

Statistics Corner

Questions and answers about language testing statistics:

Principal components analysis and exploratory factor analysis— Definitions, differences, and choices

James Dean Brown
University of Hawai'i at Manoa

QUESTION: In Chapter 7 of the 2008 book on heritage language learning that you co-edited with Kimi Kondo-Brown, there is a study (Lee and Kim, 2008) comparing the attitudes of 111 Korean heritage language learners. On page 167 of that book, a principal components analysis (with varimax rotation) describes the relationships among 16 purported reasons for studying Korean with four broader factors. Several questions come to mind. What is a principal components analysis? How does principal components analysis differ from factor analysis? What guidelines do researchers need to bear in mind when selecting "factors"? And finally, what is a varimax rotation, and why is it applied?

ANSWER: This is an interesting question, but a big one, made up of at least three sets of sub-questions: (a) What are principal components analysis (PCA) and exploratory factor analysis (EFA), how are they different, and how do researchers decide which to use? (b) How do investigators determine the number of components or factors to include in the analysis? (c) What is rotation, what are the different types, and how do researchers decide which to use? And, (d) how are PCA and EFA used in language test and questionnaire development? I will address the first one (a) in this column. And, I'll turn to the other three in subsequent columns.

What Are Principal Components Analysis and Exploratory Factor Analysis?

Principal components analysis (PCA) and exploratory factor analysis (EFA) are often referred to collectively as *factor analysis* (FA). The general notion of FA includes "a variety of statistical techniques whose common objective is to represent a set of variables in terms of a smaller number of hypothetical variables" (Kim & Mueller, 1978, p. 9). A more elaborate definition is provided by Tabachnick and Fidell (2007, p. 607):

... statistical techniques applied to a single set of variables when the researcher is interested in discovering which variables in the set form coherent subsets that are relatively independent of one another. Variables that are correlated with one another but largely independent of other subsets of variables are combined into factors.

In the study you mentioned, Lee and Kim (2008) looked at the attitudes expressed by 111 heritage and traditional learners of Korean, and then performed a PCA (with varimax rotation) on the results. The participants answered a 34-item questionnaire with both Likert-scale and open-ended questions. The PCA was used to analyze the results for 16 of the Likert-scale items on motivations for studying Korean. The researchers found that four broad factors underlay the relationships among the participants' responses to these items. How researchers go about deciding on the number of factors and why they decide to use a particular kind of rotation will be addressed in subsequent columns. However, that these researchers did find four components is evident in Table 1.

Notice in Table 1 that the wording of each of the Likert-scale items is given in the first column. The next four columns are labeled Components 1, 2, 3, and 4, and each column shows values that look suspiciously like correlation coefficients (positive and negative); that's because they are correlation coefficients. The analysis has actually generated a set of four new predicted values for each participant—one value for each of the four components (in essence these are four new hypothetical variables, Components 1, 2, 3, and 4). These values are called component scores, and they can be saved as data if the researcher so wishes. The correlation coefficients in Table 1 are the correlations between all participants' Likert-scale answers for each item (or *variable*, as they are called in FA), and these component scores. For example, the correlation between their Likert-scale

answers to the “I learn Korean to transfer credits to college” item and their component 1 scores is 0.83—a fairly high correlation, wouldn’t you say? In contrast, the correlations of those same Likert-scale answers and the Component 2, 3, and 4 scores are very low. Does that make sense? The remaining correlation coefficients can be interpreted in similar manner.

Table 1 Principal Components Analysis (with Varimax Rotation) Loadings of Motivation Items (adapted from Lee & Kim, 2008¹)

	Instrumental		Integrative		<i>h</i> ²
	Component 1: School-related	Component 2: Career-related	Component 3: Personal fulfillment	Component 4: Heritage ties	
I learn Korean to transfer credits to college.	0.83	0.01	0.17	-0.11	0.72
I learn Korean because my friend recommended it.	0.82	0.14	0.06	0.22	0.74
I learn Korean because my advisor recommended it.	0.80	0.10	0.36	0.04	0.78
I learn Korean because of the reputation of the program and instructor.	0.77	0.12	0.12	0.22	0.67
I learn Korean for an easy A.	0.69	0.18	-0.29	0.16	0.61
I learn Korean to fulfill a graduation requirement.	0.63	0.11	0.19	-0.18	0.47
I learn Korean to get a better job.	0.04	0.80	0.20	0.11	0.69
I learn Korean because I plan to work overseas.	0.26	0.80	0.11	0.10	0.75
I learn Korean because of the status of Korean in the world.	0.10	0.73	0.20	0.22	0.63
I learn Korean to use it for my research.	0.38	0.48	0.44	0.10	0.58
I learn Korean to further my global understanding.	0.16	0.33	0.71	0.08	0.65
I learn Korean because I have an interest in Korean literature.	0.18	0.04	0.64	0.13	0.56
I learn Korean because it is fun and challenging.	0.04	0.04	0.63	0.46	0.63
I learn Korean because I have a general interest in languages.	0.11	0.03	0.57	0.51	0.59
I learn Korean because it is the language of my family heritage.	-0.01	0.26	0.05	0.80	0.71
I learn Korean because of my acquaintances with Korean speakers.	0.10	0.21	0.20	0.70	0.58
% of variance explained by each factor	0.23	0.15	0.14	0.12	0.64

Extraction Method: Principal Component Analysis. Rotation Method: Varimax; Eigenvalue>1.0

Some of the correlation coefficients in Table 1 are in bold-faced italics in a larger font to emphasize them. For example, in the Component 1 column, the first six correlations (by convention, these are called *loadings* in FA) of .63 to .83 are emphasized because they are much higher than the other loadings in that same column. Similarly, the Component 2 loadings of .48 to .80 are highlighted, the Component 3 loadings of .57 to .71 are accentuated, and the Component 4 loadings of .70 and .80 are emphasized. For each of the four components, the variables with loadings that are much higher than the others in the same column are of particular interest because they are for the variables that are most highly related to the component scores.

How high is a high loading? Well, obviously, as correlation coefficients, they can range from 0.00 to 1.00 and 0.00 to -1.00, with the sign depending on the direction of the relationship. The reader can decide whether the values reported in a particular study are adequate. However, loadings below 0.30 are typically ignored in such analyses. In the study reported in Table 1, it appears that the researchers decided that a better cut-point would be 0.40 (i.e., there are values above 0.30 but below 0.40, which are not emphasized) for deciding which loadings should be interpreted. It is up to the researcher to decide on the cut point and up to the readers to decide whether they buy that cut point.

The researcher also interprets the patterns found in such analyses—a fact that can become a problem for FA. Researchers risk seeing only those patterns they want to see because they are free to interpret the results any way they like. As a result, it is particularly important that researchers be transparent in explaining how they made these decisions, and that readers carefully examine the researchers’ interpretations to make sure those interpretations make sense and are believable.

Consider Component 1 in Table 1, which is labeled “school-related.” Have a look at the first six items on the questionnaire (i.e., those loading heaviest on Component 1). Are those questions really school-related? I suppose if the “friend” is a school friend, those six questions can truly be said to all be school-related? Now, what do you think of the four

¹ Note that the authors had Components 1-4 labeled as Factors 1-4. I have changed them here to be consistent with the fact that they were performing a principal *components* analysis. Also I added the bold-faced italics (larger font) for emphasis.

items for “career-related” Component 2? Are the four items for Component 3 all related to “personal fulfillment”? Are those loading heavily on Component 4 related to “heritage ties”? I think these interpretations are pretty good,² but what do you think? That’s important too.

There are additional numbers around the edges of Table 1 that are also worth considering. In the column furthest to the right, there are *communalities* (h^2). Each of these values tells us the proportion of variance accounted for the particular variables in that row by the four components in this analysis. For instance, the 0.72 at the top right indicates that 72% of the variance in the “I learn Korean to transfer credits to college” variable is accounted for by the four components in this analysis. Clearly, this analysis is much better at accounting for the variance in some variables than in others. Which variable has the highest communality? Which has the lowest? Why is the relative value of these communalities important? It’s important because those variables with relatively high communalities are being accounted for fairly well, while those with low ones are not.

Across the bottom of Table 1, the following numbers represent the proportion of variance accounted for by each component: 0.23, 0.15, 0.14, and 0.12. These indicate that Component 1 accounts for 23% of the variance, Component 2 accounts for 15%, Component 3 accounts for 14%, and Component 4 accounts for 12% of the variance. The last number at the bottom right of Table 1 (0.64) indicates the total proportion of variance accounted for by the analysis as a whole. In other words, 64% (or just shy of 2/3rds of the variance) was accounted for by this analysis. This total proportion of variance can be calculated by either adding up the individual proportions of variance accounted for by each of the four components, or by averaging the communalities.

How are PCA and EFA Different?

Calculations for both PCA and EFA involve matrix algebra as well as matrices of Eigen vectors and Eigen values. Any explanation of this would be quite involved and not particularly enlightening for most readers of this column, so suffice it to say that both PCA and EFA depend on calculating and using matrices of Eigen vectors and values in conjunction with a matrix of the correlation coefficients all of which are based on the variables being studied.

The difference between PCA and EFA in mathematical terms is found in the values that are put in the diagonal of the correlation matrix. In PCA, 1.00s are put in the diagonal meaning that all of the variance in the matrix is to be accounted for (including variance *unique* to each variable, variance *common* among variables, and error variance). That would, therefore, by definition, include all of the variance in the variables. In contrast, in EFA, the communalities are put in the diagonal meaning that only the variance shared with other variables is to be accounted for (*excluding* variance unique to each variable and error variance). That would, therefore, by definition, include only variance that is common among the variables.

How do Researchers Decide Whether to Use PCA or EFA?

The difference between PCA and EFA in conceptual terms is that PCA analyzes variance and EFA analyzes covariance (Tabachnick and Fidell, 2007, p. 635). Thus when researchers want to analyze only the variance that is accounted for in an analysis (as in situations where they have a theory drawn from previous research about the relationships among the variables), they should probably use EFA to exclude unique and error variances, in order to see what is going on in the covariance, or common variance. When researchers are just exploring without a theory to see what patterns emerge in their data, it makes more sense to perform PCA (and thereby include unique and error variances), just to see what patterns emerge in all of the variance.

² My one reservation is that, in their interpretation and discussion, the authors overlooked the *complex* variables which loaded on two or more components: “I learn Korean to use it for my research” and “I learn Korean because I have a general interest in languages.”

For purposes of illustration, I will use data based on the 12 subtests of the *Y/G Personality Inventory (Y/GPI)* (Guilford and Yatabe, 1957) which are: social extraversion, ascendance, thinking extraversion, rathymia, general activity, lack of agreeableness, lack of cooperativeness, lack of objectivity, nervousness, inferiority feelings, cyclic tendencies, and depression. The first six scales have been shown to be extraversion measures; the last six scales have been shown to be neuroticism measures (for definitions and more information on the Y/GPI, see Robson, 1994; Brown, Robson, and Rosenkjar, 2001). The data used for this illustration are based on an English language version administered for comparison purposes to 259 students at two universities in Brazil. The descriptive statistics for this sample are shown in Table 2.

Table 2. *Descriptive Statistics for the 12 Y/GPI Scales Administered to University Students in Brazil*

Trait	M	SD	N
Social extraversion	6.56	3.62	259
Ascendance	10.10	3.71	259
Thinking extraversion	12.33	2.72	259
Rhathymia	10.33	3.73	259
General activity	5.42	3.79	259
Lack of agreeableness	7.06	2.72	259
Lack of cooperativeness	10.47	3.43	259
Lack of objectivity	9.45	3.41	259
Nervousness	11.65	4.96	259
Inferiority feelings	9.52	3.85	259
Cyclic tendencies	11.38	3.93	259
Depression	12.99	4.63	259

Table 3. *PCA and EFA (with Varimax Rotation) Loadings for the 12 Y/GPI Scales Administered in Brazil*

Variables	Rotated PCA Eigenvalues ≥ 1.00				Rotated EFA Eigenvalues ≥ 1.00			
	Comp 1	Comp 2	Comp 3	h^2	Factor 1	Factor 2	Factor 3	h^2
Social extraversion	-0.139	0.744	-0.140	0.592	-0.135	0.665	-0.118	0.474
Ascendance	-0.109	0.658	-0.099	0.455	-0.109	0.548	-0.098	0.321
Thinking extraversion	-0.091	-0.053	0.916	0.851	-0.072	-0.026	0.530	0.287
Rhathymia	0.419	0.644	0.221	0.639	0.391	0.606	0.206	0.562
General activity	-0.227	0.746	-0.068	0.613	-0.219	0.680	-0.070	0.515
Lack of agreeableness	0.150	0.638	0.304	0.522	0.119	0.540	0.197	0.345
Lack of cooperativeness	0.562	0.054	0.171	0.348	0.466	0.029	0.060	0.221
Lack of objectivity	0.693	0.067	-0.247	0.546	0.614	0.044	-0.187	0.415
Nervousness	0.798	-0.177	-0.022	0.669	0.759	-0.171	-0.021	0.606
Inferiority feelings	0.695	-0.481	0.085	0.722	0.677	-0.469	0.101	0.689
Cyclic tendencies	0.820	0.110	-0.014	0.685	0.785	0.106	-0.005	0.627
Depression	0.812	-0.232	-0.058	0.716	0.784	-0.227	-0.076	0.672
Proportion of Variance	0.295	0.225	0.093	0.613	0.259	0.182	0.037	0.478

Table 3 shows PCA and EFA analyses (with varimax rotation) and the resulting loadings for the Y/GPI administered in Brazil. Notice that the first column contains labels for the 12 scales. Then the next four columns show the results for a PCA of the data, and the last four columns show analogous results for an EFA of the same data. Notice that the patterns are very clear in both cases, but that the actual loadings differ for the PCA and EFA. Note also that the patterns of relatively strong loadings are the same for both analyses, so in that sense, it made little difference which analysis was used. However, notice also that including all of the variance in the PCA produced generally higher loadings, higher communalities, and ultimately accounted for more variance overall (61.3% as opposed to 47.8%) than the EFA (which excluded the unique and error variances). The comparison of these two analyses indicates that the unique variances (and perhaps error variances) of the variables, which are used in the PCA, are contributing to higher loadings with the components in ways that are not present in the EFA. That is, of course, worth thinking about.

In sum, the primary differences between PCA and EFA are that (a) PCA is appropriate when researchers are just exploring for patterns in their data without a theory and therefore want to include unique and error variances in the analysis, and EFA is appropriate when researchers are working from a theory drawn from previous research about the relationships among the variables and therefore want to include only the variance that is accounted for in an analysis (thereby excluding unique and error variances) in order to see what is going on in the covariance, or common variance.

Basically, researchers tend to: (a) use PCA if they are on a fishing expedition trying to find patterns in their data and have no theory to base the analysis on, or (b) use EFA if they have a well-grounded theory to base their analysis on. Generally, the second strategy is considered to be the stronger form of analysis.

Conclusion

I have shown what PCA and EFA (collectively known as factor analysis or FA) are, and in part, how they should be presented and interpreted. In the process, I have defined and exemplified loadings, communalities, proportions of variance, components, factors, PCA, and EFA. I have also explored the basic mathematical and conceptual differences between PCA and EFA, and discussed how researchers decide on whether to use PCA or EFA. However, much about FA has been left unexplained. How do researchers decide the number of components or factors to include in the analysis? For instance, how did I decide on the three components and factors shown in Table 3? Also, what is rotation, what are the different types, and how do researchers choose which type to use? For instance, what is the varimax rotation mentioned in Tables 1 and 3 (and the associated text), and why did the researchers choose it? As I mentioned above, I will address these issues in two subsequent columns.

References

- Brown, J. D., Robson, G., & Rosenkjar, P. (2001). Personality, motivation, anxiety, strategies, and language proficiency of Japanese students. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 361-398). Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai'i Press.
- Guilford, J. P., & Yatabe, T. (1957). *Yatabe-Guilford personality inventory*. Osaka: Institute for Psychological Testing.
- Kim, J. O., & Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it*. Beverly Hills, CA: Sage.
- Lee, J. S., & H. Y. (2008). In K. Kondo-Brown & J. D. Brown (Eds.), *Teaching Chinese, Japanese, and Korean heritage language students*. New York: Lawrence Erlbaum Associates.
- Robson, G. (1994). *Relationships between personality, anxiety, proficiency, and participation*. Unpublished doctoral dissertation, Temple University Japan, Tokyo, Japan.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Upper Saddle River, NJ: Pearson Allyn & Bacon.