original article

Using Principal Component Analysis to Validate Psychological Scales: Bad statistical habits we should have broken yesterday II

Stefan L.K. Gruijters
Open University of the Netherlands

There is surprisingly little justification to be found in the literature for the use of principal component analysis (PCA) for scale validation purposes – which raises the question why the practice sporadically reappears in the literature. Instead, a large body of literature suggests that PCA is inappropriate in the context of psychological construct validation (e.g., Borsboom, 2006; Edwards & Bagozzi, 2000; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Mulaik, 1990). One reason for PCA's continued appearance in the literature might be because methodological decisions often involve fast and frugal heuristics. In this instance, the continued use of PCA for validation work likely connects to the default heuristic – *if there is a default in the SPSS graphical use interface, do nothing about it* (see Borsboom, 2006). As a result, despite various substantive reasons to prefer alternatives, SPSS default procedures such as PCA continue to be reported in the literature. [1]

In this short paper, I will review some reasons why the use of PCA finds little justification in the context of validating psychological scales. I will make a case for the burgeoning use of better alternatives such as confirmatory factor analysis (CFA). This will amount to two recommendations. The arguments described here are not novel or original (e.g., Borsboom, 2006; Borsboom, Mellenbergh, & van Heerden, 2003; Costello & Osborne, 2005; Fabrigar et al., 1999; Haig, 2005; Mulaik, 1990), but worth discussing given the persistent habit to use PCA for scale validation.

Measurement of latent variables

Many researchers tend to assume that PCA is just a form of factor analysis (such as principal axis factoring), when in fact these are different methods designed for different goals (Fabrigar et al., 1999). In general, there are two different objectives when performing a component or factor analysis: 1) achieving data reduction, and 2) performing a latent variable analysis. PCA is a data reduction method and not a latent variable detection technique, but factor analysis is a latent variable technique (e.g., Borsboom, 2006). For some purposes, such as creating an index variable (e.g., socioeconomic status) from various indicators (e.g., income, education level, etc.), PCA could perhaps be a feasible technique. But when validating *psychological scales*, researchers are interested in testing *latent variables*, and not just in reducing a large number of variables to fewer indices. By extension, this makes PCA inappropriate to use in the context of latent variable analysis, which is involved when validating psychological scales. To make a case for this view, some issues need to be clarified in more depth. First, what are "latent variables" exactly?

There are myriad informal and formal definitions of latent variables (e.g., Bollen, 2002; Borsboom et al., 2003). A colloquial definition holds that latent variables are unobservable, and therefore not directly measurable. Such latent variables are held causally responsible for observable data patterns, including the correlations between items. Most researchers (ideally) have a good conceptual grasp of the latent variable they aim to measure – what it relates to, why people vary on the variable, and what sort of indicators can be used to measure the latent variable. For example, health psychologists working in the socio-cognitive tradition rely heavily on the conceptual model for *attitude* provided by expectancy-value theory, which is embedded within the Theory of Planned Behavior (TPB), Reasoned Action Approach (RAA), and so forth. These approaches assume that attitudes are created on the basis of expectancy-value weighted behavioral beliefs.

Measurement of the latent variable proceeds (indirectly) by responses on observable indicators, sometimes referred to as manifest variables. In the case of attitude, this latent variable is held to manifest itself in responses on semantic differentials (e.g., do you think exercising is good – bad, fun – not fun, important – not important). Often, psychometricians (e.g., Bollen, 2002; Borsboom et al., 2003) assume a causal model underlying measurement of latent variables (see also Gruijters & Fleuren, 2018). That is, the latent variable is conceptualized as a *cause* of response variation on observables – though there are alternative models (e.g., Fleuren, van Amelsvoort, Zijlstra, de Grip, & Kant, 2018). In the TPB for instance, variation on semantic differentials (e.g., 1= bad; 7 = good) is seen to be *caused* by individuals' attitude (it is because of attitude variation that individuals respond differently to semantic differentials). As another example, intelligence is often seen to cause variation on particular IQ-test questions; that is, the variation in test scores *reflect* (is caused by) variation in intelligence. An analogy may further clarify the causal model of measurement – in a sense, psychologists studying latent variables are in the business of estimating the size of an unobservable distant fire, by merely looking at the smoke that rises above the skyline.

The question of validity, then, involves determining whether we are looking at smoke (a scale) that is telling of one particular latent variable (e.g., attitude), or perhaps distinctive ones (e.g., affective and cognitive components), or perhaps something else entirely. A valid instrument is here defined as an instrument that measures what it claims to measure. More specifically, a test (Y) can be said to be a valid measurement of latent variable (X), if the latent variable X exists and is causing variation in item scores on test Y (Borsboom, Mellenbergh, & van Heerden, 2004). The assumption that the latent variable in question exists as a relevant psychological phenomenon is critical (*cf.* Peters & Crutzen, 2017), because one cannot measure them otherwise (Michell, 1999) – in which case, of course, no instrument could provide a valid measurement. One important prerequisite for concluding an instrument taps into an underlying latent variable is unidimensionality (one underlying factor) – because a scale cannot be said to measure attitude (and just attitude) if the data reflect more than one underlying dimension. Of course, the converse (observing unidimensionality) merely provides *evidence for* a valid instrument. It could still be the case that 'schmattitude' rather than attitude was measured. Because of this, the question of validity cannot be answered by solely scrutinizing statistics (Borsboom et al., 2004) – it requires grounding in substantive theory of what an attitude is and what sort of indicators can be used to measure it. Nonetheless, though not sufficient, unidimensionality is a necessary requirement for validity. This can be examined with a latent variable analysis.

Latent variable analysis to test measurement validity

In a latent variable analysis one tries to estimate a number of potential latent variables in an observable response pattern (i.e., an exploratory analysis), or test hypotheses about expected latent variables in a response pattern (i.e. a confirmatory analysis). Given the previously described requirement of unidimensionality, a latent variable analysis for validation purposes needs to assess whether correlations between items can be explained by a single *common cause* (a latent variable). The *principle of local independence* (e.g., Bollen, 2002; Borsboom et al., 2003) allows such a test. Local independence implies the following: If items are measuring a single latent variable (causally responsible for variation in item scores), then factoring out this common cause of variation should (approximately) render the correlations between indicators zero. Conversely, if items still correlate substantially after controlling for the effect of the common cause, then a particular instrument is likely multidimensional. This is somewhat intuitive: If variation on IQ-test items is solely caused by differences in intelligence, then controlling for the influence of intelligence should leave all IQ-test items uncorrelated. So, in order to test local independence we need a statistical procedure that is able to explain the *correlations* between items by involving latent variables as potential common causes.

Both factors and components explain correlations between items to some extent, but component analysis does a poorer job at it because it includes a portion of irrelevant variance in the analysis. Items in a scale have two main variance components, communality (shared variance) and uniqueness (item-unique variance). Shared variance refers to variance which potentially can be explained by reference to a common cause. Item-unique variance refers to variance that cannot be explained by postulating a common cause but rather (as the term suggests) implies unique sources. PCA uses *both* the shared variance and item-unique variance of items to create a number of components (e.g., Fabrigar et al., 1999). For this reason, components do not provide a good explanation of the correlations between items because correlation is *solely* related to the shared variance. Consequently, components account for more than what latent variables are supposed to account for. By including *irrelevant* item-unique variance in the analysis, the result is that components are not adequate representations of latent variables (see also Borsboom, 2006; Borsboom et al., 2003; Costello & Osborne, 2005; Fabrigar et al., 1999; Mulaik, 1990).

Factor analysis reduces the variance-covariance matrix to a number of factors by just using the estimated shared variance of items to do so. This makes factors suitable for use in a latent variable analysis – because the latent variable of interest is supposed to only explain the shared variance of items and not their unique variance. By explaining some of the shared item variance, the factor succeeds to some extent in reproducing the observed correlations between items. A perfect unidimensional model with no measurement error would completely succeed in reproducing the observed correlation between items – these items would be completely locally independent. In practice, factors will never *fully* account for the correlations between items – this left-over bit is usually referred to as residual correlation.

But does the choice of method actually matter?

Despite the differences between components and factors, it seems that often researchers determine the appropriateness of a particular analysis by informal empirical comparison. Does method B usually result in roughly similar numbers compared

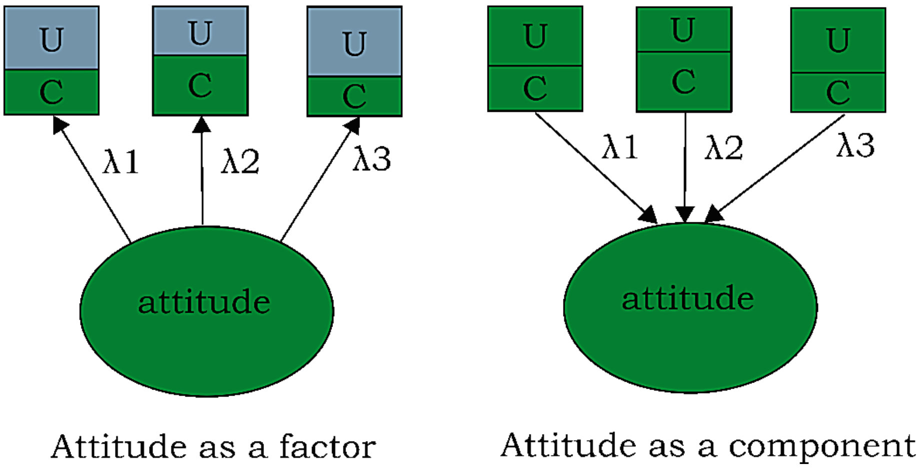

Figure 1. A simplified depiction of a factor and a component. C= communality (shared variance), U=uniqueness (item-unique variance). Left: a one-factor model of attitude. The factor is extracted while making a distinction between communality and uniqueness. Right: a one-component model. The 'attitude' component is distilled from all of the item variance, including the item-unique variance (U).

to alternative or golden-standard method A? If so, then all must be fine with method B. For instance, Field (2009) argues that the methods (component and factor analysis) 'usually result in similar solutions' (*p*. 636/637) and that 'differences arise largely from the calculation' (*p*. 638). Indeed, PCA may not always lead to different conclusions when used as an alternative to factor analysis. But, empirical similarity with factor analysis does not imply that PCA is (conceptually) appropriate for a latent variable analysis[2]. No matter how similar the results of the methods can be, in specific cases the number of estimated components can and will differ from the number of factors (e.g., Fabrigar et al., 1999). Discrepancies such as these, and – of course – because it is impossible to predict beforehand whether the methods will differ, make it worthwhile to have theoretical arguments to strengthen the choice of methodology.

Finally, another reason for not using PCA to validate a scale is because it is an exploratory approach while validation is by definition a confirmatory matter. But, the same critique applies in this instance to using methods such as principal axis factoring and other forms of exploratory factor analysis (EFA) – so for validating scales, both methods ignore hypotheses (a priori beliefs about the factor structure). In instances where researchers have a clear idea about what a scale is supposed to be measuring, there is little justification for using exploratory approaches rather than a confirmatory approach (i.e., confirmatory factor analysis). The difference between EFA and CFA lies in the former using the data to estimate a potential number of factors, whereas the latter uses a hypothesis about the number of factors to test against the data (e.g., Haig, 2005). Naturally, researchers examining the validity of a measurement instrument will have developed an instrument *in line with* a theory or model, specifying how the latent variable can be measured. CFA allows one to specify a model that aligns with the theory, and to test whether the model is feasible given the data. Compared to EFA, CFA thus allows researchers to put the theory before the observation – instead of using theory to aid with post-hoc interpretation of noisy empirical findings. EFA is, for these reasons, best seen as a method to generate theory involving latent variables, whereas CFA is a method to test *a priori* ideas about latent variables (see Haig, 2005).

Conclusion

Factor analysis provides a means to perform a latent variable analysis, because it is well-suited to explain correlations between indicators. Component analysis involves not just shared variance, but also tries to explain variance that is unique to the item. Because latent variables of interest are not supposed to account for item-unique variance, but only the shared variance, PCA is ill-suited for a latent variable analysis (see also Borsboom, 2006; Costello & Osborne, 2005; Fabrigar et al., 1999; Haig, 2005) – and is thus also not suited to validate scales. Additionally, in the context of scale validation, there are no good reasons to take an exploratory 'going in blind' approach when one has a priori beliefs about the factor structure. By specifying a factor structure to be tested in a CFA, one is in a position to use theory to guide empirical tests rather than vice versa.

Adequate measurement is a prerequisite for replicable research – this makes it important for researchers to assess the quality of their measurements using appropriate procedures. Two recommendations for research in health psychology follow: 1) do not resort to PCA for latent variable analysis and scale validation specifically, and 2) use CFA to test measurement hypotheses rather than EFA.

Footnotes

1. Some note that a component analysis is computationally less demanding than a factor analysis. Before the advent of modern computers it was more feasible to conduct a PCA, which may be one historical factor explaining its initial popularity (see Costello & Osborne, 2005).

2. Another illustration of the 'empirical similarity' argument can be found in discussions surrounding the (mis)use of coefficient alpha to estimate reliability. Defenders of coefficient alpha often point out that in many contexts, despite making some (usually) unrealistic assumptions, alpha closely approximates other internal consistency indices. By extension, it is argued, the choice between alpha and its alternatives must be trivial. This is problematic reasoning for researchers who do arrive at different conclusions with regard to internal consistency, depending on the index that was used.

Acknowledgements

The author is thankful to Bram Fleuren, Ilja Croijmans, and Gjalt-Jorn Peters for stimulating comments on aspects of this paper, but the author is responsible for any errors or omissions.

References

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605–34. http://doi.org/10.1146/annurev.psych.53.100901.135239

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425–440. http://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203–219. http://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review, 111*(4), 1061–1071. http://doi.org/10.1037/0033-295X.111.4.1061

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*. Retrieved from https://pareonline.net/getvn.asp?v=10&n=7

Edwards, J. R., & Bagozzi, R. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–174. http://doi.org/10.1037//1082-989X.5.2.155

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299. http://doi.org/10.1037/1082-989X.4.3.272

Field, A. (2009). *Discovering statistics using SPSS* (third edition). London: Sage Publications, Inc.

Fleuren, B. P. I., van Amelsvoort, L. G. P. M., Zijlstra, F. R. H., de Grip, A., & Kant, Ij. (2018). Handling the reflective-formative measurement conundrum: a practical illustration based on sustainable employability. *Journal of Clinical Epidemiology, 103*, 71–81. http://doi.org/10.1016/j.jclinepi.2018.07.007

Gruijters, S. L. K., & Fleuren, B. P. I. (2018). Measuring the Unmeasurable: The Psychometrics of Life History Strategy. *Human Nature, 29*(1), 33–44. http://doi.org/10.1007/s12110-017-9307-x

Haig, B. D. (2005). Exploratory Factor Analysis, Theory Generation, and Scientific Method. *Multivariate Behavioral Research, 40*(3), 303–329. http://doi.org/10.1207/s15327906mbr4003_2

Michell, J. (1999). *Measurement in psychology*. Cambridge, U.K.: Cambridge University Press.

Mulaik, S. A. (1990). Blurring the Distinctions Between Component Analysis and Common Factor Analysis. *Multivariate Behavioral Research, 25*(1), 53–59. http://doi.org/10.1207/s15327906mbr2501_6

Peters, G.-J. Y., & Crutzen, R. (2017). Pragmatic Nihilism: How a Theory of Nothing can Help Health Psychology Progress. *Health Psychology Review*, 1–39. http://doi.org/10.1080/17437199.2017.1284015

Stefan L.K. Gruijters
Open University of the Netherlands
General Psychology, Faculty of Psychology & Education Science
mail@stefangruijters.nl