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 Fokkema¹ and Samuel Greiff²

¹Department of Methods and Statistics, University of Leiden, The Netherlands,
²Cognitive Science and Assessment, University of Luxembourg, Esch, Luxembourg

We regularly receive papers at *EJPA* where a principal component analysis (PCA) or an 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 zero, but please note that some values approach .2 and $-.2$, which would be interpreted as a small to medium effect size.

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.

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.

As mentioned before, we see evidence for a two-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 original PCA.

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 section “Some Nitpicking All the Same.”

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

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

For the CFA, we specify two latent factors and specify loadings in accordance with Table 1. That is, Items x3, x11, x15, x17, and x23 are assumed to load on the first factor, whereas Items 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: $\chi^2(26) = 16.925$; $p = .911$; 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

Table 1. Standardized component loadings

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

Boldfaced loadings have absolute values > .40.

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

actually generated from a population model of zero correlations: the factor (co)variances and the p -value clearly indicate that they do not differ significantly from zero.

Interpretation

The little experiment above has shown us that performing PCA and CFA on the same data can indeed have dire consequences: it yields deceptively optimistic model fit indices and parameter estimates. One may wonder: How could we

Table 3. Estimated factor (co)variances

Factor	(Co)variance (<i>SE</i>)	<i>p</i> -value
Factor 1	0.037 (0.040)	.362
Factor 2	0.065 (0.049)	.187
Factor 1 and 2	0.001 (0.008)	.907

obtain such excellent model fit indices with data that were generated as to be uncorrelated? The answer is twofold:

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 known as overfitting, which yields inflated estimates of model fit and parameter estimates. Obviously, we are not the first to write about the topic of overfitting. 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 structural equation modeling (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 published in *EJPA* 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, that a rough visual scree test is not the state of the art when deciding on the number of components 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 yield 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 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 datasets.
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 dataset.
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. <https://doi.org/10.1097/00006842-200405000-00021>
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.
- Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, 5, 155–174. <https://doi.org/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. <https://doi.org/10.1027/1015-5759/a000450>
- 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. <https://doi.org/10.1027/1015-5759/a000419>
- 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. <https://doi.org/10.1177/1745691617693393>
- Ziegler, M. (2014). Comments on item selection procedures. *European Journal of Psychological Assessment*, 30, 1–2. <https://doi.org/10.1027/1015-5759/a000196>

Marjolein Fokkema

Department of Methods and Statistics
University Leiden
Wassenaarseweg 52
2333 AK Leiden
The Netherlands
E-mail m.fokkema@fsw.leidenuniv.nl

Samuel Greiff

Cognitive Science and Assessment
University of Luxembourg
6, rue Richard Coudenhove-Kalergi
4366 Esch-sur-Alzette
Luxembourg
E-mail samuel.greiff@uni.lu