

Title:

Principle Component and Factor Analysis; Analytic Strategies to Increase Content Validity of Questionnaire Factors.

Authors:

Manouchehr Afshinnia: Head of Department of Statistics, Isfahan University, Iran.

Farsad Afshinnia: Resident of Internal Medicine, Brookdale University Hospital and Medical Center, Brooklyn, New York.

Corresponding author:

*Farsad Afshinnia MD
7 Hegeman Avenue, Apt 6C
Brooklyn, NY 11212
Email: afshinnia@hotmail.com*

Key Words:

Content Validity, Questionnaire, Factor Analysis, Principle Component.

Summary:

Background:

In most of the questionnaire based surveys and KAP studies, there are different topics of interest. Most of the times, information on each topic is gathered by more than one question. As a result a large number of questions may be found in a questionnaire. However, the answers to some questions related to a particular topic of a questionnaire may actually be more informative or even present a better structural model of data collection in other topics of the same questionnaire.

Objective:

To use principle component and factor analysis as analytic strategies to generate new factors, including set of questions with better content validity.

Methods:

This analytic strategy is based on the highest correlation coefficients between questions within a proposed factor, and the lowest coefficients among questions between different factors. The criteria of using factor analysis in this study are: 1- Lack of multicollinearity, 2- Lack of singularity of any couple of variables, 3- Kaiser-Mayer-Okin (KMO) statistic greater than 0.5 and 4- Significant level of Bartlett test of sphericity. The method of factor component determination is varimax rotation.

Conclusion:

After conducting the analysis on questions of a questionnaire, the number of questions in each factor may change in decreasing or increasing order so that, better content validity may be achieved in each factor.

Background:

Most of the times, information on each topic of a questionnaire is gathered by more than one question. As a result a large number of questions may be found in a questionnaire. However, the answers to some questions related to a particular topic of a questionnaire may actually be more informative or even present a better structural model of data collection in other topics of the same questionnaire. Furthermore, evaluation of content validity of questionnaire factors is based on the comments of experts in that particular field. For these reasons, there is always uncertainty about relative adequacy of this validity due to lack of quantitative measures. The objective of this study is to use principle component and factor analysis as analytic strategies to generate new factors, including set of questions with better content validity.

Validity:

Now a days, data collection of a large number of researches in different divisions of social and human sciences are based on questionnaires. There are different types of questionnaires. Some of them are self administered, some are filled by an interviewer, some consist of open questions, while others may include closed ended questions. However, in any type of standard questionnaires, common certain features are essential which are reliability and validity. Reliability of a questionnaire is defined as reproducibility of it and can be calculated by different methods such as, Cronbach's alpha, Split half, Guttman, Parallel and Strict Parallel methods (1,2). Their calculation is beyond the scope of this manuscript.

Validity is defined as the accuracy and reliability of a test or a questionnaire, generally (i.e. the extent to which the test or the questionnaire measures what it is supposed to measure) (3). However, statistical, internal, external and construct validities have been defined as more specified types of validity since 1976. Statistical validity indicates that the relationship between variables are not simply by accident or by chance (4,5). External validity shows if the test, variable or the questionnaire can evaluate or be representative of what is supposed to be measured in general population or similar groups (6). For example how much fasting plasma glucose level is representative of metabolic control of diabetes. On the other hand, internal validity shows how much the measurements or observations are representative of the actual values which the tests, variables or questionnaires are measuring (7). For example how the observed value of fasting plasma glucose is representative of actual plasma glucose. Construct validity includes face, criterion and content validity (11). Face validity shows whatever understood from the questionnaire is the same by anybody who reads it (8). Criterion validity presents the degree of similarity between the results of a test or a questionnaire when compared with similar other questionnaires (9,10). Content validity indicates the whole spectrum of knowledge, attitude, beliefs and behaviors which are covered by questionnaire (6,12).

General view:**a. Principle component :**

In social sciences, a single variable is rarely able to define a particular aspect of human or a society, therefore it is needed to evaluate these aspects

by variety of questions. In this case, the questions are not necessarily the same but evaluate a particular aspect of a factor which may not be evaluated or be just partially evaluated by other questions in that factor. As a result, a factor which evaluates a particular feature, consists of variables with significant correlations so that, it is a better representative of that feature compared with each question alone. Principle component is a linear combination of the original variables which produces a single summary (factor) containing information from all variables. Principle components are generated by two requirements placed on the linear combinations. They are: 1- The variance of the first principle component is greatest, the second has the next largest, the third the next largest and so forth. 2- At the same time each principle component generates transformed values that are uncorrelated with the values generated by any other principle component (13).

Formally, a principle component is a linear combination of k multivariate measurements and theoretically there are as much principle components as variables in a set of data. So, the first, second and third principle components are shown as:

$$\begin{aligned} P_{(1)} &= a_1x_1 + a_2x_2 + \dots + a_kx_k && \text{(Equation series 1)} \\ P_{(2)} &= b_1x_1 + b_2x_2 + \dots + b_kx_k \\ P_{(3)} &= c_1x_1 + c_2x_2 + \dots + c_kx_k \end{aligned}$$

The criteria of maximum variability and zero correlation produce the following equations for the first principle component:

$$\begin{aligned} a_1S^2_1 + a_2S^2_{21} + a_3S^2_{31} + \dots + a_kS^2_{k1} &= a_1\lambda_1 && \text{(Equation series 2)} \\ a_1S_{12} + a_2S^2_2 + a_3S^2_{32} + \dots + a_kS^2_{k2} &= a_2\lambda_1 \\ &\vdots \\ a_1S_{1k} + a_2S^2_{2k} + a_3S^2_{3k} + \dots + a_kS^2_k &= a_k\lambda_1 \end{aligned}$$

in which S^2 is variance of the first variable. These equations are usually solved by computer to produce values for coefficients a_1, a_2, \dots, a_k , making it possible to calculate the first principle component. This process is repeated to produce second, third and any number of principle components up to a maximum of k , the number of variables per observation. The values of $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_k$ are also produced and play an important role, because these values indicate the variance associated with each principle component (i.e. $\lambda = S^2_p$). An individual principle component summarize some but not all of the information contained in the original multivariate measurement. Therefore other principle components are used to explain the remaining variability of the model as much as possible. The worth of each summary is directly related to its variability so that, the greater its variance the more information in its content. Usually, the worth of each principle component is judged by the proportion of total variation of multiple variables. It is:

$$100 \lambda_1 / (\lambda_1 + \lambda_2 + \dots + \lambda_k) \quad (\text{Equation 3})$$

So the percent of the variability explained by the first m principle components is:

$$100 \sum_{k=1}^m (\lambda_k / \lambda) \quad (\text{Equation 4})$$

in which λ is the total variance of the variables.

b. Factor analysis:

Factor analysis models are based on equations:

$$y_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{iv}F_v + e_i \quad \text{for } i = 1, 2, \dots, p \quad (\text{Equation 5})$$

where F_1, F_2, \dots, F_v are the latent variables or common factors with the number of factors v being less than the number of variables p . The term e_i is an unexplained component. The basic premise of factor analysis is that the measured variables are correlated because they are constructed from common components. The greater the contribution from common components, the higher will be the correlation among the measured variables. It is the structure of matrix of correlations among the measured variables that provides the information from which the number of common factors and weighting a_{ij} are determined (16).

c. Principle component vs. Factor analysis:

Most parts of the analyses and computations are the same. Both techniques attempt to find a low-dimensional representation of multivariate data to aid in understanding the structure of data. Principle component analysis is, however, essentially a simple rotation of the axis of the multivariate data, whereas factor analysis postulate a statistical model to explain the observed correlations or covariances. In many cases in practice, however, both approaches will lead to similar results (17).

d. Applications:

From what is described above, the following possibilities are three different applications of principle components and factor analysis:

- 1- The first principle component can be used as a univariate summary of original multivariate measurements.
- 2- The principle component coefficients can detect meaningful combinations of the original multivariate measurements.

- 3- The first several principle components can be employed as a graphical tool to search for natural clustering of multivariate observations, or similarly, as a way to explore k -variable data for extreme values.

Methods:

After designing the questionnaire, a pilot study should be done to gather enough data to begin the analyses. The following steps and roles are considered while analyzing (14, 15):

In the first step correlation matrix should be calculated. Then singularity of the correlation matrix has to be considered. No two variables should be exactly the same. If two variables, say X_1 and X_2 are the same, it is difficult to estimate their coefficients in a regression line or in principle components, because for any number of X_1 , its coefficient is exactly the same as X_2 , if they are calculated separately. However, if we use both in a model we see that the coefficients are not uniquely determined. In real life, it suggests that the two questions of a questionnaire may be the same but in different formats and one can be deleted. So the whole matrix has to be singular. For that reason, Determinant of correlation matrix has to be greater than zero.

Lack of multicollinearity: If two or more variables are highly correlated it is difficult to estimate their coefficients in a principle component, because of the above reason. In real life, it means that in a series of highly correlated questions of a questionnaire, they may gather similar information and one question can be used instead.

Kaiser-Mayer-Okin (KMO) statistic has to be greater than 0.5. The KMO measure of sampling adequacy tests whether the partial correlations among variables are small. Value greater than 0.5 indicates sampling adequacy.

Significant level of Bartlett test of sphericity. Bartlett's test of sphericity tests whether the correlation matrix is an identity matrix, which would indicate that the factor model is inappropriate. So, when the result is not statistically significant, there is no correlation between the variables and no further analysis can be done.

Standardization: If the scales of the answers to different questions are the same then the absolute values can be used. For example upon using a questionnaire with Likert spectrum from 1 to 10. However if the questions are not homogenous regarding their scales, the variables have to be standardized with mean = 0 and standard deviation = 1, which make the analyses and interpretations so much easier.

It is hard to say how many factors is the appropriate number of factors for the analysis. However, the variables of a factors with $\lambda_k < 0.1$ λ usually do not gather together to be representative of a particular subject, because their weak correlation with each other. A rough change in the slope of scree plot is another suggestion to show appropriate number of the factors. Scree plot, plots λ of each factor vs. their number. Usually the factor with the greatest drop of λ is the last one.

Upon using different methods of factor extraction including principle component, unweighted least squares, generalized least squares, maximum likelihood, principle axis factoring and alpha factoring, there is not any particular preference for using any of them. Most of the times similar results are achieved with any approach. However, at the end of analysis with each method, one may prefer a particular

method because of more informative structure of the factors, if they differ in different methods of extraction.

After initial extraction of the factors, it is needed to adjust the model of each factor so that, the greatest correlation of each variable with each factor would be achieved. One of the most satisfactory methods is varimax. Varimax uses the idea of maximizing the sum of the variances of the squares of loadings of the factors. Other methods are quartimax, equamax, promax and direct oblimin which do not present the same results, necessarily. At the end, content of each factor is determined by the highest factor score coefficients in factor score coefficient matrix.

Discussion:

A prime use of factor analysis has been in the development of both the theoretical constructs for an area and the operational representatives for the theoretical construct. In other words, a prime use of factor analysis requires reifying the factors i.e. converting them into or to regard them as a concrete thing (14). In psychology, how would one deal with an abstract concept such as aggression? On a questionnaire a variety of possible “aggression” questions may be used. If most or all of them have high loadings on the same factor, and other questions thought to be unrelated to aggression had low loadings, one might identify that factor with aggression. Further the highest loadings might identify operationally the questions to be used to examine this abstract concept, even if there is no actual relationship between the question and the area of the concept. Since most of the times, our knowledge is of the original variables, without a unique set of variables loading a factor, interpretation is difficult. So, we have not to forget that factor analysis is not everything and is not the only source of determination of construct of an abstract concrete. The factor must only be interpreted by an individual with extensive background in the substantive area, because a particular question might not be an appropriate member of a factor, even if the analysis say so.

Similarly, not every factor which is suggested by factor analysis is worthy or interpretable. These factors are made because of effect of particular variables such as, time, place or special events in a portion of population which is not representative of the general population. They present a biased relationship rather than presenting actual construct of a factor. For example, forming a factor including wearing brown shoes and chewing gum due to their association, because in one of two companies where the KAP study is done brown shoes and gum are given to the employees for free.

The idea of using factor analysis, necessitates enough sample size through a pilot study. This impose a great limitation, especially when the number of questions in the questionnaire is too much. Because the higher the number of questions the greater should be the sample size to provide multivariate analysis. However, it is a good idea, if one is trying to make a standard questionnaire for sequential studies, particularly when cases from other populations and more other variables are used. In this case the analysis may be even strengthened.

Factor analysis does not change the content of the whole questionnaire. It may actually change the content of each factor or even suggests new factors due to new classification of the observed variables of a questionnaire. Theoretically, the first factor has the greatest variance. In another word, it has the most informative structure of the whole questionnaire. So, it may be a violation to the content validity of other

factors of a questioner, if not interpreted holistically. Regarding different methods of extraction and rotation, there is not any particular role of preference, but one may use any of them just based on comparing results of different methods.

In summary, factor analysis is a valuable tool in pilot studies of questionnaire based researches or even in sequential studies to make standard factors with appropriate content validity.

References:

- 1- Cronbach LJ. Coefficient alpha and internal structure of tests. *Psychometrika*. 1951;16:297-334.
- 2- Maxim PS. *Quantitative research methods in the social sciences*. Oxford. 1999; pp 233-250.
- 3- Robert G. Sharrar. *General principles of epidemiology*. In: Brett J. Cassens. *Preventive medicine and public health*. 2nd edition. Williams and Wilkins Inc. 1992; pp 1-28.
- 4- Kvale S. *Interviews. An introduction to qualitative research interviewing*. SAGE. 1996; pp 210-228.
- 5- Mc-Burney DH. *Research methods*. 3rd edition. Brooks/Cole. 1994; pp 119-140.
- 6- Wakefield JF. *Educational psychology, learning to be a problem solver*. Houghton Mifflin Company. 1996; pp 591-638.
- 7- Hulley SB, Cumminngs S. *Designing clinical research*. 1995; pp 63-85.
- 8- Peers I. *Statistical analysis for education and psychology researches*. Flamer press. 1996; pp 1-16.
- 9- Samuda RJ, Feuerstein R, Kaufman AS, Lewis JE, Sternberg RJ and Associates. *Advances in cross-cultural assessment*. SAGE. 1998; pp56-99.
- 10- Neuman WL. *Social research methods*. 3rd ed. ALLYN and BACON. 1997; pp 131-175.
- 11- Osterlin SJ. *Constructing test items: Multi-choice constructed response, performance and other formats*. 2nd edition. KAP. 1997; pp 59-105.
- 12- Armitage P, Berry G. *Statistical methods in medical research*. Blackwell. 2nd ed. 1987; pp 296-357.
- 13- Selvin S. *Practical biostatistical methods*. Duxbury press. 1995; pp 221-245.
- 14- Fisher LD, Van Belle G. *Biostatistics, A methodology for the health sciences*. John Wiley & Sons. 1993; pp 496-595 & 692-762.
- 15- Afshinnia F. *Applied data analysis*. Isfahan University of Medical Sciences Press. 1999; pp 189-203.
- 16- McPherson G. *Applying and interpreting statistics*. 2nd ed. Springer. 2001; pp 449-511.
- 17- Everitt B, Rabe-Hesketh S. *Analyzing medical data using S-Plus*. Springer. 2001; pp 381-400.