How performing PCA and CFA on the same data equals trouble

Overfitting in the assessment of internal structure and some editorial thoughts on it

Marjolein Fokkema, University of Leiden, The Netherlands

Samuel Greiff, University of Luxembourg, Luxembourg

We regularly receive papers at EJPA where a principal component analysis (PCA) or exploratory factor analysis (EFA)[1] is performed, followed by a confirmatory factor analysis (CFA) on the same (or partially overlapping) data. On the one hand, we are thankful for these submissions as they simplify the often tedious editorial task, by providing good grounds for on-desk rejection (see also Greiff & Ziegler, 2017). But when such grounds for rejection are all too regularly employed, they may instill a feeling of unease in the editor: Am I turning into a sour, nitpicking bureaucrat? Am I too strict and stuck with my own ideas of what good science is? Can we not let the data speak for itself?

To confront such feelings of unease, we wanted to see whether the consequences of performing PCA and CFA on the same dataset are indeed so dire and justify rejection. To this end, we ran an experiment with simulated data and would like to share the results in this editorial. With the results of the simulation in mind, we will give some editorial advice on how authors can avoid trouble coming along with combining PCA and CFA. The R code and output for the experiment are provided in the online supplementary material.

Experiment

We randomly generated values for 25 completely uncorrelated, standard normally distributed item scores for 300 observations each. For illustrational purposes, we first calculated the inter-item correlations of these items that, importantly, are uncorrelated in the population. Figure 1 depicts a histogram of the resulting correlations. The sample correlations are indeed distributed around a mean of 0, but note that some values approach .2 and -.2, which would be interpreted as a small to medium effect size.

[1] We fully agree with readers who take offense in the confusion of principal component analysis and exploratory factor analysis. The two are different techniques, involving different assumptions, estimation methods, and different interpretations of the results. We will discuss the differences in the subsection 'Some nitpicking all the same'

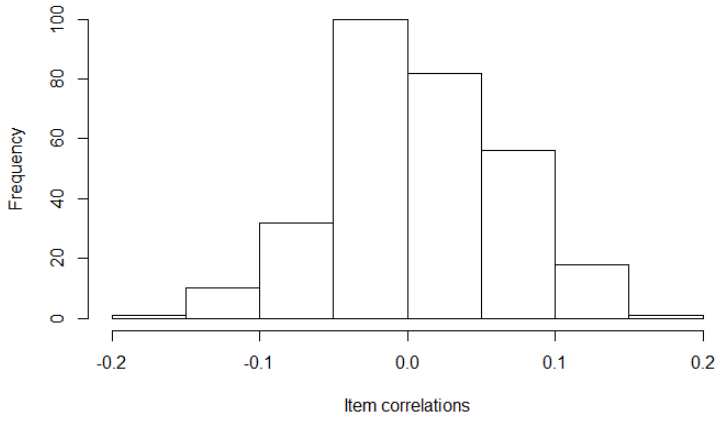


Figure 1. Histogram of inter-item correlations for the 25 item scores.

For the sake of the example, let us forget that we know the variables are in fact uncorrelated in the population. Instead, we take the position of a researcher who just sees the data and performs a PCA on them. Figure 2 depicts the resulting scree plot.

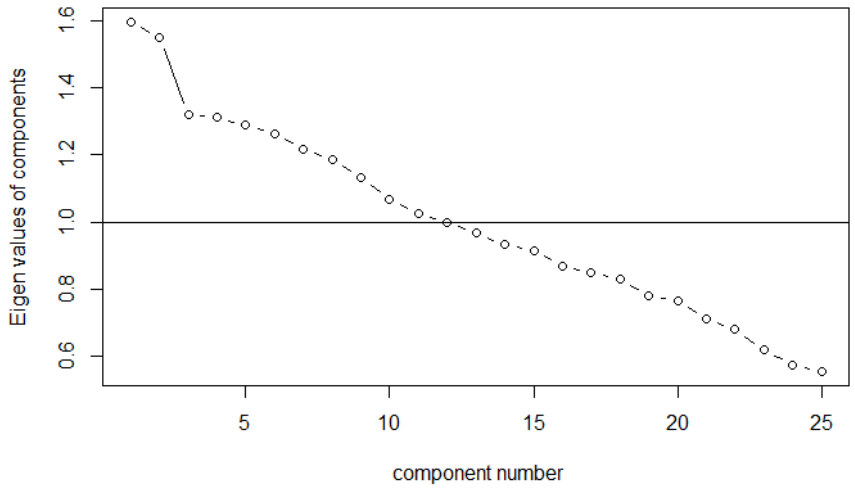


Figure 2. Scree plot resulting from PCA on the 25 uncorrelated items.

We perform a rough visual scree test (Cattell, 1966) to select the number of components to retain. That is, taking a look at Figure 2, we select components until the eigenvalues show a sudden drop and start to level off. We therefore proceed with a two-component solution. The varimax rotated loadings of the two-component solution are presented in Table 1. We retain items with loadings > .40. Many items do not correlate that strongly with either component and, against the backdrop of these results, should be discarded. In addition, some of the items may need reverse coding, because of negative loadings.

Table 1

Standardized component loadings. Boldfaced loadings have absolute values > .40.

Item	Component 1	Component 2
x1	0.19	-0.25
x2	0.10	**0.44**
x3	**0.40**	0.13
x4	-0.02	**-0.58**
x5	-0.19	-0.22
x6	0.15	0.10
x7	-0.03	-0.15
x8	-0.16	0.25
x9	0.14	-0.04
X10	0.39	-0.02
x11	**0.41**	0.03
x12	-0.21	-0.23
x13	-0.04	**0.42**
x14	-0.17	0.38
x15	**0.48**	0.01
x16	-0.23	0.10
x17	**0.43**	0.10
x18	0.11	-0.17
x19	-0.06	0.05
x20	0.22	0.04
x21	-0.33	-0.05
x22	0.09	0.24
x23	**0.44**	-0.11
x24	-0.18	0.12
x25	0.01	**0.53**

As mentioned before, we see evidence for a 2-factor solution after the initial PCA. In order to further validate this result, we now proceed with conducting a CFA – of note, in doing so we use the same sample of 300 observations on which we conducted the PCA.

For the CFA, we specify two latent factors and specify loadings in accordance with Table 1. That is, x3, x11, x15, x17 and x23 are assumed to load on the first factor, whereas x2, x4, x13 and x25 are assumed to load on the second factor. The remaining items are omitted from the model. Furthermore, we allow the correlation between factors to be estimated freely, as is common in psychological research.

Strikingly, the resulting model shows very good fit to the data. The chi-square test and other fit indices indicate excellent fit: χ^2 (26) = 16.925; *p* = .911; CFI = 1.000; RMSEA < .001; SRMR = .034.

The estimated factor loadings are presented in Table 2. The absolute values of the standardized loadings range from .19 to .49, indicating substantial correlations between the items and factors, which could be interpreted as further evidence for the appropriateness of the two-factor model. The *p* values of the loadings range from .039 to .144. Although only one loading is significant at the alpha = .05 level, the *p* values are still rather low, given that the null hypothesis is true in the population (the true correlation between items is zero, so the true correlation between items and factors must also be zero).

Table 3 presents the estimated factor (co)variances, which provide the clearest indication that the data were actually generated from a population model of zero correlations: the factor (co)variances are and the *p* value clearly indicate that they do not differ significantly from zero.

Table 2

Estimated factor loadings and standard errors.

Factor	Item	Loading (SE)	*p* value	Standardized loading
Factor 1	X3	1.00 (-----)	-----	.189
	X11	1.84 (1.22)	.131	.349
	X15	1.83 (1.21)	.131	.343
	X17	1.52 (1.04)	.144	.280
	X23	1.66 (1.12)	.140	.294
Factor 2	X2	1.00 (-----)	-----	.267
	X4	-2.01 (1.06)	.059	-.490
	X13	0.89 (0.50)	.075	.234
	X25	1.51 (0.73)	.039	.365

Table 3

Estimated factor (co)variances

Factor	(co)variance (SE)	*p* value
Factor 1	0.037 (0.040)	.362
Factor 2	0.065 (0.049)	.187
Factor 1 & 2	0.001 (0.008)	.907

Interpretation

The little experiment above has shown us that performing PCA and CFA on the same data can indeed have dire consequences: It yields deceivingly optimistic model fit indices and parameter estimates. One may wonder how we could obtain such excellent model fit indices with data that were generated as to be uncorrelated? The answer is two-fold:

Firstly, because of capitalizing on chance characteristics of the data. We performed an exploratory analysis, found patterns that in reality only reflected sampling fluctuations and used those as a hypothesis for a confirmatory analysis on the same data. This is also called overfitting, which yields inflated estimates of model fit and parameter estimates. Obviously, we are not the first to write about this topic. In fact, a vast body of literature has been devoted to overfitting, or capitalizing on the idiosyncrasies of the sample at hand. Excellent further readings on this relevant topic are Babyak (2004) or Yarkoni and Westfall (in press).

Secondly, model fit indices in SEM are a function of how well sample covariances are reproduced by the fitted model. If the sample covariances are small (relative to the sample variances), which they are in our example, it will be easy for any model to reproduce them well and show excellent model fit. For some important further thoughts on model fit see the recent editorial by Greiff and Heene (2017).

We should note that the results we obtained in the experiment are not coincidental: replicating the same procedure for generating and analyzing data as above yields similar, overly optimistic results in terms of model fit, parameter estimates and test statistics. Of note, with increasing sample size, the risk of overfitting decreases, as sample correlations approximate the population mean more and more closely as sample size increases.

Of course, objections may be raised to our experiment. For example, that a Kaiser-Meyer-Olkin test should be performed prior to performing PCA (maybe also prior to CFA), that most loadings were not statistically significant so the model fit is not that good, that selecting 9 items out of 25 is quite extreme, or that a zero-correlation population model is not representative for psychological research. We agree with such objections. In fact, here we merely aimed at providing an example of how exploration and confirmation using the same data yields overly optimistic, misleadingly meaningful results. Applying such procedures to datasets from real-world studies will also yield overly optimistic results.

Some nitpicking all the same

Admittedly, this little experiment does not provide a rigorous test of our being nitpicking bureaucrats or not (though some may argue that the mere fact of undertaking this endeavor proves that we are). Therefore, we would like to also stress here that PCA should never be referred to as (exploratory) factor analysis. Regularly, manuscripts submitted to EJPA state that factor analysis was performed, while the method section reports the use of

PCA. Although PCA and EFA share similarities, they are mathematically and conceptually different: principal components represent a parsimonious summary of the item scores, whereas common factors are assumed to underlie or cause the observed item scores. In other words, whereas EFA implies a reflective measurement model, PCA implies a formative measurement model. A reflective model assumes a direct effect from the construct on the item scores, while a formative model assumes item scores to be the causes of a construct. Both views on psychological assessment can be equally valid and often yield similar parameter estimates, but they are very different from a psychometric perspective. An enlightening and in-depth discussion of such measurement models is provided by Edwards and Bagozzi (2000).

Furthermore, we would like to stress that assessing the internal structure of a psychological measure through exploratory analyses is in most cases uncalled for. Although this point has already been stressed in earlier editorial(s) (e.g., Ziegler, 2014), exploratory approaches such as PCA still regularly appear to be the first weapon of choice of researchers who want to assess internal structure. The message therefore bears repeating: an exploratory approach is appropriate when the number of factors and the allocation of items to factors are unknown. In most cases, however, measures were designed to capture specific (sub)constructs, providing a clear hypothesis that would best be tested with a confirmatory technique.

Some advice for authors, reviewers and editors

Obviously, overfitting yields unreliable results and should be avoided, both on a general level and, of course, also for submissions to EJPA. This editorial is meant to increase awareness of this issue and, at the same time, to offer guidance to authors considering submission of their work to EJPA. Therefore, we will conclude this editorial with some initial thoughts we suggest that authors follow:

1. Refrain from performing exploratory and confirmatory analyses on the same dataset as this yields high danger of overfitting, in particular in smaller data sets;
2. Refrain from performing exploratory analyses to assess internal structure as much as possible;
3. If the goal of your study requires both exploration and confirmation, and sample size is large enough, perform each on separate data, for instance by splitting the data set;
4. When evaluating the results of a CFA, do not only focus on model fit indices, but also inspect parameter estimates (factor loadings, factor (co) variances, residual variances), their standard errors and *p* values.

References

Babyak, M. A. (2004). What you see may not be what you get. A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*, 411-421. doi: 10.1097/00006842-200405000-00021

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*, 155-174. doi: 10.1037//1082-989x.5.2.155

Greiff, S. & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment, 33*, 313-317.

Greiff, S. & Ziegler, M. (2017). How to make sure your paper is desk rejected. A practical guide to rejection in EJPA. *European Journal of Psychological Assessment 33*, 75-78. doi: 10.1027/1015-5759/a000419

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Yarkoni, T., & Westfall, J. (in press). Choosing prediction over explanation in psychology. Lessons from machine learning. *Perspectives in Psychological Science*. Retrieved from http://jakewestfall.org/publications/Yarkoni_Westfall_choosing_prediction.pdf. doi: 10.1177/1745691617693393

Ziegler, M. (2014). Comments on item selection procedures. *European Journal of Psychological Assessment, 30*, 1-2. doi: 10.1027/1015-5759/a000196

How performing PCA and CFA on the same data equals trouble

Supplementary material

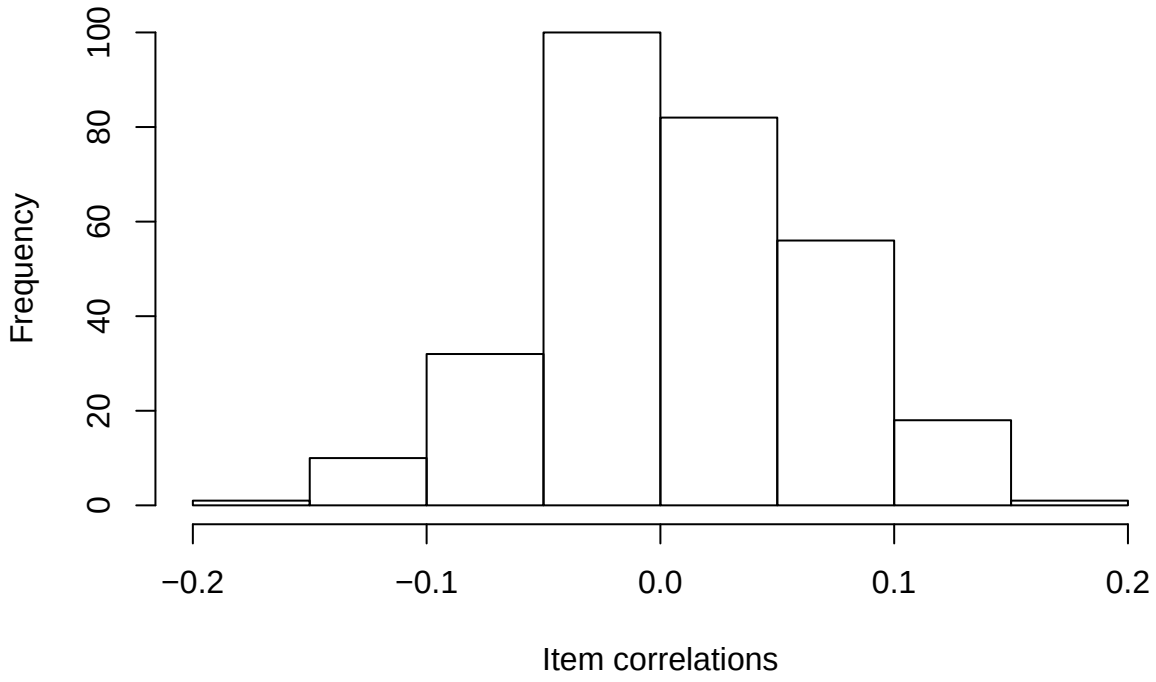
Data generation

We start by setting the random seed to make sure the results are random but can be reproduced exactly. We generate a data matrix with 25 uncorrelated columns (variables $X1$ through $X15$) and 300 rows (observations):

```
set.seed(403612)
n <- 300
p <- 25
data <- matrix(rnorm(n*p), nrow = n, ncol = p)
colnames(data) <- paste0("x", 1:p)
```

We take a look at the item correlations, first:

```
hist(cor(data)[upper.tri(cor(data))], main = "", xlab = "Item correlations")
```

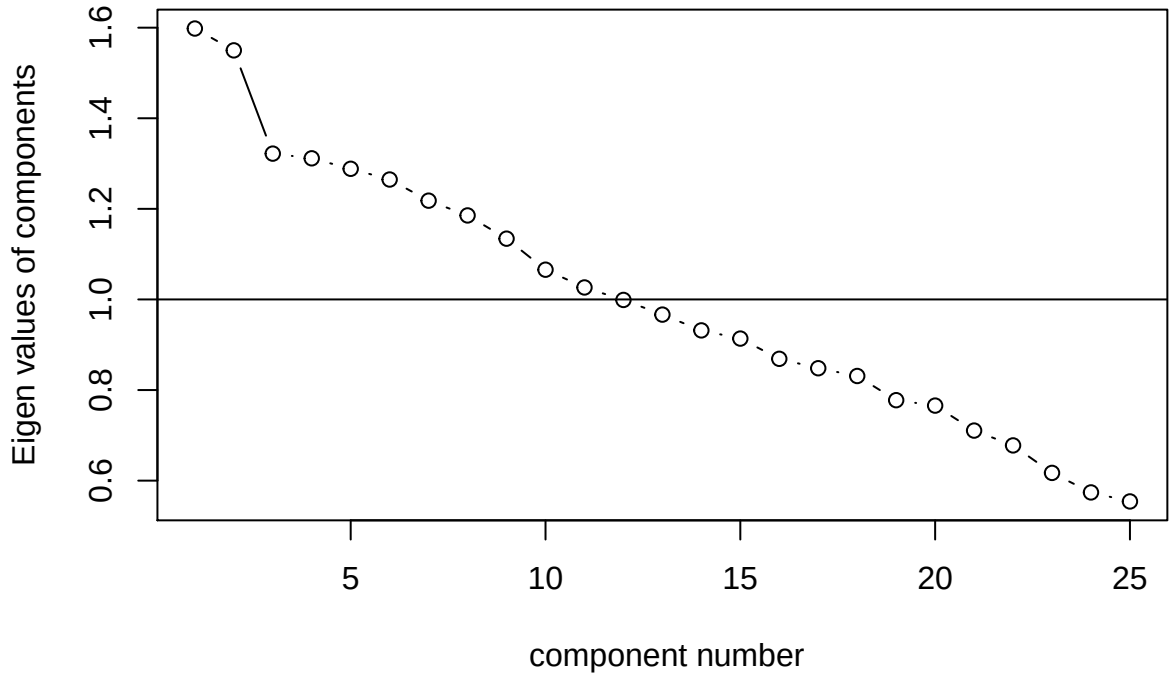


Although the data were generated from a population model of zero correlations, some sample correlations deviate substantially from zero, some even approaching .2 and -.2. For now, we forget that we know the variables are in fact uncorrelated and perform a PCA on the data.

PCA

We perform the PCA using the R package psych (Revelle 2017):

```r
library(psych)
VSS.scree(data, main = "")
```



We employ the elbow criterion, where we add componenents until the eigenvalues (or variance explained) shows a sharp decrease and levels off. Therefore, we continue by requesting the two-component solution for further exploration and interpretation. We employ the default rotation method, which is varimax:

```r
principal(data, nfactors = 2)
```

```
## Principal Components Analysis
## Call: principal(r = data, nfactors = 2)
## Standardized loadings (pattern matrix) based upon correlation matrix
##       RC1   RC2     h2   u2 com
## x1   0.19 -0.25 0.0985 0.90 1.8
## x2   0.10  0.44 0.2050 0.80 1.1
## x3   0.40  0.13 0.1797 0.82 1.2
## x4  -0.02 -0.58 0.3340 0.67 1.0
## x5  -0.19 -0.22 0.0856 0.91 1.9
## x6   0.15  0.10 0.0310 0.97 1.7
## x7  -0.03 -0.15 0.0229 0.98 1.1
## x8  -0.16  0.25 0.0887 0.91 1.7
## x9   0.14 -0.04 0.0213 0.98 1.2
## x10  0.39 -0.02 0.1559 0.84 1.0
## x11  0.41 -0.03 0.1702 0.83 1.0
## x12 -0.21 -0.23 0.0986 0.90 2.0
## x13 -0.04  0.42 0.1762 0.82 1.0
## x14 -0.17  0.38 0.1719 0.83 1.4
```

```
## x15  0.48  0.01 0.2341 0.77 1.0
## x16 -0.23  0.10 0.0653 0.93 1.4
## x17  0.43  0.10 0.1959 0.80 1.1
## x18  0.11 -0.17 0.0427 0.96 1.7
## x19 -0.06  0.05 0.0061 0.99 1.9
## x20  0.22  0.04 0.0503 0.95 1.1
## x21 -0.33 -0.05 0.1097 0.89 1.1
## x22  0.09  0.24 0.0660 0.93 1.3
## x23  0.44 -0.11 0.2092 0.79 1.1
## x24 -0.18  0.12 0.0471 0.95 1.8
## x25  0.01  0.53 0.2821 0.72 1.0
##
##                        RC1  RC2
## SS loadings           1.59 1.55
## Proportion Var        0.06 0.06
## Cumulative Var        0.06 0.13
## Proportion Explained  0.51 0.49
## Cumulative Proportion 0.51 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  834  with prob <  1.1e-63
##
## Fit based upon off diagonal values = -0.38
```

We will retain items with compnent loadings $> .40$. The first component consist of x3, x11, x15, x17 and x23, while the second component consists of x2, x4, x13 and x25. We discard the remaining items.

CFA

We use the structure we obtained through PCA as the input for a CFA on the very same data. We perform the CFA using the R package lavaan (Rosseel 2012):

```
library(lavaan)
model <- '
  F1 =~ x3 + x11 + x15 + x17 + x23
  F2 =~ x2 + x4 + x13 + x25
'
fit <- cfa(model, data)
```

We request some model fit indices:

```
fitmeasures(fit, c("chisq", "df", "pvalue", "cfi", "srmr", "rmsea"))
```

```
##  chisq     df pvalue    cfi   srmr  rmsea
## 16.925 26.000  0.911  1.000  0.034  0.000
```

All indices report excellent model fit! We request the parameter estimates:

```
summary(fit, standardized = TRUE)
```

```
## lavaan (0.5-23.1097) converged normally after  74 iterations
##
##   Number of observations                           300
```

```
##
##   Estimator                                         ML
##   Minimum Function Test Statistic               16.925
##   Degrees of freedom                                26
##   P-value (Chi-square)                           0.911
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Standard Errors                             Standard
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   F1 =~
##     x3                1.000                               0.192    0.189
##     x11               1.842    1.219    1.511    0.131    0.353    0.349
##     x15               1.826    1.209    1.510    0.131    0.350    0.343
##     x17               1.523    1.042    1.462    0.144    0.292    0.280
##     x23               1.659    1.123    1.477    0.140    0.318    0.294
##   F2 =~
##     x2                1.000                               0.256    0.267
##     x4               -2.006    1.061   -1.890    0.059   -0.513   -0.490
##     x13               0.893    0.502    1.779    0.075    0.228    0.234
##     x25               1.513    0.734    2.061    0.039    0.387    0.365
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   F1 ~~
##     F2                0.001    0.008    0.117    0.907    0.019    0.019
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .x3                0.987    0.087   11.304    0.000    0.987    0.964
##    .x11               0.900    0.104    8.629    0.000    0.900    0.878
##    .x15               0.915    0.105    8.752    0.000    0.915    0.882
##    .x17               1.002    0.100   10.055    0.000    1.002    0.922
##    .x23               1.068    0.109    9.797    0.000    1.068    0.914
##    .x2                0.852    0.081   10.556    0.000    0.852    0.929
##    .x4                0.831    0.156    5.317    0.000    0.831    0.760
##    .x13               0.901    0.082   11.001    0.000    0.901    0.945
##    .x25               0.975    0.115    8.476    0.000    0.975    0.867
##     F1                0.037    0.040    0.911    0.362    1.000    1.000
##     F2                0.065    0.049    1.320    0.187    1.000    1.000
```

```
residuals(fit)
```

```
## $type
## [1] "raw"
##
## $cov
##     x3     x11    x15    x17    x23    x2     x4     x13    x25
## x3   0.000
## x11  0.022  0.000
## x15 -0.033 -0.017  0.000
## x17 -0.057  0.001  0.055  0.000
```

```
## x23  0.066  0.006 -0.011 -0.035  0.000
## x2  -0.012 -0.057  0.051  0.074  0.030  0.000
## x4  -0.057 -0.002 -0.009 -0.002  0.073  0.012  0.000
## x13  0.070  0.005 -0.015 -0.023 -0.040  0.006 -0.011  0.000
## x25  0.076  0.048 -0.024 -0.063 -0.035  0.014 -0.002 -0.022  0.000
##
## $mean
##  x3 x11 x15 x17 x23  x2  x4 x13 x25
##   0   0   0   0   0   0   0   0   0
```

All but one standardized loadings show absolute values $> .20$. All loadings show p-values $< .15$. The factor variances indicate that maybe there are no latent factors underlying the observed variables.

References

Revelle, William. 2017. *psych: Procedures for Psychological, Psychometric, and Personality Research.* Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Rosseel, Yves. 2012. "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software* 48 (2): 1–36. http://www.jstatsoft.org/v48/i02/.