IN REVIEW, 7. November 2016

Open Quantitative Sociology & Political Science.

Some new methods for exploratory factor analysis of socioeconomic data

Emil O. W. Kirkegaard[1]

Abstract
Some new methods for factor analyzing socioeconomic data are presented, discussed and illustrated with analyses of new and old datasets.

A general socioeconomic factor (S) was found in a dataset of 47 French-speaking Swiss provinces from 1888. It was strongly related (r's .64 to .70) to cognitive ability as measured by an army examination. Fertility had a strong negative loading (r -.44 to -.67). Results were similar when using rank-transformed data.

The S factor of international rankings data was found to have a split-half factor reliability of .93, that of the general factor of personality extracted from 25 OCEAN items .55, and that of the general cognitive ability factor .68 based on 16 items from the *International Cognitive Ability Resource*.

Key words: general socioeconomic factor, S factor, exploratory factor analysis, research methods, Switzerland, reliability, intelligence, cognitive ability, IQ

1. Introduction

Exploratory factor analysis[2] is a method for finding underlying dimensions, called *factors*, in datasets. Because factor analysis is so useful for many purposes, it is widely used in many different sciences e.g.

[1] Ulster Institute for Social Research, United Kingdom. Email: emil@emilkirkegaard.dk

[2] Principal components analysis is included here. Some readers will bark and say that it is mathematically very different (Everitt & Dunn, 2001). That may be true, but in practice the results are nearly the same as using regular factor analytic procedures (Jensen & Weng, 1994; Kirkegaard, 2014b).

climatology, chemistry, biology, geology, psychology, computer science and sociology. The application of the method is essentially limited only to those areas of science where one needs to look for underlying patterns among variables.

Factor analysis was first invented to be used in psychometrics to analyze cognitive ability data (Cudeck & MacCallum, 2007; Spearman, 1904). Cognitive ability data are easy to analyze in the sense that better performance on one test is always correlated with better performance on another test (called *the positive manifold*) and the interest is mainly in the first factor because this has most or all of the predictive validity (Dalliard, 2013; Jensen, 1998; Malcolm J. Ree, Carretta, & Green, 2003). This means that when a single factor is extracted, all the factor loadings are positive. This has certain simplifying implications for methodology.[3]

In a previous publication (Kirkegaard, 2014b), I used factor analysis to analyze international country rankings. I found that there is a strong general factor that corresponds to the concept of general socioeconomic well-being or performance (which I called *S*). Briefly put, this means that in most, but not all, cases desirable outcomes have positive loadings and undesirable outcomes have negative loadings. This finding has since been replicated in a number of other datasets covering intra-national regions and persons grouped in various ways (Kirkegaard, 2014a, 2015e, 2015a; Kirkegaard & Tranberg, 2015).

The purpose of this paper is to review methods presented in earlier papers and to introduce new ones that were developed for the factor analysis of socioeconomic data. Some of these were presented in various earlier papers and some are new. The reason to have a single review paper of the methods is that otherwise the reader would be forced to read a number of other papers to learn about the respective methods. While the methods were developed for the purpose of studying socioeconomic data, they may be useful for other types of data as well.

2. Jensen's method with reversing

(This method was first used in (Kirkegaard, 2014b).)

Sometimes methods developed to be used for factor analysis of data that has a positive manifold do not work well when used on data where there can be genuine negative loadings. Arthur Jensen invented what he called *the method of correlated vectors* (I will refer to this as *Jensen's method*) to examine whether a given criterion variable was related to the general cognitive ability factor (g) or whether it was related to other parts of the variance (Jensen, 1985, 1998; Jensen & Reynolds, 1982). The method consists of calculating the correlation between the indicators' g-loading and the vector of the criterion variable's correlations with each indicator. The theory being that if the criterion variable is related to general cognitive ability approximated by the g factor, then the subtests that measure this trait better should correlate more strongly with the criterion variable as well. Likewise, if a criterion variable is

[3] Davide Piffer pointed out the exception with elementary cognitive tests (Jensen, 2006). These have negative correlations to the other tests because shorter response times mean better performance. Sometimes researchers reverse the response times (multiply by -1) to preverse the all positive correlation matrix. These tests are rarely used and so the problems with the negative correlations they cause are rarely present.

related to the g factor, but to other parts of the variance, then the correlation should be negative. A large and growing number of studies have used this method to examine the relationship between the g factor and variables such as brain size (Rushton & Ankney, 2009), group differences (Frisby & Beaujean, 2015; Jensen, 1985; McDaniel & Kepes, 2014; te Nijenhuis, van den Hoek, & Armstrong, 2015), the Flynn effect (te Nijenhuis & van der Flier, 2013), test training/re-taking gains (te Nijenhuis, van Vianen, & van der Flier, 2007), and education related gains (te Nijenhuis, Jongeneel-Grimen, & Kirkegaard, 2014). For criticism of the method, see e.g. Ashton and Lee (2005).

The method is generally applicable. I have previously used it to examine the relationship between S and other variables with strongly positive results, e.g. (Kirkegaard, 2014b). One problem with this is that the Jensen coefficient (the resulting correlation from applying the method) is sensitive to whether there are variables with negative loadings or not.[4] If there are, then the Jensen coefficient will be inflated towards ±1 because the presence of the negative loadings greatly increases the variance. However, the negative loading of the variables depends on arbitrary choices made by coders. For instance, one could use a variable called *percent of youth with at least a high school education* in which case one would get a positive loading. One could also code the same data negatively and call it *percent of youth without high school or better.* The loading of the indicator will depend on which way one coded it, and this affects the application of Jensen's method. Thus, if one wanted, one could strategically recode about half the variables in an S factor analyses and so inflate the Jensen coefficient to near ±1. But results should not depend on arbitrary coding choices made by researchers, especially not when they can be gamed.

A simple solution is to recode the variables such that higher values correspond to desirable outcomes. This works well in many cases, but not all. In some cases (e.g. population density, fertility, economic inequality), there is no clear answer with regards to which direction is the desirable one. Thus, subjective judgment calls affect the results which is undesirable.

Instead another method was chosen to deal with the problem: reverse all indicators with negative loadings (called *reversing*). However, in the published studies where this was done, comparison figures without reversing were not included. Thus, in order to illustrate the method, a new S factor analysis is presented below.

2.1. An S factor analysis of 47 French-speaking provinces in 19th century Switzerland

R[5] includes many datasets to use for testing and illustrative purposes. One such dataset concerns 47 French-speaking provinces in Switzerland and dates from around 1888. The dataset contains the following variables:

Fertility	Ig, ‘common standardized fertility measure’

[4] Thanks to Marc Dalliard who was the first to notice this problem. See the peer review thread for the first S factor paper at http://openpsych.net/forum/showthread.php?tid=77.

[5] R is a statistical computing language. See https://www.r-project.org/about.html.

- Agriculture % of males involved in agriculture as occupation
- Examination % draftees receiving highest mark on army examination
- Education % education beyond primary school for draftees.
- Catholic % ‘catholic’ (as opposed to ‘protestant’).
- Infant Mortality live births who live less than 1 year.

(Quoted from the dataset description. Use ?swiss in R to see.)

Four of these are clearly socioeconomic indicators: fertility, workers in agriculture, secondary educational attainment and infant mortality. The last two variables are demographic and cognitive, which we may consider criterion variables in this analysis. Note that the cognitive variable is a threshold measure of a (presumably) normal distribution of cognitive ability. This results in a non-linear transformation (La Griffe du Lion, 2001, 2007) that is expected to decrease the correlations to some unknown degree.

The 4 S indicators were factor analyzed and Figure 1 shows the factor loadings.[6]

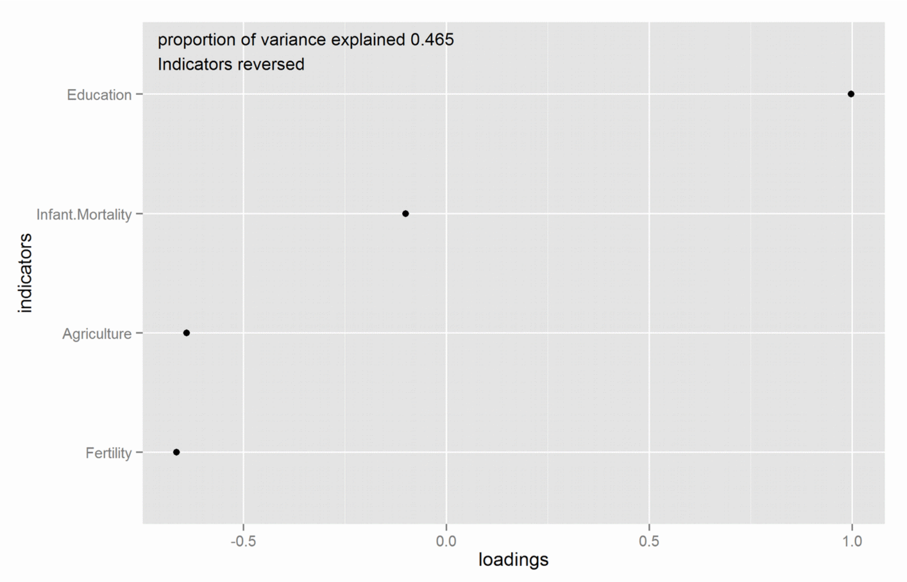

Figure 1: S factor loadings for the Swiss dataset.

Note that the figure includes a note in the top left corner that the indicators are reversed. This is because when factor analysis is performed, the factor is turned such that most loadings are positive by default. This is generally what one wants. However, in an S factor analysis, to align loadings and scores with theory, undesirable indicators should have negative loadings and desirable positive. If one analyses a dataset that includes mostly undesirable variables, the result will probably be that these are assigned positive loadings. To reverse this, one can multiply the scores and loadings by -1.

This factor turning can result in problems if one examines different subsets of a larger dataset as we

[6] Extracted with default settings. The scores were highly stable across all method parameter choices, all r's >.99.

will see later in this paper.

Table 1 shows the intercorrelations between the variables.

	Fertility	**Agriculture**	**Examination**	**Education**	**Catholic**	**Infant Mortality**	**S**	**S_noGe**	**S_rank**
Fertility	1	0.35	-0.65	-0.66	0.46	0.42	-0.67	-0.57	-0.49
Agriculture	0.35	1	-0.69	-0.64	0.40	-0.06	-0.64	-0.60	-0.66
Examination	-0.65	-0.69	1	0.70	-0.57	-0.11	0.70	0.64	0.68
Education	-0.66	-0.64	0.70	1	-0.15	-0.10	1	1	0.81
Catholic	0.46	0.40	-0.57	-0.15	1	0.18	-0.16	-0.21	-0.27
Infant Mortality	0.42	-0.06	-0.11	-0.10	0.18	1	-0.10	-0.05	-0.05
S	-0.67	-0.64	0.70	1	-0.16	-0.10	1	1	0.81
S_noGe	-0.57	-0.60	0.64	1	-0.21	-0.05	1	1	0.88
S_rank	-0.49	-0.66	0.68	0.81	-0.27	-0.05	0.81	0.88	1

Table 1: Intercorrelations in the Swiss dataset. S_noGe and S_rank are explained in Section 3.2.

Cognitive ability (*Examination*) correlates .70 with S, a large correlation in line with results from many other regional studies e.g. as found when analyzing regions of India, Brazil and Italy (Kirkegaard, 2015h, 2015d, 2015g). There is a small negative correlation with Catholic% for S, but a fairly large one for cognitive ability. If we ignore the fact that there are too few variables here for Jensen's method to work well (the precision is too low given only $N_{indicator}$=4; (Cumming, 2012)), then we could use the method to examine whether the observed correlations between the extracted factor scores and the criterion variables could plausibly be attributed to the underlying trait.

Figures 2 and 3 show Jensen's method when applied to the S x Catholic% relationship with and without reversing.

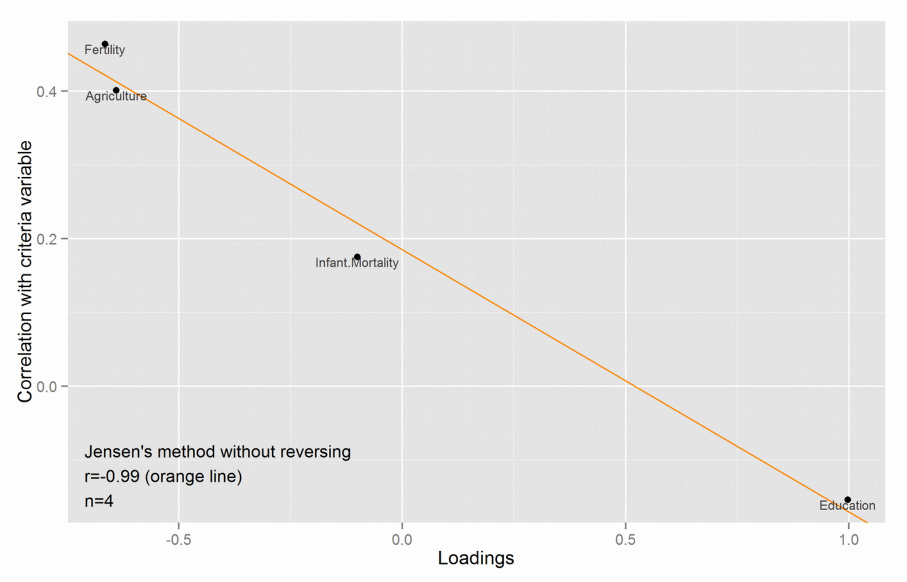

Figure 2: Jensen's method applied to S x Catholic%. Without reversing.

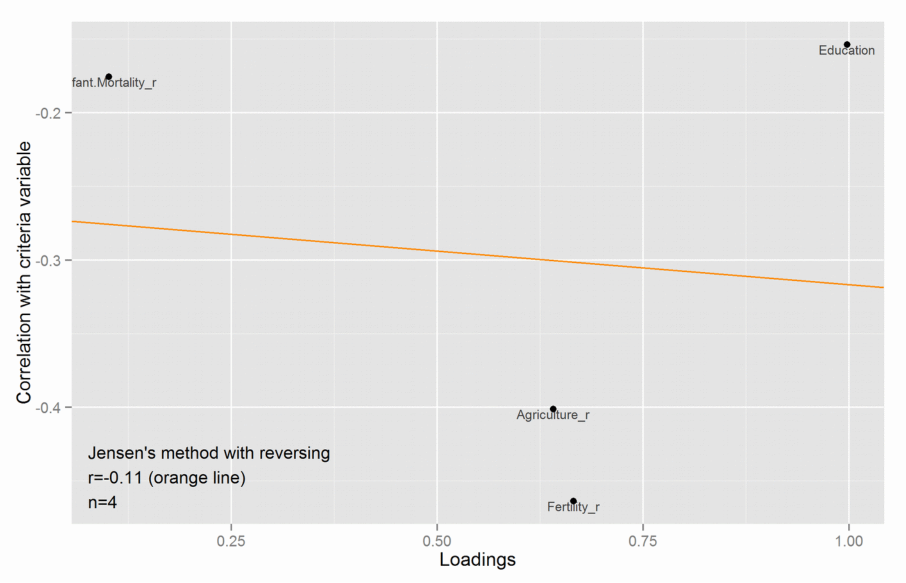

Figure 3: Jensen's method applied to S x Catholic%. With reversing.

We see a drastic change depending on whether reversing is used or not. The strong negative correlation seen in Figure 2 was entirely due to the negative loadings of the non-Education variables inflating the variance. The standard deviation of loadings in the first case is .78 and .37 in the second.

Figures 4 and 5 show Jensen's method when applied to the S x cognitive ability relationship.

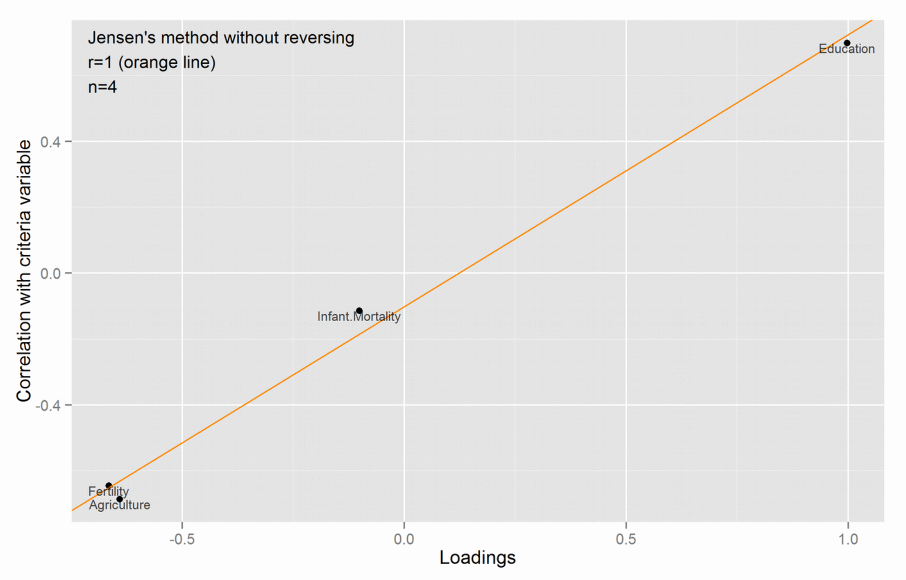

Figure 4: Jensen's method applied to S x cognitive ability. Without reversing.

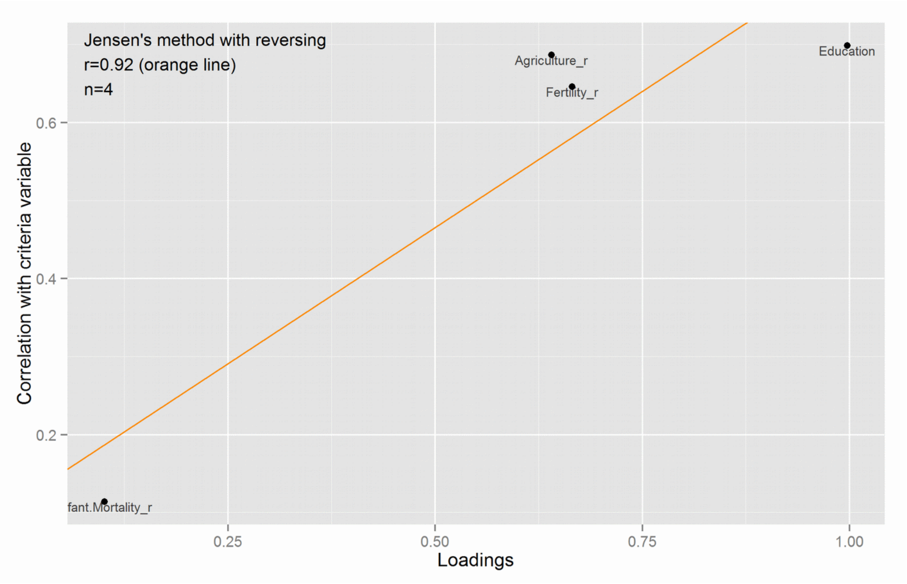

Figure 5: Jensen's method applied to S x cognitive ability. With reversing.

For the relationship to cognitive ability, however, reversing made little difference: the coefficient changed from 1.00 to .92. In other words, the strong positive relationship was not due to negative loadings inflating the variance. Of course, because the number of indicators is only 4, not much certainty can be ascribed to this finding (analytic 95% confidence interval: -.35 to 1.00).

To sum up, this study replicated the usual S factor findings for a new dataset. The strong negative loading of fertility in a dataset from the 19th century is interesting, suggesting that dysgenic selection was already in effect (Clark, 2007; Lynn, 1996). Caution is advised for this interpretation because the data are analyzed at the aggregate level. It is possible that the individual level relationship is different from the aggregate level (ecological fallacy).

3. Identifying structural outliers

(This method was first used in (Kirkegaard, 2015c).)

Exploratory factor analysis attempts to identify one or more underlying dimensions in the data. Some cases however might not exhibit the same structure as the majority of the cases.[7] For instance, a case might have high crime rates, high use of social benefits as well as high income and high educational attainment despite the loadings for these being negative and positive respectively. Such patterns are often seen for cases that consist mostly of one large city (Carl, 2015; Kirkegaard, 2015e). I previously called this phenomenon *mixedness* because the indicators of these cases give a decidedly mixed picture of the case, but it seems more suitable to use the term *structural outlier* (Kirkegaard, 2015c).

Previously, I developed two metrics for measuring the degree of structural outlierliness. The first metric, the mean absolute residual, is calculated as follows:

[7] Another option is that the data are actually composed of two or more sub-populations with markedly different factor structure. This scenario is not considered further in the present paper but warrants further study.

1. Extract the general factor.
2. Extract the factor scores.
3. For each indicator:
    1. Regress the indicator on the factor scores.
    2. Calculate and save the standardized residuals.
4. Calculate the mean absolute residual by each case.

The idea is that if a case follows the general structure of the data, then we should be able to predict that case's scores on the indicator variables from the factor score. The metric calculated above is a measure of this predictability. A value of 0 means that indicator scores are exactly predictable if we know the factor scores.

The second metric, change in factor size, is calculated as follows:

1. Extract the general factor from the complete dataset.
2. For every case:
    1. Create a subset of the dataset where this case is excluded.
    2. Extract the general factor from the subset.
    3. Extract the proportion of variance explained and save it.
    4. Calculate the difference in the proportion of variance to the factor analysis using the complete dataset and save it.

The idea here is that highly mixed cases decrease the size of the general factor because they don't fit the pattern well. The opposite is also possible, namely that they fit the structure ‘too well’ and so inflate the factor size.

A variant of the second metric is the undirectional version, where we are just interested in changes that have a large effect on the recovered factor structure and for this reason we use take the absolute value.

3.1. Two variants of a new metric based on changes in factor loadings

It is possible that exclusion of a given case results in a large changes to the indicator loadings but in such a way that the size of the general factor is not changed. Such cases would be undetectable by the change in factor size metric described above. The extent to which it changes the structure to be similar to itself will also determine the extent to which the first metric will fail to capture the effect. Thus, a new metric is proposed based on the change in loadings themselves. The the two variants of the metric is calculated as follows:

1. Extract the general factor from the complete dataset and save the factor loadings.
2. For every case:

1. Create a subset of the dataset where this case is excluded.
2. Extract the general factor from the subset.
3. Extract the factor loadings from the analysis and save them.
4. Calculate the mean and max absolute factor loadings change compared with the full analysis and save these values.

The metric scores cases by their influence on the factor loadings of the dataset, both on average and in their strongest impact.

3.2. Structural outliers in the Swiss dataset

As an example, the three metrics described above were applied to the Swiss dataset analyzed previously.

The Swiss data contain some relatively weak outliers. There are no strong outliers on the mean absolute residuals (MAR) metric. On the change in factor size (CFS), V. De Geneve is an outlier, with a value of -.045. In other words, this case increases the factor size by 4.5%points which is substantial for a dataset with 47 cases. This is the city district of Geneva, so it is not surprising it is an outlier. With respect to both mean and max absolute loading change (MeanALC and MaxALC), Geneva is also an outlier with Sierre and Neuchatel also having fairly large effects on the overall loadings.

Histograms of the structural outlierness metrics are shown in Figures 6-10.

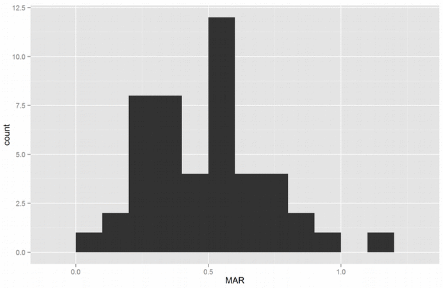

Figure 6: MAR in the Swiss dataset.

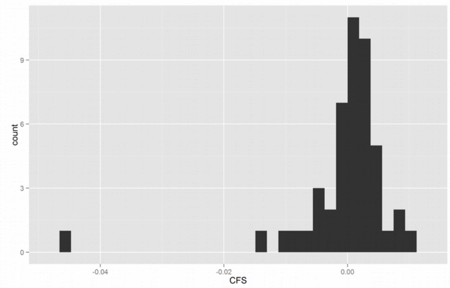

Figure 7: CFS in the Swiss dataset.

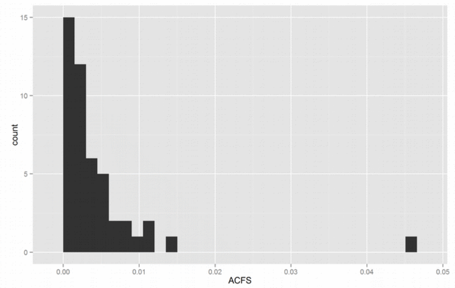

Figure 8: ACFS in the Swiss dataset.

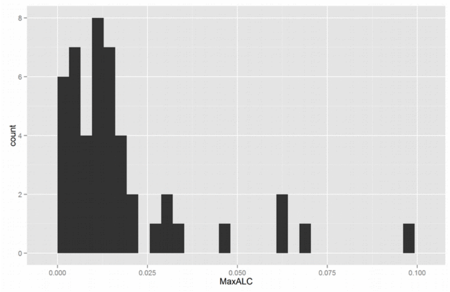

Figure 9: MaxALC in the Swiss dataset.

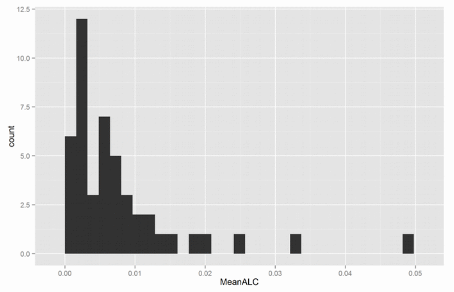

Figure 10: MeanALC in the Swiss dataset.

Since two different methods confirmed that Geneva was a likely outlier, a parallel dataset without this case was created for further analysis.

3.3. Which metric is to be preferred?

Conceptually, the metrics are somewhat distinct, so depending on the goal of the analysis, a particular metric may be the best suited for the task. But if the goal is to identify structural outliers in general, it's not clear which one is to be preferred. My current practice is using all of them and then comparing their results. Sometimes, one indicator may give divergent results and so one has to pay extra attention to see if one can figure out why. A general approach is to factor analyze the indicators to get a single structural outlierness score for each case. When doing this, it's important to choose only one of the variants of a given metric. E.g. do not use both mean and maximum absolute loading change, pick one of them.

Lastly, a nice feature of the mean absolute residuals method is that it allows one to see which indicators cause a given case to be a structural outlier. The other methods do not allow for this possibility.

4. Robust factor analysis

(This method was first used in (Kirkegaard, 2015e).)

A limitation of the methods presented above is that they merely identify the outlier cases, if any. They don't offer a way to include them in an analysis without disrupting the results. One way to do this is to employ some kind of method that is less affected by outliers. The rank-order correlation (Spearman's rho) is such a method when it comes to correlations. So since factor analysis is based on the correlations between variables, one option is to convert the dataset into rank-transformed data and then factor analyze it.

Often it will be a good idea to conduct both standard factor analysis and ranked factor analysis so that one may compare the results, e.g. as in (Kirkegaard, 2015e).

Figure 11 shows the factor loadings for standard factor analysis, standard factor analysis without Geneva and factor analysis on the rank-transformed data.

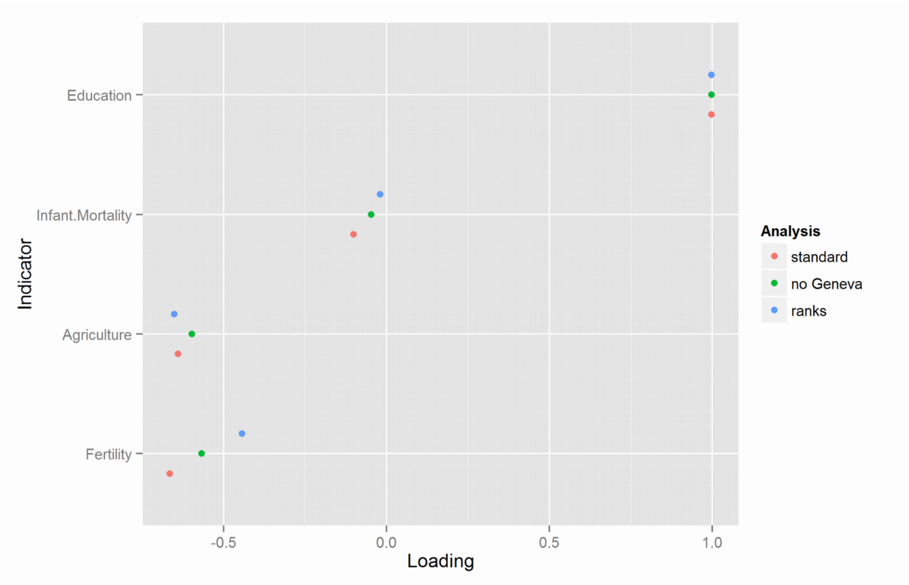

Figure 11: Factor loadings in the Swiss data using three factor analysis methods.

In this case we see that the results are very stable across methods.

We can look back at Table 1 and see that the S scores correlated between .81 and 1.00, suggesting high but not great stability across extraction methods. The correlation with cognitive ability, however, was robust (.64 to .70).

Figures 12 to 15 show the scatterplots of the data.

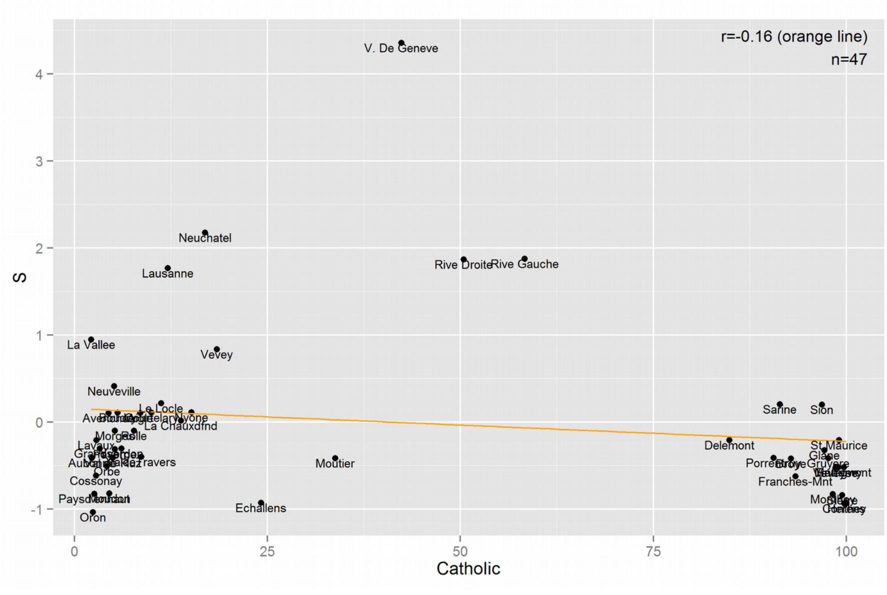

Figure 12: Scatterplot of S and Catholic% in the Swiss dataset.

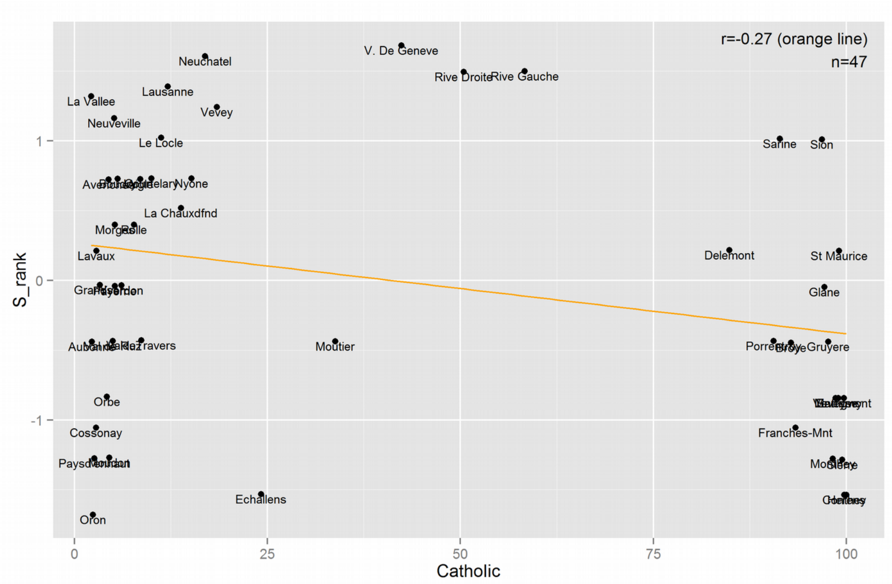

Figure 13: Scatterplot of S_rank and Catholic% in the Swiss dataset.

Both Figures 12 and 13 show a upside-down U-shaped relationship, which is very curious.

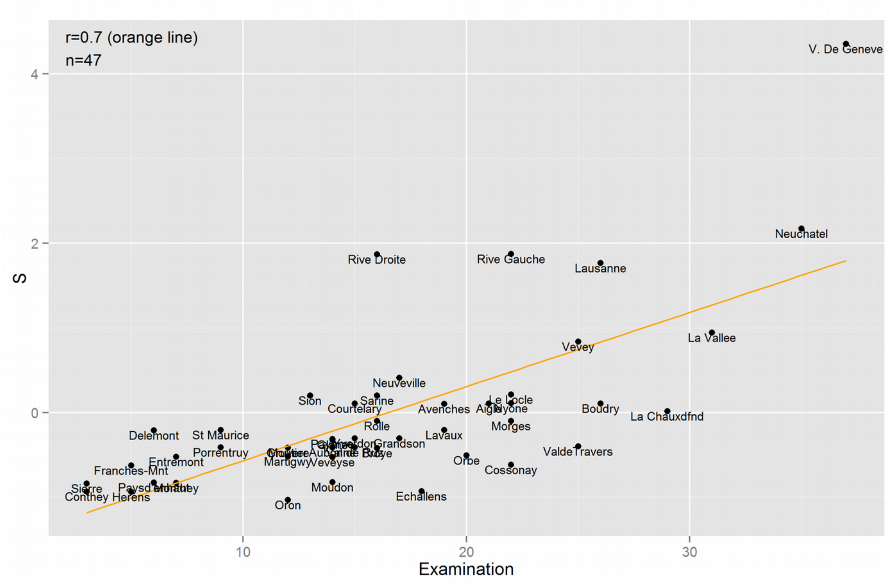

Figure 14: Scatterplot of S and cognitive ability in the Swiss dataset.

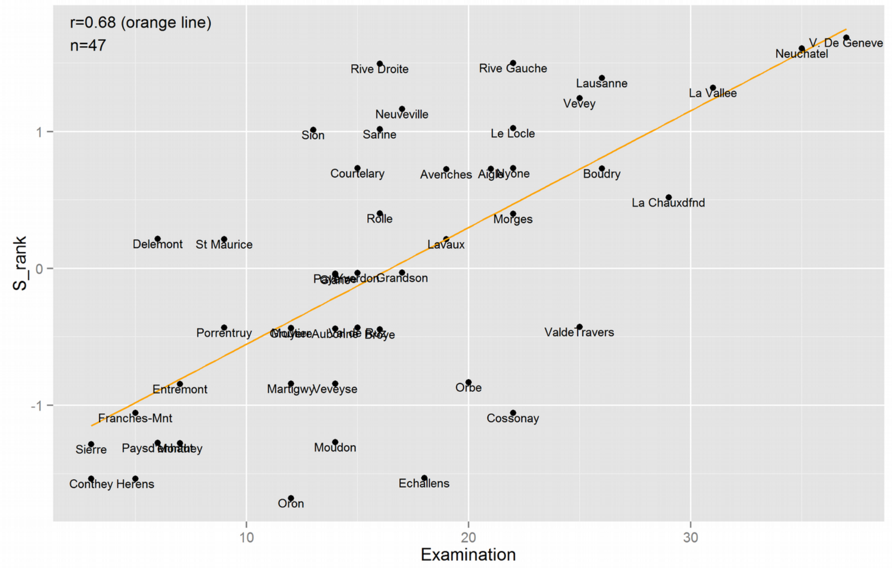

Figure 15: Scatterplot of S_rank and cognitive ability in the Swiss dataset.

We see that Geneva appeared to be inflating the correlation, but when switching to rank-transformed data, the overall pattern actually become more linear.

5. Redundant indicators

(This method was first used in (Kirkegaard, 2015b).)

One problem with extracting general factors from datasets is that if the selection of indicators is not representative of all possible indicators, the general factor will be 'colored' or 'contaminated' with variance from one of more group factors that are represented in the data (Carroll, 1993, p. 596; Jensen, 1998, p. 85). Jensen states that the 'coloring' of the general factor is one reason to prefer hierarchical factor analysis over direct factor extraction methods (Jensen, 1998). He cites no references or data for this however, so it is not clear whether this is a good recommendation. A large simulation study would probably be able to settle the issue.

Meanwhile, we might try to avoid the problem by trying to ensure that our sample of indicators is representative. Because group factors arise when groups of indicators are more strongly correlated than one would expect based on their general factor loadings, one can use this to try to avoid them. One simple idea is to avoid including variables that are very highly correlated.

In datasets of socioeconomic measures, such very highly correlated pairs often represent one construct measured in both genders (e.g. *mean income* by gender), or measured in a negative and positive fashion (e.g. *percent of the population employed* and *percent of the population receiving unemployment benefits*). Sometimes, two variables may simply be reverse coded copies of each other (e.g. *percent of population with at least high school* and *percent of population without high school*). In the case of gender-split variables, they may or may not be highly correlated. If they are, then averaging them to obtain one measure is reasonable. If they are not, however, it is instructive to include both as they may have differential relationships to the general factor or criterion variables which could be of importance.

A similar fact applies to variables that measure the same or very similar constructs positively and negatively. For instance, whether employment rate and use of unemployment benefits are very strongly correlated or not depends on whether persons in different regions have the same tendency to seek benefits when not employed. They may or may not and this might reveal an interesting pattern that could otherwise be missed.

Because the number of intercorrelations between all indicators increases quickly as a function of the number of indicators, it is practical to have an algorithm that automatically excludes variables. An algorithm was developed and works as follows:

1. Calculate the correlation between every pair of indicators.
2. Sort the indicator pairs by their absolute correlation.
3. If any correlation between a pair of indicators exceeds the threshold, the second indicator in the pair is removed, then go to step (1).

The algorithm finishes when no pair of indicators have a correlation that reach the threshold for removal. Previous studies have used a threshold of .90 which seems to work well.

6. Bootstrapping indicators, reliability and factor analysis

Since we can't include every conceivable indicator of S, there will be sampling problems with the indicators as well. Worse, a single included indicator may have a large effect on the results. This is often seen in analyses with a small number of S indicators, such as in the analysis of Swiss data above ($N_{indicator}$ = 4), where one indicator is (nearly) perfectly aligned with the extracted S factor (Kirkegaard, 2015i; Kirkegaard & Tranberg, 2015). Jensen and others have called this *psychometric sampling error (Dragt, 2010; Jensen, 1993; Kranzler & Jensen, 1991),* but a more general term would be *indicator sampling error.*

Bootstrapping is a statistical method that involves creating new random samples from the obtained sample, fitting a model to it, saving the model fitting parameters and finally calculating descriptive statistics for the distribution of model parameters. This is an alternative way to finding confidence intervals for parameter values that does not involve the usual parametric statistical assumptions. Unfortunately, we cannot use bootstrapping for factor analysis for indicators, since bootstrapping would result in duplicated indicators which result in coloring of the general factor.

However, we can split the sample of indicators randomly into two subsets and then extract the factor in each sample independently, and finally correlate the factor scores. In fact, in my first published S factor study (Kirkegaard, 2014b), I used sample splitting to show that the S factor scores were fairly stable as long as one had a reasonable number of indicators. For instance, the average *absolute* correlation of S factors extracted from a random selection of 5 indicators with the S factor extracted from all 54/42 indicators was about .90 (see Section 6.1 for an explanation). Similarly, if one chooses two sets of 5 indicators at random, the S factors extracted from them correlated .76-.80 on average (using absolute values to nullify the fact that the factors are sometimes reversed).

The method described is an extension of Cronbach's alpha, which is the mean correlation of summed scores for all possible split-halves. Because S factors can have negative loadings, using summed scores would not work well as the positively and negatively loaded variables would cancel each other out instead of aggregating.

The original R code to run these analyses, however, was not written well and would not easily be re-useable for another dataset. It was probably also for this reason that I have neglected to examine indicator sampling stability in the later studies. To rectify this, I have written a function that repeatedly splits the dataset into two random subsets of the indicators, extracts the first factor, correlates the scores and saves the results.

6.1. S score reliability in the international dataset

To illustrate the method, the data used in the international S factor paper were reanalyzed (Kirkegaard, 2014b). The data consist of 96 socioeconomic indicators of which 54 come from the Social Progress Index 2014 dataset (http://www.socialprogressimperative.org/data/spi) and 42 from the Democracy

Ranking 2013 dataset (http://democracyranking.org/). Complete data is available for only 70 cases. A closer look at the missing data reveals that many cases are missing only a few (3 or less) datapoints. These datapoints were imputed using deterministic imputation using *irmi()* from the VIM package (Templ, Alfons, Kowarik, & Prantner, 2015). This resulted in 105 cases with complete data.

A quick reliability test is to extract the national S factor again and check the correlation with the published scores which were calculated using a somewhat different method.[8] Bartlett's scoring method was used because it has been found to work well in datasets with low case $n_{cases}/n_{indicators}$ ratios, even ratios <1 (Kirkegaard, 2015b). The correlation between the new and previously published S scores was .997, so there was negligible method variance.

The split-half factor analysis reliability algorithm was run 500 times and a histogram of the results is shown in Figure 16.

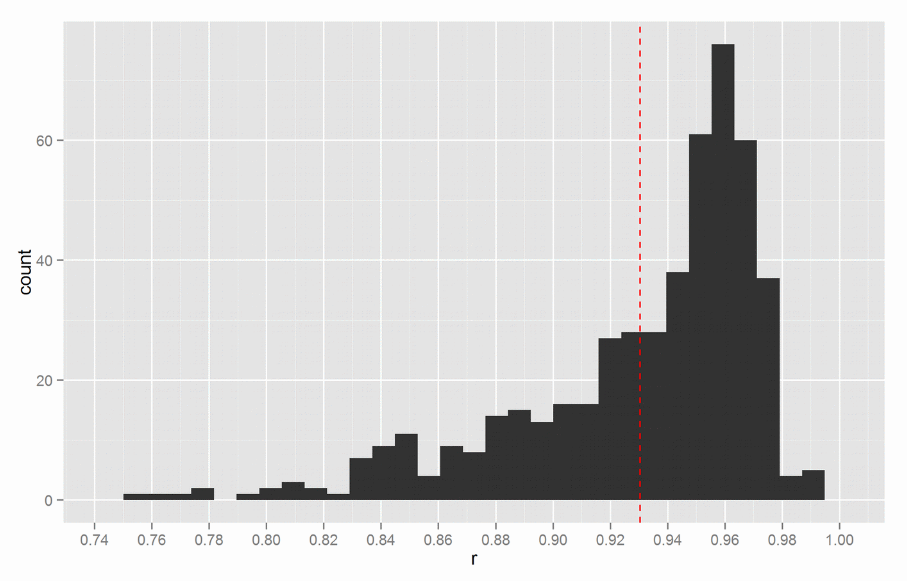

Figure 16: Histogram of split-half factor reliability runs for S from 96 indicators. N=500. Mean = .93, median .95.

All runs produced strong correlations showing that indicator sampling error is an unlikely error source for the S factor in this dataset. In other words, there is an indifference of the indicators used to measure it (Jensen, 1998).

6.2. General factor of personality score reliability in a 25 item dataset

The general factor of personality (GFP) is another recent child of factor analytic methodology (DeYoung, 2006; Digman, 1997; Musek, 2007). It is similar to the S factor in that involves the use of negative factor loadings. Briefly speaking, the general factor of personality is a proposed measure of 'good personality' or social effectiveness (van der Linden, Dunkel, & Petrides, 2016) and can be extracted from personality data in a similar fashion as the general cognitive ability factor can be extract

[8] The published scores was an average of two S factor analyses, one carried out on each dataset. This improves the coverage of countries somewhat, but decreases the number of indicators in each analysis.

from ability data.

A dataset with 2800 cases with data for 25 OCEAN/big five[9] items is included in the *psych* package (Revelle, 2015). The items are from the International Personality Item Pool and comes from the Synthetic Aperture Personality Assessment project (Revelle, Wilt, & Rosenthal, 2010).

The same method as before was used, also with 500 runs. The histogram of split-half factor correlations is shown in Figure 17.

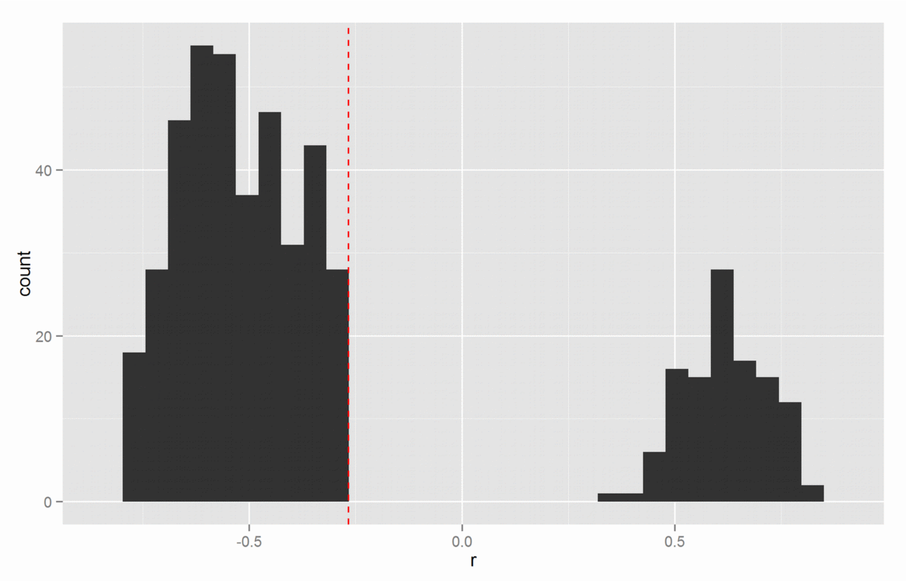

Figure 17: Histogram of split-half factor reliability runs for GFP from 25 OCEAN items. N=500.

Here we see very different results. The results are split between two distributions. Why? It is because sometimes, the split of the indicators is done such that one sample has a majority of indicators with a negative loading. The factor extraction method then reverses them so that the factor has mostly positive loadings. This means that the factor scores will also be reversed resulting in negative correlations between the factors.

One solution to this problem would be to use absolute values as was done in the original S factor study. This however will bias the reliability upwards in some cases. If there genuinely is no reliability (our measurement is pure noise), then then the factor correlations will be normally distributed around 0. If we then take the absolute values, they will all be positive and indicate that there is some reliability when there isn't. More generally, this will happen whenever the reliability distributions overlap with 0. When this is not the case, there is no bias. In the above figure, we see that the two reliability distributions do not overlap 0. Figure 18 shows the histogram using absolute values.

[9] These are: O=Openness

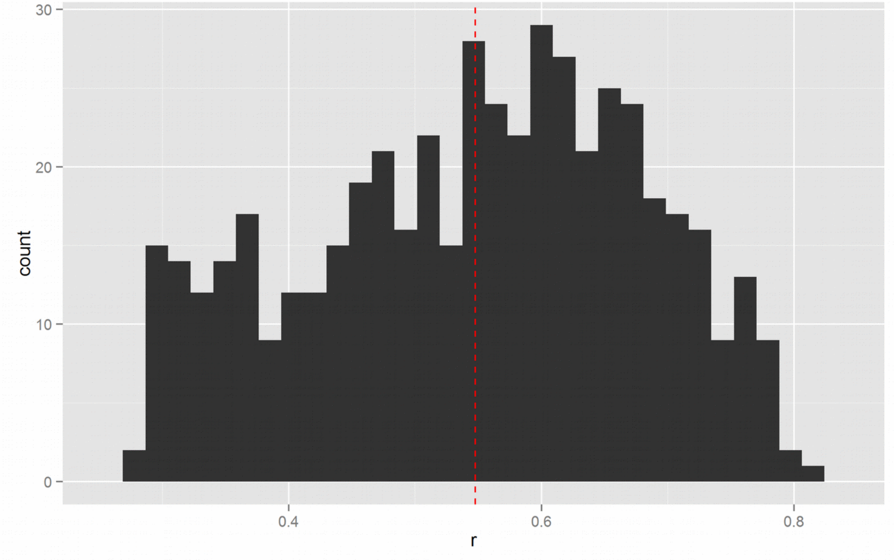

Figure 18: Histogram of split-half factor reliability runs for GFP from 25 OCEAN items. N=500. Absolute values. Mean = .55, median .56.

In other words, for GFP in this dataset, we see medium reliability across indicators.

6.3. General cognitive ability factor reliability in a dataset of 16 items

The 16 items are part of the *International Cognitive Ability Resource* test, a public domain test being developed (Condon & Revelle, 2014; Kirkegaard & Nordbjerg, 2015). The dataset is part of the *psych* package for R and contains complete data for 1248 cases (Revelle, 2015). The 16 items includes 4 verbal reasoning items, 4 alphanumeric series items, 4 matrix items and 4 3D-rotation items.

Since all loadings are positive, the split-half method is not predicted to be better than existing methods for examining internal reliability. In general, it has been found that it does not matter whether scores are derived using unit weights, factor loadings or even random numbers (M. J. Ree, Carretta, & Earles, 1998).

Figure shows the intercorrelations histogram 19. This was based on classical factor analysis because it has been found that item response theory (IRT) based and classical scores correlate near 1, making it unnecessary to use the more computationally costly IRT method (Kirkegaard, 2015f; Kirkegaard & Nordbjerg, 2015).

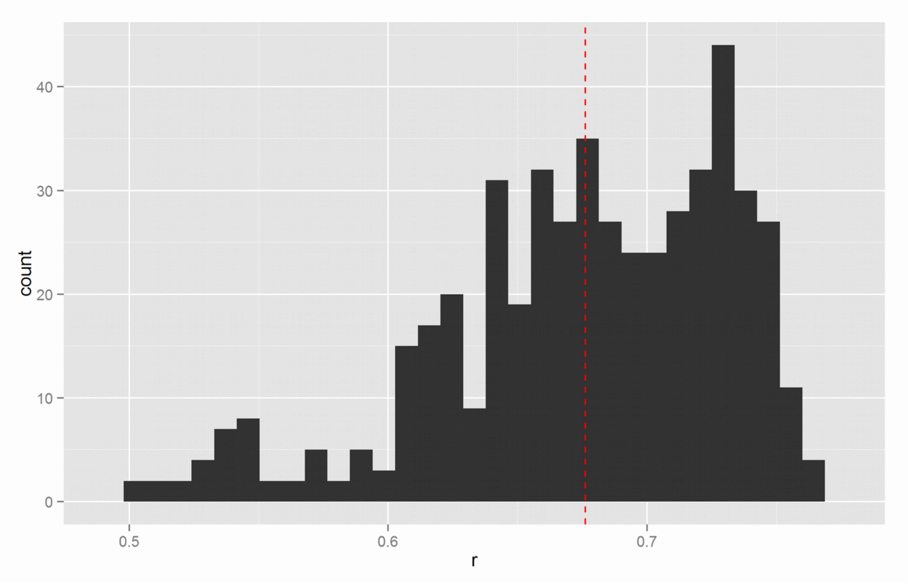

Figure 19: Histogram of split-half factor reliability runs for the GCA factor from the 16 ICAR items. N=500. Mean = .68, median .68.

By comparison, Cronbach's alpha is .83. Presumably the decrease is due to the use of factor loadings.

7. Discussion and conclusion

A number of the methods reviewed in this paper are still in the experimental stage and lack large simulation studies that back up their effectiveness. Still, the methods have been used in a number of analyses seemingly without major issues, which does give some confidence in them.

Many unresolved methodological problems with regards to exploratory factor analysis of socioeconomic data remain. First, how to meta-analysis S factor studies. Each study contains a unique batch of indicators that do not overlap perfectly or sometimes at all between studies, and it is unclear how such heterogeneous data can be aggregated in a sensible way. Second, what is the best way to generalize the current structural outlierliness methods to multi-dimensional data. Third, it is unknown whether using hierarchical or bi-factor analysis can help alleviate the problem with group factor contamination of unbalanced datasets. Fourth, it is not clear when one should use rank-ordered data to include outlier cases or when one should use interval data and exclude them. Other options include winsorizing the outlying datapoints and including them in the standard interval data analysis.

Supplementary material and acknowledgments

Functions to calculate and plot Jensen's method can be found in my personal R package, *kirkegaard*. It is found on GitHub: https://github.com/Deleetdk/kirkegaard.

The R source code and data files can be found at the Open Science Framework repository: https://osf.io/3npj8/files/.

Thanks to L. J. Zigerell, Davide Piffer and Noah Carl for reviewing the paper.

References

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence, 33*(4), 431–444. https://doi.org/10.1016/j.intell.2004.12.004

Carl, N. (2015). IQ and socioeconomic development across Regions of the UK. *Journal of Biosocial Science,* 1–12. https://doi.org/10.1017/S002193201500019X

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. Retrieved from https://www.google.com/books?hl=en&lr=&id=i3vDCXkXRGkC&oi=fnd&pg=PR7&dq=Carroll,+1993+human+cognitive+abilities&ots=3b3O4R_IKc&sig=wOss3EHXu37Q3_OZV9Due_3wyFg

Clark, G. (2007). *A farewell to alms: a brief economic history of the world*. Princeton: Princeton University Press.

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64. https://doi.org/10.1016/j.intell.2014.01.004

Cudeck, R., & MacCallum, R. C. (Eds.). (2007). *Factor analysis at 100: historical developments and future directions*. Mahwah, N.J: Lawrence Erlbaum Associates.

Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Dalliard. (2013). Is Psychometric g a Myth? *Human Varieties*. Retrieved from http://humanvarieties.org/2013/04/03/is-psychometric-g-a-myth/

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology, 91*(6), 1138–1151. https://doi.org/10.1037/0022-3514.91.6.1138

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73*(6), 1246–1256.

Dragt, J. (2010). Causes of group differences studied with the method of correlated vectors: A psychometric meta-analysis of Spearman’s hypothesis. Retrieved from http://dare.uva.nl/cgi/arno/show.cgi?fid=176083

Everitt, B. S., & Dunn, G. (2001). *Applied multivariate data analysis* (2. ed). Chichester: Wiley.

Frisby, C. L., & Beaujean, A. A. (2015). Testing Spearman’s hypotheses using a bi-factor model with WAIS-IV/WMS-IV standardization data. *Intelligence*, *51*, 79–97. https://doi.org/10.1016/j.intell.2015.04.007

Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests: Spearman’s hypothesis. *Behavioral and Brain Sciences*, *8*(02), 193. https://doi.org/10.1017/S0140525X00020392

Jensen, A. R. (1993). Psychometric g and achievement. In *Policy perspectives on educational testing* (pp. 117–227). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-94-011-2226-9_4

Jensen, A. R. (1998). *The g factor: the science of mental ability*. Westport, Conn.: Praeger.

Jensen, A. R. (2006). *Clocking the mind: mental chronometry and individual differences* (1st ed). Amsterdam ; Boston ; London: Elsevier.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, *3*(4), 423–438. https://doi.org/10.1016/0191-8869(82)90007-1

Jensen, A. R., & Weng, L.-J. (1994). What is a good g? *Intelligence*, *18*(3), 231–258. https://doi.org/10.1016/0160-2896(94)90029-9

Kirkegaard, E. O. W. (2014a). Crime, income, educational attainment and employment among immigrant groups in Norway and Finland. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2014/10/crime-income-educational-attainment-and-employment-among-immigrant-groups-in-norway-and-finland/

Kirkegaard, E. O. W. (2014b). The international general socioeconomic factor: Factor analyzing international rankings. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2014/09/the-international-general-socioeconomic-factor-factor-analyzing-international-rankings/

Kirkegaard, E. O. W. (2015a). An S factor among census tracts of Boston. *The Winnower.* Retrieved from https://thewinnower.com/papers/an-s-factor-among-census-tracts-of-boston

Kirkegaard, E. O. W. (2015b). Examining the S factor in Mexican states. *The Winnower.* Retrieved from https://thewinnower.com/papers/examining-the-s-factor-in-mexican-states

Kirkegaard, E. O. W. (2015c). Finding mixed cases in exploratory factor analysis. *The Winnower.* Retrieved from https://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 https://thewinnower.com/papers/indian-states-g-and-s-factors

Kirkegaard, E. O. W. (2015e). IQ and socioeconomic development across Regions of the UK: a reanalysis. *The Winnower.* Retrieved from https://thewinnower.com/papers/1419-iq-and-socioeconomic-development-across-regions-of-the-uk-a-reanalysis

Kirkegaard, E. O. W. (2015f). Opinions about nuclear energy and global warming, and wordsum intelligence. *The Winnower.* Retrieved from https://thewinnower.com/papers/opinions-about-nuclear-energy-and-global-warming-and-wordsum-intelligence

Kirkegaard, E. O. W. (2015g). S and G in Italian regions: Re-analysis of Lynn's data and new data. *The Winnower.* Retrieved from https://thewinnower.com/papers/s-and-g-in-italian-regions-re-analysis-of-lynn-s-data-and-new-data

Kirkegaard, E. O. W. (2015h). The S factor in Brazilian states. *The Winnower.* Retrieved from https://thewinnower.com/papers/the-s-factor-in-brazilian-states

Kirkegaard, E. O. W. (2015i). The S factor in the British Isles: A reanalysis of Lynn (1979). *The Winnower.* Retrieved from https://thewinnower.com/papers/the-s-factor-in-the-british-isles-a-reanalysis-of-lynn-1979

Kirkegaard, E. O. W., & Nordbjerg, O. (2015). Validating a Danish translation of the International Cognitive Ability Resource sample test and Cognitive Reflection Test in a student sample.

Open Differential Psychology. Retrieved from http://openpsych.net/ODP/2015/07/validating-a-danish-translation-of-the-international-cognitive-ability-resource-sample-test-and-cognitive-reflection-test-in-a-student-sample/

Kirkegaard, E. O. W., & Tranberg, B. (2015). What is a good name? The S factor in Denmark at the name-level. *The Winnower*. Retrieved from https://thewinnower.com/papers/what-is-a-good-name-the-s-factor-in-denmark-at-the-name-level

Kranzler, J. H., & Jensen, A. R. (1991). Unitary g: Unquestioned postulate or Empirical fact? *Intelligence*, *15*(4), 437–448. https://doi.org/10.1016/0160-2896(91)90005-X

La Griffe du Lion. (2001, July). Pearbotham's Law on the Persistence of Achievement Gaps. Retrieved August 16, 2015, from http://www.lagriffedulion.f2s.com/adverse.htm

La Griffe du Lion. (2007, January). Intelligence, Gender and Race. Retrieved August 16, 2015, from http://www.lagriffedulion.f2s.com/g.htm

Lynn, R. (1996). *Dysgenics: genetic deterioration in modern populations*. Westport, Conn: Praeger.

McDaniel, M. A., & Kepes, S. (2014). An Evaluation of Spearman's Hypothesis by Manipulating g Saturation. *International Journal of Selection and Assessment*, *22*(4), 333–342. https://doi.org/10.1111/ijsa.12081

Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality*, *41*(6), 1213–1233. https://doi.org/10.1016/j.jrp.2007.02.003

Ree, M. J., Carretta, T. R., & Earles, J. A. (1998). In Top-Down Decisions, Weighting Variables does Not Matter: A Consequence of Wilks' Theorem. *Organizational Research Methods*, *1*(4), 407–420. https://doi.org/10.1177/109442819814003

Ree, M. J., Carretta, T. R., & Green, M. T. (2003). The ubiquitous role of g in training. *The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen*, 261–274.

Revelle, W. (2015). psych: Procedures for Psychological, Psychometric, and Personality Research (Version 1.5.4). Retrieved from http://cran.r-project.org/web/packages/psych/index.html

Revelle, W., Wilt, J., & Rosenthal, A. (2010). Individual Differences in Cognition: New Methods for Examining the Personality-Cognition Link. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of Individual Differences in Cognition* (pp. 27–49). New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4419-1210-7_2

Rushton, J. P., & Ankney, C. D. (2009). Whole Brain Size and General Mental Ability: A Review. *The International Journal of Neuroscience*, *119*(5), 692–732. https://doi.org/10.1080/00207450802325843

Spearman, C. (1904). “General Intelligence,” Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

te Nijenhuis, J., Jongeneel-Grimen, B., & Kirkegaard, E. O. W. (2014). Are Headstart gains on the g factor? A meta-analysis. *Intelligence*, *46*, 209–215. https://doi.org/10.1016/j.intell.2014.07.001

te Nijenhuis, J., van den Hoek, M., & Armstrong, E. L. (2015). Spearman’s hypothesis and Amerindians: A meta-analysis. *Intelligence*, *50*, 87–92. https://doi.org/10.1016/j.intell.2015.02.006

te Nijenhuis, J., & van der Flier, H. (2013). Is the Flynn effect on g?: A meta-analysis. *Intelligence*, *41*(6), 802–807. https://doi.org/10.1016/j.intell.2013.03.001

te Nijenhuis, J., van Vianen, A. E. M., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, *35*(3), 283–300. https://doi.org/10.1016/j.intell.2006.07.006

Templ, M., Alfons, A., Kowarik, A., & Prantner, B. (2015, February 19). VIM: Visualization and Imputation of Missing Values. CRAN. Retrieved from http://cran.r-project.org/web/packages/VIM/index.html

van der Linden, D., Dunkel, C. S., & Petrides, K. V. (2016). The General Factor of Personality (GFP) as social effectiveness: Review of the literature. *Personality and Individual Differences*, *101*, 98–105. https://doi.org/10.1016/j.paid.2016.05.020