

Exploratory factor analysis and principal component analysis in clinical studies: Which one should you use?

1 | INTRODUCTION

Factor analysis covers a range of multivariate methods used to explain how underlying factors influence a set of observed variables. When research aims to identify these underlying factors, exploratory factor analysis (EFA) is used. In contrast, when the aim is to test whether a set of observed variables represents the underlying factors, in accordance with an existing conceptual basis, confirmatory factor analysis is performed. EFA has many similarities with a commonly used data reduction technique called principal component analysis (PCA). These similarities, along with using the related terms *factor* and *component* interchangeably, contribute to confusion in analysis. The difficulty in identifying the appropriate use of statistical methods and their application and interpretation impacts clinical and research implications (Beavers et al., 2013; Tabachnick & Fidell, 2001). We acknowledge previous articles in nursing journals offering guidance on the use of factor analysis (Gaskin & Happell, 2014; Watson & Thompson, 2006).

Exploratory factor analysis and PCA are commonly used techniques to express multivariate data with fewer dimensions. The aim of these techniques is to summarize a set of original variables into a smaller set of factors or components that maximize the possible information and variation from the data in the original variables (Meyers, Gamst, & Guarino, 2013). EFA focuses on interrelationships between variables, and hence covariance is used to identify factors, while PCA uses the variance to identify components. In this editorial, we identify some essential methodological considerations that must be taken into account when using these techniques and compare their application using the examples of 'hospitalization stress' and 'hospitalization-related stressors'.

2 | PRINCIPLES OF EFA AND PCA

Exploratory factor analysis is a statistical technique used to simplify complex datasets by examining the pattern of correlations (or covariances) among observed variables (Kline, 1994). EFA is particularly useful in investigating complex concepts which are not easily measurable such as mental health and quality of life. EFA includes the concept of a *latent factor* that exerts influence on observed variables (Basto & Pereira, 2012). The aim is to concisely represent

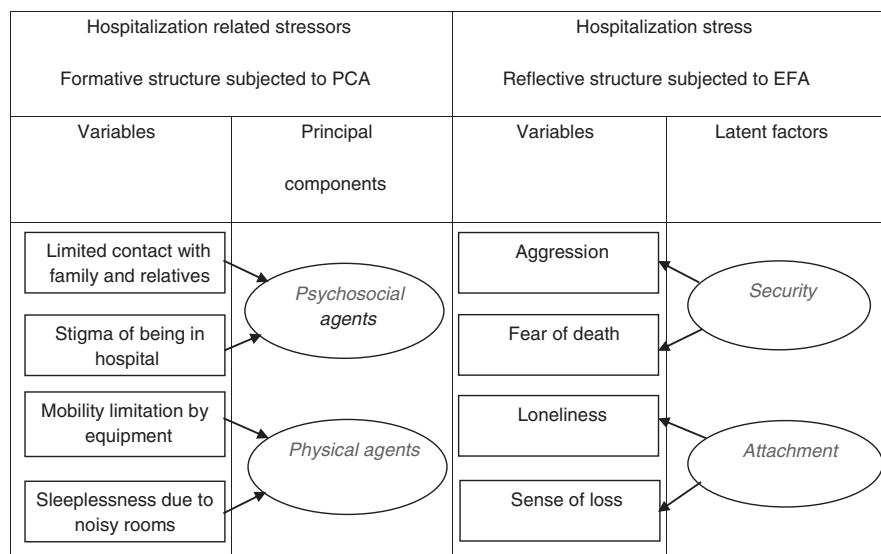
interrelationships to aid conceptualization of a set of latent constructs underlying a battery of measured variables. The information from the original measured variables is presented in a smaller number of derived factors (Gorsuch, 2014). The key objective is to extract the maximum common variance from the variables to arrange them under common factors to understand how much each variable contributes to each factor. The proportion of variance which can be explained by a set of factors which are common to the other observed variables is called *communality*. The degree of communality provides information to decide whether a particular factor should be retained. There is also a unique variance to that variable not explained by the factors, known as *uniqueness*.

Principal component analysis is used to simplify complex data by identifying a small number of *principal components* which capture the maximum variance. These components are linear combinations of the original variables. PCA and EFA achieve data simplification by identifying the number of components and factors, respectively, which explain the set of observed variables (*Component/Factor retention*). This choice involves a trade-off between parsimony (retaining fewer components/factors) and completeness (explaining more variance). Some other applications of EFA, in addition to data reduction, are analysis of multiple indicators, measurement and validation of complex constructs, and development and/or assessment of psychometric properties of new scales (Boateng, Neilands, Frongillo, Melgar-Quinonez, & Young, 2018; Gorsuch, 2014).

The relationship between an observed variable and a component/factor is expressed by a *factor loading* (ranging from 0 to 1), which measures the amount of the variance in the variable explained by the component/factor. A factor loading of >0.4 generally indicates that the variable can be attributed to the factor (Cuttillo, 2019). A factor loading matrix shows the relationship between the factors and the original variables, with components/factors typically named by the common attributes of the set of variables with which they are most correlated. Neither EFA nor PCA provide a unique solution, as component/factor rotation allows for an infinite number of possible representations. The rotation can be chosen to maximize simplicity, interpretation, and/or replicability (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Two common types of rotation are orthogonal rotation (e.g., Varimax and Quartimax rotation), where the components/factors remain uncorrelated with each other, and oblique rotation (e.g., Promax rotation), which allows for correlation.

In Figure 1 we provide a graphical representation of the relationship between observed and latent variables when working with two

FIGURE 1 Illustrative example showing relationships between components/factors and variables in PCA and EFA approaches. EFA, exploratory factor analysis; PCA, principal component analysis



apparently similar concepts of *hospitalization stress* and *hospitalization-related stressors* in two studies with different objectives. Research assessing *hospitalization stress* may use a large number of potential variables (e.g., loneliness, aggression, sense of loss, fear of death). Factor analysis may identify two underlying factors, *security* and *attachment* to which the variables load, with two variables loading to each factor. If one variable is hypothesized to be more related to one factor than another, this quantitative distinction can also be checked by EFA (Gorsuch, 2014). Alternatively, for a study on *hospitalization-related stressors*, there may be a large set of situations in hospital settings that may be associated with perceiving stress during the hospital stay. For a study measuring four variables, (a) mobility limitation due to connected equipment; (b) limited contact with family and relatives; (c) stigma of being in hospital; and (d) sleeplessness due to noisy rooms, PCA may have identified two principal components *physical agents* and *psychosocial agents* representing the four variables. The left side of Figure 1 shows PCA as a data reduction process identifying two principal components, whereas the right side shows EFA as a structure identification process comprising two latent factors.

3 | DIFFERENCES BETWEEN EFA AND PCA

Exploratory factor analysis and PCA are related but conceptually distinct techniques (Basto & Pereira, 2012). PCA reduces the number of variables extracting the essence of the dataset by creating principal components, whereas EFA uncovers the constructs underlying the data and identifies latent factors to explain the data. In the examples shown in Figure 1, EFA identified two underlying factors that account for variability of variables assessing patient stress, whereas PCA reduced the measured hospitalization stressors into two principal components.

The focus of EFA is the relationship among the variables, whereas PCA has more emphasis on data reduction than interpretation. PCA aims to explain the maximum amount of the total variance in the variables by analysing all of the observed variance, while in

EFA, only the shared covariance between the variables is analysed (Schneeweiss & Mathes, 1995). PCA is undertaken when there is sufficient correlation among the original variables. EFA is appropriate when we expect that there is a latent trait or unobservable characteristics among the observed variables. EFA and PCA also have different model assumptions regarding the data structure.

There are reasons that encourage researchers to use PCA rather than EFA. There are circumstances (e.g., where the error variances are small) in which PCA could be considered as a good approximation of EFA leading to yield similar output statistics (Rao & Sinharay, 2007). Another reason for increased use of PCA is that it is usually the default option in some statistical software packages increasing its use despite other approaches (Basto & Pereira, 2012; Hooper, 2012). An awareness of the differences between PCA and EFA allows for alignment between statistical approach and research objectives, and ensures appropriate interpretation of results (Santos et al., 2019).

Both EFA and PCA procedures identify patterns regardless of clinical knowledge behind those variables. These procedures can be used when the researcher has limited information with regard to the latent structure (Lever, Krzywinski, & Altman, 2017; Pett, Lackey, & Sullivan, 2003) which may lead to less attention to the theoretical knowledge needed to select the appropriate procedure. Returning to Figure 1, hospitalization stress and hospitalization-related stressors may seem similar, but the objectives of the study and nature of the observable variables determine which technique is appropriate. Where relevant clinical knowledge exists, this should be used as a guiding approach to any analysis, regardless of any existing or likely latent structure.

4 | INTERPRETING FACTORS AND PRINCIPAL COMPONENTS

The results of EFA simply set out a number of factors, the meaning of which has to be deduced from the variables which load

to the respective factors (Gorsuch, 2014). Instrument evaluation should distinguish between structures that are *reflective* (when factors affect or explain *effect indicators*) and *formative* (when components are formed but do not affect *cause indicators*). The first structure constructs the *scale* and the second constructs the *index*, known as reflective and formative measures, respectively. It is important to know that PCA identifies a formative structure and is conceptually inappropriate for effect indicators. Evaluation studies may inappropriately assume a reflective structure, and hence use EFA, where a formative structure is required. It is worth noting that the use of PCA does not imply the existence of a formative structure nor does using EFA imply an existing reflective structure, as both models could erroneously be used to analyse the same data and even yield similar results (Rao & Sinharay, 2007).

The researchers also need background knowledge to decide whether they are working with reflective or formative structures. In Figure 1, patient characteristic indicators of hospitalization stress have been treated as reflective indicators and have been subjected to EFA, whereas hospital setting characteristics are treated as formative indicators of perceived stress during hospital stay (*hospitalization-related stressors*) and have been subjected to PCA. All interpretations of factors/components based on loadings should be validated against external criteria (Gorsuch, 2014). If data reduction is the goal of analysis and the researcher is willing to have fewer dimensions through calculating weighted sums of indicators, PCA is the appropriate method and in this case, observed variables could not be considered as manifestations of components (Widaman, 1993).

In exploratory studies, the primary aim of the analysis is to examine the dataset to obtain the 'best estimate' of the components or latent factors to model the structure (Bro & Smilde, 2014). It should be noted that factors in EFA should be interpreted as explanatory rather than causal. For PCA, the principal components may be challenging to interpret, especially in high-dimensional databases (Allen & Maletic-Savatic, 2011; Chao, Wu, Wu, & Chen, 2018).


5 | CONCLUSION

Although both EFA and PCA are similar, they have different applications and interpretation. EFA is used to understand the underlying factors that are responsible for a set of observed variables, whereas PCA is used when the aim is data reduction. Given the problematic nature of causal language, a careful consideration of statistical procedure choice and research evidence reporting is important to minimize misinterpretation to better support the veracity of knowledge development (Thapa, Visentin, Hunt, Watson, & Cleary, 2020). As EFA and PCA have conceptual and statistical differences, attention to their characteristics is required to support accurate use and reporting so keep this in mind when you are deciding which one to use for your data needs.

CONFLICT OF INTEREST


No conflict of interest was declared by the authors in relation to the editorial itself. Note that Roger Watson is JAN Editor-in-Chief.

Mousa Alavi PhD, Associate Professor¹ 

Denis C. Visentin PhD, Senior Lecturer² 

Deependra K. Thapa MPH, MSc, Researcher² 

Glenn E. Hunt PhD, Associate Professor³ 

Roger Watson RN, PhD, FAAN, Professor⁴ 

Michelle Cleary RN, PhD, Professor² 

¹Department of Psychiatric Nursing, School of Nursing and Midwifery, Isfahan University of Medical Sciences, Isfahan, Iran

²College of Health and Medicine, University of Tasmania, Sydney, NSW, Australia

³Discipline of Psychiatry, Concord Clinical School, The University of Sydney, Sydney, NSW, Australia

⁴Faculty of Health Sciences, University of Hull, Hull, UK

Correspondence

Mousa Alavi, PhD, Department of Psychiatric Nursing, School of Nursing and Midwifery, Isfahan University of Medical Sciences, Hezarjarib Avenue, Isfahan, Iran.

Email: m_alavi@nm.mui.ac.ir

ORCID

Mousa Alavi  <https://orcid.org/0000-0003-4847-2915>

Denis C. Visentin  <https://orcid.org/0000-0001-9961-4384>

Deependra K. Thapa  <https://orcid.org/0000-0002-5689-0837>

Glenn E. Hunt  <https://orcid.org/0000-0002-8088-9406>

Roger Watson  <https://orcid.org/0000-0001-8040-7625>

Michelle Cleary  <https://orcid.org/0000-0002-1453-4850>

REFERENCES

- Allen, G. I., & Maletic-Savatic, M. (2011). Sparse non-negative generalized PCA with applications to metabolomics. *Bioinformatics*, 27(21), 3029–3035. <https://doi.org/10.1093/bioinformatics/btr522>
- Basto, M., & Pereira, J. M. (2012). An SPSS R-menu for ordinal factor analysis. *Journal of Statistical Software*, 46(4), 1–29. <https://doi.org/10.18637/jss.v046.i04>
- Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research, and Evaluation*, 18(1). <https://doi.org/10.7275/qv2q-rk76>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6. <https://doi.org/10.3389/fpubh.2018.00149>
- Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6(9), 2812–2831. <https://doi.org/10.1039/C3AY41907J>
- Chao, Y. S., Wu, H. C., Wu, C. J., & Chen, W. C. (2018). Principal component approximation and interpretation in health survey and bio-bank data. *Frontiers in Digital Humanities*, 5. <https://doi.org/10.3389/fdigh.2018.00011>
- Cuttillo, L. (2019). Parametric and multivariate methods. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of*

- bioinformatics and computational biology (pp. 738–746). Amsterdam, the Netherlands; Oxford, UK; Cambridge, MA: Elsevier.
- 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. <https://doi.org/10.1037/1082-989X.4.3.272>
- Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, 51(3), 511–521. <https://doi.org/10.1016/j.ijnurstu.2013.10.005>
- Gorsuch, R. L. (2014). *Factor analysis, classic edition*. New York, NY: Routledge Taylor & Francis.
- Hooper, D. (2012). Exploratory factor analysis. In H. Chen (Ed.), *Approaches to quantitative research – Theory and its practical application: A guide to dissertation students* (pp. 125–136). Cork, Ireland: Oak Tree Press.
- Kline, P. (1994). *An easy guide to factor analysis*. London, UK: Routledge. <https://doi.org/10.4324/9781315788135>
- Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods*, 14, 641–642. <https://doi.org/10.1038/nmeth.4346>
- Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation* (2nd ed.). Thousand Oaks, CA: SAGE Publications.
- Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. Thousand Oaks, CA: SAGE Publications, Inc. <https://doi.org/10.4135/9781412984898>
- Rao, C. R., & Sinharay, S. (Eds.) (2007). *Handbook of statistics, volume 26: Psychometrics*. Amsterdam, the Netherlands: Elsevier.
- Santos, R. O., Gorgulho, B. M., Castro, M. A., Fisberg, R. M., Marchioni, D. M., & Baltar, V. T. (2019). Principal component analysis and factor analysis: Differences and similarities in nutritional epidemiology application. *Revista Brasileira de Epidemiologia*, 22, E190041. <https://doi.org/10.1590/1980-549720190041>
- Schneeweiss, H., & Mathes, H. (1995). Factor analysis and principal components. *Journal of Multivariate Analysis*, 55(1), 105–124. <https://doi.org/10.1006/jmva.1995.1069>
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.
- Thapa, D. K., Visentin, D. C., Hunt, G. E., Watson, R., & Cleary, M. (2020). Being honest with causal language in writing for publication. *Journal of Advanced Nursing*. <https://doi.org/10.1111/jan.14311>
- Watson, R., & Thompson, D. R. (2006). Use of factor analysis in *Journal of Advanced Nursing: Literature review*. *Journal of Advanced Nursing*, 55(3), 330–341. <https://doi.org/10.1111/j.1365-2648.2006.03915.x>
- Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, 28(3), 263–311. https://doi.org/10.1207/s15327906mbr2803_1