorer with a convenient download function for state-level data. I used Google Translate to find my way around the site.

Using the data explorer, I selected a broad range of variables. The goal was to cover most important areas of socioeconomic development and avoid variables of little importance or which are heavily influenced by local climate factors (e.g. amount of rainforest). The following variables were selected:

1. Gini coefficient
2. Activity rate age 25-29
3. Unemployment rate age 25-29
4. Public sector workers%

5. Farmers%
6. Service sector workers%
7. Girls age 10-17 with child%
8. Life expectancy
9. Households without electricity%
10. Infant mortality rate
11. Child mortality rate
12. Survive to 40%
13. Survive to 60%
14. Total fertility rate
15. Dependancy ratio
16. Aging rate
17. Illiteracy age 11-14 %
18. Illiteracy age 25 and above %
19. Age 6-17 in school %
20. Attendence higher education %
21. Income per capita
22. Mean income lowest quintile
23. Pct extremely poor
24. Richest 10 pct income
25. Bad walls%
26. Bad water and sanitation%
27. HDI
28. HDI income
29. HDI life expectancy
30. HDI education
31. Population
32. Population rural

Variables were available only for three time points: 1991, 2000 and 2010. I selected all three with intention of checking stability of results over different time periods.

Most data was already in an appropriate per unit measure so it was not necessary to do extensive conversions as with the Mexican data (Kirkegaard, 2015a). I calculated fraction of the population living in rural areas by dividing the rural population by the total population.

Note that the data explorer also has data at a lower level, that of municipals. It could be used in the future to see if the S factor holds for a lower level of aggregate analysis.

**S factor loadings**
I split the data into three datasets, one for 1991, 2000 and 2010.

I extracted S factors using the *fa()* function with default parameters from the psych package (Revelle, 2015).

*S factor in 1991*
Due to missing data, there were only 21 indicators available for this year. The data could not be imputed since it was missing for all cases for these variables. The loadings plot is shown in Figure 1.

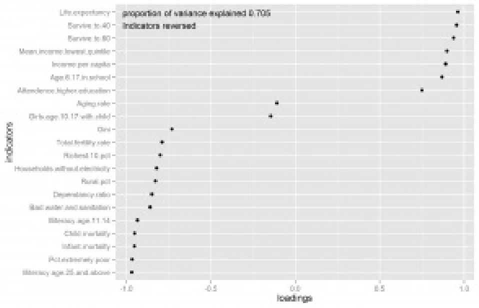

*Figure 1: Loadings plot for S factor for the data from 1991*

All indicators were in the expected direction aside from perhaps “aging rate”, which is somewhat unclear and would perhaps be expected to have a positive loading.

*S factor in 2000*
Less missing data, 26 variables. Loadings plot shown in Figure 2.

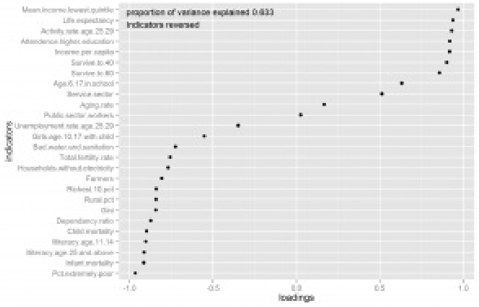

*Figure 2: Loadings plot for S factor for the 2000 data*

All indicators were in the expected direction.

*S factor for 2010*
27 variables. Factor analysis gave an error for this dataset which means that I had to remove at least one variable.[1] This left me with the question of which variable(s) to exclude. Similar to the previous analysis for Mexican states (Kirkegaard, 2015a), I used an automatic method. After removing one variable, the factor analysis worked and gave no warning. The excluded variable was child mortaility which was correlated near perfectly with another variable (infant mortality, r=.992), so little indicator sampling error should be introduced because of this deletion. The loadings plot is shown in Figure 3.

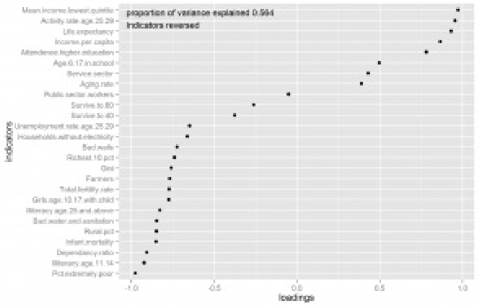

*Figure 3: Loadings plot for S factor for the 2010 data, minus one variable*

Oddly, survival to 60 and 40 now have negative loadings, altho one would expect them to correlate highly with life expectancy which has a loading near 1. In fact, the correlations between life expectancy and the survival variables was -.06 and -.21, which makes you wonder what these variables are measuring. Excluding them does not substantially change results, but does increase the amount of variance explained to .60.

Out of curiosity, I also tried versions where I deleted 5 and 10 variables, but this did not change much in the plots, so I won't show them. Interested readers can consult the source code.

**Mixed cases**

To examine whether there are any cases with strong mixedness — cases that are incongruent with the factor structure in the data — I developed two methods which are presented elsewhere (Kirkegaard, 2015c). Briefly, the first method measures the mixedness of the case by quantifying how predictable indicator scores are from the factor score for each case (mean absolute residual, MAR). The second quantifies how much the size of the general factor changes after exclusion of each individual case (improvement in proportion of variance, IPV). Both methods were found to be useful at finding a strongly mixed case in simulation data.

I applied both methods to the Brazilian datasets. For the second method, I had to create two additional reduced datasets since the factor analysis could not run with the resulting combinations of cases and indicators.

There are two ways one can examine the results: 1) by looking at the top (or bottom) most mixed cases for each method; 2) by looking at the correlations between results from the methods. The first is interesting if Brazilian state-level inequality in S has particular interest, while the second is more relevant for checking that the methods really work — they should give congruent results if mixed cases are present.

*Top mixed cases*

For each method and each analysis, I extracted the names of the top 5 mixed states. They are shown in Table 1.

| | **Position_1** | **Position_2** | **Position_3** | **Position_4** | **Position_5** |
|---|---|---|---|---|---|
| **m1.1991** | Amapá | Acre | Distrito Federal | Roraima | Rondônia |
| **m1.2000** | Amapá | Roraima | Acre | Distrito Federal | Rondônia |
| **m1.2010.1** | Roraima | Distrito Federal | Amapá | Amazonas | Acre |
| **m1.2010.5** | Roraima | Distrito Federal | Amapá | Acre | Amazonas |
| **m1.2010.10** | Roraima | Distrito Federal | Amapá | Acre | Amazonas |
| **m2.1991** | Amapá | Rondônia | Acre | Roraima | Amazonas |
| **m2.2000.1** | Amapá | Rondônia | Roraima | Paraíba | Ceará |
| **m2.2010.2** | Amapá | Roraima | Distrito Federal | Pernambuco | Sergipe |
| **m2.2010.5** | Amapá | Roraima | Distrito Federal | Piauí | Bahia |
| **m2.2010.10** | Distrito Federal | Roraima | Amapá | Ceará | Tocantins |

*Table 1: Top 5 mixed cases by method and dataset*

As can be seen, there is quite a bit of agreement across years, datasets, and methods. If one were to do a more thoro investigation of socioeconomic differences across Brazilian states, one should examine these states for unusual patterns. One could do this using the residuals for each indicator by case from the first method (these are available from the FA.residuals() in psych2). A quick look at the 2010.1 data for Amapá shows that the state is roughly in the middle regarding state-level S (score = -.26, rank 15 of 27), Farmers do not constitute a large fraction of the population (only 9.9%, rank 4 only behind the states with large cities: Federal district, Rio de Janeiro, and São Paulo). Given that farmers% has a strong negative loading (-.77) and the state's S score, one would expect the state to have relatively more farmers than it has, the mean of all states for that dataset is 17.2%.

Much more could be said along these lines, but I rather refrain since I don't know much about the country and can't read the language very well. Perhaps a researchers who is a Brailizian native could use the data to make a more detailed analysis.

*Correlations between methods and datasets*
To test whether the results were stable across years, data reductions, and methods, I correlated all the mixedness metrics. Results are in Table 2.

| | **m1.1991** | **m1.2000** | **m1.2010.1** | **m1.2010.5** | **m1.2010.10** | **m2.1991** | **m2.2000.1** | **m2.2010.2** | **m2.2010.5** |
|---|---|---|---|---|---|---|---|---|---|
| **m1.1991** | | | | | | | | | |
| **m1.2000** | 0.88 | | | | | | | | |
| **m1.2010.1** | 0.81 | 0.85 | | | | | | | |
| **m1.2010.5** | 0.77 | 0.87 | 0.98 | | | | | | |
| **m1.2010.10** | 0.70 | 0.79 | 0.93 | 0.96 | | | | | |
| **m2.1991** | 0.48 | 0.64 | 0.45 | 0.48 | 0.40 | | | | |
| **m2.2000.1** | 0.41 | 0.58 | 0.34 | 0.39 | 0.27 | 0.87 | | | |
| **m2.2010.2** | 0.53 | 0.63 | 0.66 | 0.66 | 0.51 | 0.58 | 0.68 | | |
| **m2.2010.5** | 0.32 | 0.49 | 0.60 | 0.64 | 0.51 | 0.49 | 0.59 | 0.86 | |
| **m2.2010.10** | 0.42 | 0.44 | 0.66 | 0.65 | 0.59 | 0.32 | 0.44 | 0.75 | 0.76 |

*Table 2: Correlation table for mixedness metrics across datasets and methods.*

There is method specific variance since the correlations within method (topleft and bottomright squares) are stronger than those across methods. Still, all correlations are positive, Cronbach's alpha is .87, Guttman lambda 6 is .98 and the mean correlation is .61.

**S and HDI correlations**
*HDI*
Previous S factor studies have found that HDI (Human Development Index) is basically a non-linear proxy for the S factor (Kirkegaard, 2014, 2015a). This is not surprising since the HDI is calculated from longevity, education and income, all three of which are known to have strong S factor loadings. The actual derivation of HDI values is somewhat complex. One might expect them simple to average the three indicators, or extract the general factor, but no. Instead they do complicated things (WHO, 2014).

For longevity (life expectancy at birth), they introduce ceilings at 25 and 85 years. According to data from WHO (WHO, 2012), no country has values above or below these values altho Japan is close (84 years).

For education, it is actually an average of two measures: years of education by 25 year olds and expected years of schooling for children entering school age. These measures also have artificial limits of 0-18 and 0-15 respectively.

For gross national income, they use the log values and also artificial limits of 100-75,000 USD.

Moreover, these are not simply combined by standardizing (i.e. rescaling so the mean is 0 and standard deviation is 1) the values and adding them or taking the mean. Instead, a value is calculated for every indicator using the following formula:

$$\text{dimension index} = \frac{\text{actual value} - \text{minimum value}}{\text{maximum value} - \text{minimum value}}$$

*Equation 1: HDI index formula*

Note that for education, this formula is used twice and the results averaged.

Finally, the three dimensions are combined using a geometric mean:

$$\text{HDI} = (\text{Health} \cdot \text{Education} \cdot \text{Income})^{\frac{1}{3}}$$

*Equation 2: HDI index combination formula*

The use of a geometric mean as opposed to the normal arithmetic mean, is that if a single indicator is low, the overall level is strongly reduced, whereas with the arithmetic, only the sum of the indicators matter, not the standard deviation of them. If the indicators have the same value, then the geometric and arithmetic mean have the same value.

For instance, if indicators are .7, .7, .7, the arithmetic mean is .7+.7+.7=2.1, 2.1/3=.7 and the geometric $.7^3$=0.343, $0.343^{1/3}$=.7. However, if indicators are 1, .7, .4, then the arithmetic mean is 1+.7+.4=2.1, 2.1/3=.7, but geometric mean is 1*.7*.4=0.28, $0.28^{1/3}$=0.654 which is a bit lower than .7.

*S and HDI correlations*

I used the previously extracted factor scores and the HDI data. I also extracted S factors from the HDI datasets (3 variables)[2] to see how these compared with the complex HDI value derivation. Finally, I correlated the S factors from non-HDI data, S factors from HDI data, HDI values and cognitive ability scores. Results are shown in Table 2.

| | HDI.1991 | HDI.2000 | HDI.2010 | HDI.S.1991 | HDI.S.2000 | HDI.S.2010 | S.1991 | S.2000 | S.2010.1 | S.2010.5 | S.2010.10 | CA2012 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **HDI.1991** | | 0.95 | 0.92 | 0.98 | 0.95 | 0.93 | 0.96 | 0.93 | 0.86 | 0.89 | 0.90 | 0.59 |
| **HDI.2000** | 0.97 | | 0.97 | 0.94 | 0.99 | 0.96 | 0.94 | 0.98 | 0.93 | 0.95 | 0.96 | 0.66 |
| **HDI.2010** | 0.94 | 0.98 | | 0.93 | 0.98 | 0.99 | 0.93 | 0.97 | 0.94 | 0.97 | 0.98 | 0.65 |
| **HDI.S.1991** | 0.98 | 0.96 | 0.94 | | 0.95 | 0.94 | 0.98 | 0.92 | 0.84 | 0.88 | 0.90 | 0.54 |
| **HDI.S.2000** | 0.97 | 1.00 | 0.98 | 0.97 | | 0.97 | 0.95 | 0.98 | 0.92 | 0.95 | 0.97 | 0.65 |
| **HDI.S.2010** | 0.95 | 0.98 | 0.99 | 0.95 | 0.98 | | 0.94 | 0.96 | 0.94 | 0.96 | 0.97 | 0.66 |
| **S.1991** | 0.96 | 0.96 | 0.94 | 0.97 | 0.97 | 0.96 | | 0.92 | 0.86 | 0.90 | 0.91 | 0.60 |
| **S.2000** | 0.95 | 0.98 | 0.96 | 0.93 | 0.99 | 0.97 | 0.97 | | 0.96 | 0.98 | 0.98 | 0.69 |
| **S.2010.1** | 0.89 | 0.94 | 0.94 | 0.86 | 0.94 | 0.95 | 0.92 | 0.97 | | 0.99 | 0.96 | 0.76 |
| **S.2010.5** | 0.91 | 0.95 | 0.96 | 0.89 | 0.96 | 0.97 | 0.93 | 0.98 | 0.99 | | 0.98 | 0.72 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **S.2010.10** | 0.93 | 0.96 | 0.98 | 0.93 | 0.97 | 0.98 | 0.93 | 0.97 | 0.96 | 0.98 | | 0.71 |
| **CA2012** | 0.67 | 0.73 | 0.71 | 0.60 | 0.72 | 0.74 | 0.69 | 0.78 | 0.81 | 0.79 | 0.75 | |

*Table 3: Correlation matrix for S, HDI and cognitive ability scores. Pearson's below the diagonal, rank-order above.*

All results were very strongly correlated no matter which dataset or scoring method was used. Cognitive ability scores were strongly correlated to all S factor measures. The best estimate of the relationship between S factor and cognitive ability is probably the correlation with S.2010.1, since this is the dataset cloest in time to the cognitive dataset and the S factor is extracted from the most variables. This is also the highest value (.81), but that may be a coincidence.

It is worth noting that the rank-order correlations were somewhat weaker. This usually indicates that an outlier case is increasing the Pearson correlation. To investigate this, I plot the S.2010.1 and CA2012 variables, see Figure 4.

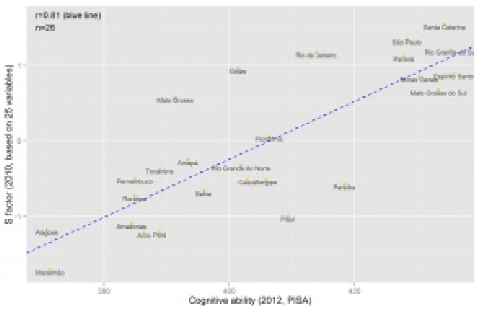

*Figure 4: Scatter plot of S factor and cognitive ability*

The scatter plot however does not seem to reveal any outliers inflating the correlation.

**Method of correlated vectors**

To examine whether the S factor was plausibly the cause of the pattern seen with the S factor scores (it is not necessarily), I used the method of correlated vectors with reversing. Results are shown in Figures 5-7.

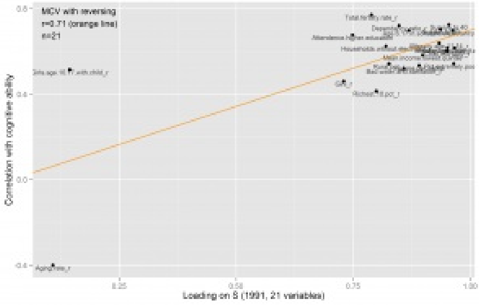

*Figure 5: MCV for the 1991 dataset*

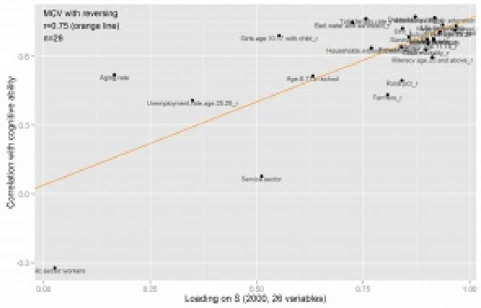

*Figure 6: MCV for the 2000 dataset*

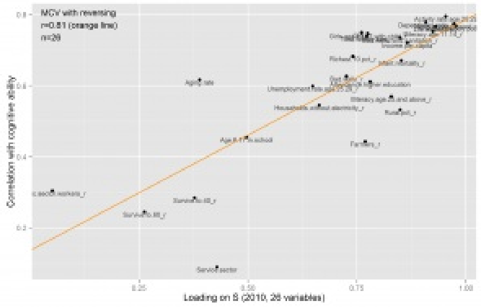

*Figure 7: MCV for the 2010 dataset*

The first result seems to be driven by a few outliers, but the second and third seems decent enough. The numerical results were fairly consistent (.71, .75, .81).

**Discussion and conclusion**

Generally, the results were in line with earlier studies. Sizeable S factors were found across 3 (or 6 if one counts the mini-HDI ones) different datasets, which explained 56 to 71% of the variance. There seems to be a decrease over time, which is intriguing as it is may eventually lead to the 'destruction' of the S factor. It may also be due to differences between the datasets across the years, since they were not entirely comparable. I did not examine the issue in depth.

Correlations of S factors and HDIs with cognitive ability were strong ranging from .60 to .81 depending on which year, analysis, dataset is chosen, and whether one uses the HDI values. Correlations were stronger when they were from the larger datasets, which is perhaps because they were better measures of latent S. MCV supported the interpretation that the latent S factor was primarily responsible for the association (r's .71 to .81).

Future studies should examine to which degree cognitive ability and S factor differences are explainable by ethnracial factors e.g. racial ancestry as done by e.g. (Kirkegaard, 2015b).

*Limitations*

There are some problems with this paper:

- I cannot read Portuguese and this may have resulted in including some incorrect variables.
- There was a lack of crime variables in the datasets, altho these have central importance for sociology. None were available in the data source I used.

**Supplementary material**

R source code, data and figures can be found in the Open Science Framework repository.

**References**

- Almeida, L. S., Lemos, G., & Lynn, R. (2011). Regional Differences in Intelligence and per Capita Incomes in Portugal. *Mankind Quarterly*, *52*(2), 213.
- Kirkegaard, E. O. W. (2014). The international general socioeconomic factor: Factor analyzing

international rankings. *Open Differential Psychology*. Retrieved from openpsych.net/ODP/2014/09/the-international-general-socioeconomic-factor-factor-analyz\ning-international-rankings/

- Kirkegaard, E. O. W. (2015a). Examining the S factor in Mexican states. *The Winnower*. Retrieved from thewinnower.com/papers/examining-the-s-factor-in-mexican-states
- Kirkegaard, E. O. W. (2015b). Examining the S factor in US states. *The Winnower*. Retrieved from thewinnower.com/papers/examining-the-s-factor-in-us-states
- Kirkegaard, E. O. W. (2015c). Finding mixed cases in exploratory factor analysis. *The Winnower*. Retrieved from thewinnower.com/papers/finding-mixed-cases-in-exploratory-factor-analysis
- Kirkegaard, E. O. W. (2015d). Indian states: G and S factors. *The Winnower*. Retrieved from thewinnower.com/papers/indian-states-g-and-s-factors
- Kirkegaard, E. O. W. (2015e). S and G in Italian regions: Re-analysis of Lynn's data and new data. *The Winnower*. Retrieved from thewinnower.com/papers/s-and-g-in-italian-regions-re-analysis-of-lynn-s-data-and-new-d\nata
- Kirkegaard, E. O. W. (2015f). The S factor in China. *The Winnower*. Retrieved from thewinnower.com/papers/the-s-factor-in-china
- Kirkegaard, E. O. W. (2015g). The S factor in the British Isles: A reanalysis of Lynn (1979). *The Winnower*. Retrieved from thewinnower.com/papers/the-s-factor-in-the-british-isles-a-reanalysis-of-lynn-1979
- Kura, K. (2013). Japanese north–south gradient in IQ predicts differences in stature, skin color, income, and homicide rate. *Intelligence*, *41*(5), 512–516. doi.org/10.1016/j.intell.2013.07.001
- Lynn, R. (1979). The social ecology of intelligence in the British Isles. *British Journal of Social and Clinical Psychology*, *18*(1), 1–12. doi.org/10.1111/j.2044-8260.1979.tb00297.x
- Lynn, R. (1980). The social ecology of intelligence in France. *British Journal of Social and Clinical Psychology*, *19*(4), 325–331. doi.org/10.1111/j.2044-8260.1980.tb00360.x
- 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. doi.org/10.1016/j.intell.2009.07.004
- Lynn, R. (2012). North-South Differences in Spain in IQ, Educational Attainment, per Capita Income, Literacy, Life Expectancy and Employment. *Mankind Quarterly*, *52*(3/4), 265.
- Lynn, R., & Cheng, H. (2013). Differences in intelligence across thirty-one regions of China and their economic and demographic correlates. *Intelligence*, *41*(5), 553–559. doi.org/10.1016/j.intell.2013.07.009
- Lynn, R., Sakar, C., & Cheng, H. (2015). Regional differences in intelligence, income and other socio-economic variables in Turkey. *Intelligence*, *50*, 144–149. doi.org/10.1016/j.intell.2015.03.006
- Lynn, R., & Yadav, P. (2015). Differences in cognitive ability, per capita income, infant mortality, fertility and latitude across the states of India. *Intelligence*, *49*, 179–185. doi.org/10.1016/j.intell.2015.01.009
- McDaniel, M. A. (2006). State preferences for the ACT versus SAT complicates inferences about SAT-derived state IQ estimates: A comment on Kanazawa (2006). *Intelligence*, *34*(6), 601–606. doi.org/10.1016/j.intell.2006.07.005
- Revelle, W. (2015). psych: Procedures for Psychological, Psychometric, and Personality Research (Version 1.5.4). Retrieved from cran.r-project.org/web/packages/psych/index.html
- Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: the homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, *21*(5), 667–706. doi.org/10.1002/per.634
- Templer, D. I., & Rushton, J. P. (2011). IQ, skin color, crime, HIV/AIDS, and income in 50 U.S. states. *Intelligence*, *39*(6), 437–442. doi.org/10.1016/j.intell.2011.08.001
- WHO. (2012). Life expectancy. Data by country. Retrieved from apps.who.int/gho/data/node.main.688?lang=en
- WHO. (2014). Human Development Report: Technical notes: Calculating the human development indices—graphical presentation. WHO. Retrieved from hdr.undp.org/sites/default/files/hdr14_technical_notes.pdf

## Footnotes

[1] Error in min(eigens$values) : invalid ‘type’ (complex) of argument.

[2] Factor loadings for HDI factor analysis were very strong, always >.9.