Preventive Medicine 67 (2014) 141–146

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Review

A primer on the use of cluster analysis or factor analysis to assess co-occurrence of risk behaviors

Hedwig Hofstetter [*], Elise Dusseldorp, Pepijn van Empelen, Theo W.G.M. Paulussen

TNO (Netherlands Organization for Applied Scientific Research), Expertise Group Life Style, The Netherlands

ARTICLE INFO

Available online 15 July 2014

Keywords:
Health behaviors
Multiple risk behavior
Clusters
Factors

ABSTRACT

Objective: The aim of this paper is to provide a guideline to a universal understanding of the analysis of co-occurrence of risk behaviors. The use of cluster analysis and factor analysis was clarified.

Method: A theoretical introduction to cluster analysis and factor analysis and examples from literature were provided. A representative sample ($N = 4395$) of the Dutch population, aged 16–40 and participating from fall 2005 to spring 2006, was used to illustrate the use of both techniques in assessing the co-occurrence of risk behaviors.

Results: Whereas cluster analysis techniques serve to focus on particular clusters of individuals showing the same behavioral pattern, factor analysis techniques are used to assess possible groups of interrelated health-risk behaviors that can be explained by an unknown common source. Choice between the techniques partly depends on the research question and the aim of the research, and has different implications for inferences and policy.

Conclusion: By integrating theory and results from an illustrative example, a guideline has been provided that contributes towards a systematic approach in the assessment of co-occurrence of risk behaviors. Following this guideline, a better comparison between outcomes from various studies is expected, leading to improved effectiveness of multiple behavior change interventions.

Contents

Introduction

Over the past decades, there has been growing interest in research on associations of lifestyle-risk behaviors (see, for example, Bailey et al., 2006; de Vries et al., 2008; Prochaska, 2008; Pronk et al., 2004). Many studies have focused on four major lifestyle-risk factors, namely physical inactivity, smoking, drinking and nutrition or diet (e.g., Bailey et al., 2006; Conry et al., 2011; de Vries et al., 2008; Heroux et al., 2012; Laska et al., 2009; Lippke et al., 2012; Poortinga, 2007; Schuit et al., 2002; Van Nieuwenhuizen et al., 2009). Other factors have also been examined, such as psychological stress (Dodd et al., 2010), delinquency behavior (Van Nieuwenhuizen et al., 2009), drug use (Faeh

[*] Corresponding author at: TNO, Expertise Group Life Style, P.O. Box 3005, 2301 DA Leiden, The Netherlands.
E-mail address: hedwig.hofstetter@tno.nl (H. Hofstetter).

http://dx.doi.org/10.1016/j.ypmed.2014.07.007
0091-7435/© 2014 Elsevier Inc. All rights reserved.

et al., 2006; Van Nieuwenhuizen et al., 2009), and unsafe sex (Van Nieuwenhuizen et al., 2009). These lifestyle-risk factors are major but preventable causes of morbidity and mortality.

Two popular statistical techniques used in studies on co-occurrence of risk behaviors are cluster analysis and factor analysis. The underlying logic of both techniques is dimension reduction (i.e., summarizing information on multiple variables into just a few variables), but they do so in very different ways. Cluster analysis techniques reduce the number of individuals into a smaller number of profiles (i.e., clusters of people) by assessing the interrelationships between individuals. The goal of factor analysis techniques is to reduce the number of variables into components (i.e., factors of behaviors). In factor analysis, groups of behaviors that are interrelated due to a common underlying factor (also called latent variable or construct) are identified.

Although often not clear to researchers or applied researchers, the choice of technique has implications for the results and conclusions that can be drawn. Researchers must therefore carefully consider which technique can answer which questions. Unfortunately, literature about multiple behaviors has shown that terminology is not consistent, and that confusing inferences are drawn from the various statistical techniques. Nigg et al. (2002), for example, stated that "health behaviors often cluster". The same phrase was used in a study by de Vries et al. (2008), who explored "clusters of health behaviors". They confusingly reported in their results that: "The distribution of these groups of behaviors resulted in three clusters of people …". Dodd et al. (2010) stated that: "… research has shown that health behaviors often coexist and that there is clear evidence of clustering". The authors hypothesized that with their results they would support health professionals in their understanding of "how behaviors cluster together". They analyzed their data using a cluster analysis method. In their discussion, it was stated that "the cluster analysis clearly demonstrates patterns between the behaviors".

As these examples show, there is a need for clarification of terminology, the choice of statistical techniques, and inferences that can be drawn from these techniques. Without such clarification, comparison between multiple risk behavior studies is hampered (Heroux et al., 2012; Poortinga, 2007). A systematic approach is desirable to facilitate a universal understanding of research concerning multiple health behaviors (de Vries et al., 2008). Such an approach leads to the envisaged straightforward link between research question, statistical technique, and conclusion. In this paper we take a first step towards framing a guideline for multiple behavior research, by clarifying terminology and by providing a clear differentiation between statistical techniques, research questions that can be answered by the techniques, and inferences that can be drawn. By using theory and an illustrative example, we will show that each of the statistical techniques has different implications for inferences and policy.

Firstly, we will provide a short theoretical introduction to cluster analysis and factor analysis, and cite examples from multiple behavior studies in which the techniques were used. Subsequently, an illustrative example is given in which both factor analysis and cluster analysis techniques are implemented to the same dataset. To conclude, we will integrate findings from the literature and our example and guide the reader in choosing the most appropriate analysis technique to meet his or her needs.

Cluster analysis

Theory

Cluster analysis is an exploratory technique used to classify people into a preferably small number of groups based upon their scores on observed variables. The underlying model is discrete: in the end individuals belong to one and only one cluster. Basically, five steps can be identified in cluster analysis, namely 1) selection of a sample of individuals to be clustered, 2) definition of a set of variables used to measure the individuals in the sample, 3) computation of the similarities between the individuals, 4) use of a cluster analysis method to create groups of similar individuals, and 5) interpretation of results.

The result of the first two steps is a data matrix consisting of n individuals (represented in rows) measured on p variables (the columns of the matrix). A visual representation of data from two behavioral variables, for example the number of hours per week of physical exercise (variable x_1) and the number of alcohol units consumed per week (variable x_2), is shown in a two-dimensional space in Fig. 1. The figure clearly shows two clusters of individuals: one cluster with people who consume large amounts of alcohol and spend little time on physical activity per week, and a homogenous subgroup of individuals who exercise more and drink less alcohol.

The dimensionality of the space is determined by the number of variables used to describe the individuals. For seven variables, for example, data is represented as a seven-dimensional space. In the third step of the cluster analysis, the coordinates in space are examined by means of a dissimilarity measure. This dissimilarity measure, such as a distance measure, expresses the relationships between individuals given their values on a set of variables. The distance between cases i and j can, for example, be computed by squaring the difference between the value on variable p for cases i and j, and by summing these squared differences over all variables (e.g., physical exercise and alcohol consumption [Aldenderfer and Blashfield, 1984]). The smaller the distance value, the more cases i and j are alike. These distance values are then summarized into an $n \times n$ dissimilarity (e.g., distance) matrix, with n representing the individuals.

In the fourth step, a clustering method is used to create clusters of similar individuals based on this $n \times n$ matrix. Several families of methods are available, each representing a different view on the creation of groups. Popular clustering methods in social and medical sciences are hierarchical clustering and latent cluster analysis. Hierarchical clustering uses the $n \times n$ matrix to sequentially merge the most similar individuals. Many possible merging rules are available for this (e.g., single linkage, complete linkage), all aiming to measure the distance between individual observations. Contrary to this standard ad-hoc clustering technique, latent cluster analysis (Vermunt and Magidson, 2002) is a model-based clustering approach. This technique does not use a

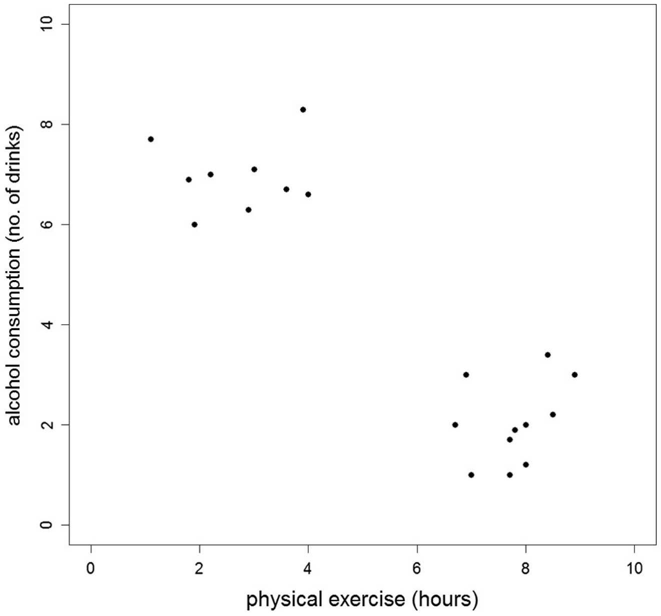

Fig. 1. Visual depiction of cluster analysis in a two-dimensional space. Two behaviors, physical exercise and alcohol consumption, are represented on the X- and Y-axes, respectively. Each dot represents an individual.

dissimilarity matrix, but clusters individuals using a probability-based classification. The observed variables are considered to be indicators for an unobserved latent variable, and the association between the observed responses can be fully explained by the small number of latent clusters. Cluster-specific response probabilities are calculated, representing the likelihood of showing a particular behavior.

The choice between clustering methods depends largely on the choice of classification, the measurement level of the variables, sample size, and preference of scientific disciplines. Hierarchical clustering techniques can handle quantitative, binary, or count data, and performs well with smaller sample sizes. Latent cluster analysis offers great flexibility, and is particularly suitable for variables with different measurement levels, large samples, longitudinal data, and multilevel data. For more in-depth reading material on cluster analysis, the reader is referred to books/chapters by, among other authors, Aldenderfer and Blashfield (1984), Everitt et al. (2011), and Vermunt and Magidson (2002).

The final step in cluster analysis, the interpretation of results, is the most fundamental step. The outcome of cluster analysis is not simply a set of clusters; active understanding and interpretation of results is required. Several criteria are available for assessing the fit of cluster models, for example the likelihood ratio chi-squared statistic L^2, Akaike Information Criterion (Akaike, 1973), and Bayesian Information Criterion (Schwarz, 1978). The result of cluster analysis is a relatively small number of clusters of individuals that resemble each other and that are different in some respects from individuals in other clusters. Names are assigned to these clusters, often denoting the most notable findings in the data. The final cluster assignment is presented in a nominal variable with its categories referring to the clusters. This variable can subsequently be used to examine which characteristics are shared by individuals within the same cluster. In addition, cluster allocation can serve as a predictor of other behaviors not included in the generation of the cluster solution.

Examples from literature

In multiple risk behavior research, cluster analysis is a popular technique used to identify subgroups sharing the same behaviors and/or other characteristics (see, for example, Carlerby et al., 2012; Conry et al., 2011; Flannery et al., 2003; Hagoel et al., 2002; Lippke et al., 2012). Dodd et al. (2010), for example, used cluster analysis to investigate the prevalence and clustering of five lifestyle-risk factors within a UK higher education institution. The researchers found three distinct clusters of people based on the lifestyle factors psychological stress, physical activity, fruit and vegetable intake, binge drinking, and smoking. An unhealthy/high risk group cluster contained individuals with high psychological distress, low physical exercise, low fruit and vegetable intake, and relatively many occasional and regular smokers. Individuals showing an opposite lifestyle, that is, low psychological distress, a high level of physical activity, high fruit and vegetable intake, moderate alcohol consumption, and non-smokers, were clustered into a healthy/low risk group. The third cluster distinguished individuals that were moderate in most lifestyle factors, that is, a moderate amount of psychological distress, physical exercise, and fruit and vegetable intake, but with a relatively high proportion of regular smokers.

Another example is a study by Hagoel et al. (2002), in which cluster analysis was carried out to group women by similarities in their health behaviors and to characterize their lifestyles in these terms. Data included were, among other things, domains of smoking, diet, physical exercise, and periodic medical checkups. Comparable to Dodd et al. (2010), the researchers found three clusters of individuals. The health-promoting lifestyle cluster is similar to the healthy/low risk cluster found by Dodd et al. (2010): this cluster included women showing health-promoting behaviors such as healthy diet, exercising, attending periodic medical checkups, and the avoidance of risk behaviors such as smoking. A second cluster was called the inactive cluster, and contained women who are distinct in their low level of physical exercise. Most remarkable about women in the ambivalent cluster is their smoking behavior: they all either currently smoke or had smoked in the past. They also adhere less to a healthy diet than women in the other two clusters, and are moderate in exercising and attending medical checkups. The authors also investigated the characteristics of each cluster on the basis of (among other things) demographic variables, and used cluster assignment as a predictor of another health behavior (mammography screening).

A study by Conry et al. (2011) focused on how key health-related behaviors (physical activity, alcohol consumption, smoking, diet), quality of life, and mental health are distributed in a national population of Irish adults. Six clusters were found, where the healthy lifestyle cluster is similar to the healthy clusters found in the studies by Dodd et al. (2010) and Hagoel et al. (2002). Other clusters found were labeled temperate, former smoker, mixed lifestyle, physically inactive, and multiple risk factors.

Factor analysis

Theory

In multiple behavior research, researchers often wish to represent the relationships among observed behavioral variables. With only a few observed variables, simple correlation coefficients can be used for this goal. However, when a researcher is confronted with many more variables, the interpretation of the pattern of correlations can become very cumbersome. The patterns of associations must then be explored empirically, with factor analysis methods being appropriate procedures. The primary goal of these methods is to explore the associations among p observed variables and to summarize these relationships into a smaller number of m new variables (latent constructs), with $m < p$ (Velicer and Jackson, 1990). Scores for the m new variables can be used to replace the original p observed scores.

Two popular procedures are common factor analysis (Zwick and Velicer, 1982) and principal component analysis (Hotelling, 1933; Pearson, 1901). In spite of some mathematical differences, which are beyond the scope of this paper, the choice between these two procedures is not very critical for practical purposes. Velicer and Jackson (1990) reviewed several studies in which different types of component and factor analysis were compared, and concluded that discrepancies between the two procedures are rare when the same number of latent constructs are extracted.

A factor can be described as an unobservable (i.e., latent) variable, that can account for the correlations or part of the correlations in the observed data. Common factors are latent variables that account for variability in at least two of the observed variables. Unique factors represent measurement error and variability in each of the observed variables that has nothing in common with any of the other variables. The purpose of factor analysis is to estimate the pattern of relations between common factor(s) and each of the observed variables. A visual explanation of this is given in Fig. 2.

Factors are a weighted sum of the observed variables, where the obtained weights depend on the analysis performed (e.g., confirmatory or exploratory analysis[1]). Once the factors are determined, correlations between these factors and the observed variables are computed. Variables having a high correlation with a particular factor are said to be highly linearly related to that factor, and are interpreted as being part

[1] A broad distinction is made between exploratory and confirmatory factor analysis. Although both methods are used to grasp the underlying factor structure of the data, they each have their own objective. Confirmatory factor analysis is primarily used for theory-testing. Common factors are 'known', and it is examined if the a priori model of the underlying structure of the data is supported by the data. Exploratory factor analysis is used for theory-building, as common factors are unknown and are searched for. We focus on the latter type of factor analysis.

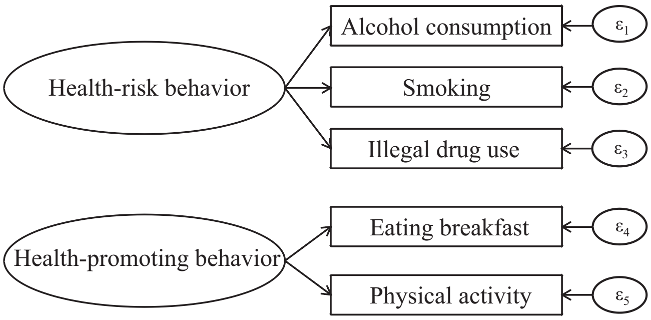

Fig. 2. Visual depiction of factor analysis. An oval represents a latent factor and a rectangle represents an observed variable. Unique factors are depicted by ε. Factors create variables, as represented by the arrows.

of the same underlying common source.[2] The proportion of variance explained by the obtained factors is a measure of how well the defined factors describe the original observed data. The result of factor analysis is multiple numeric variables, each representing a factor score locating a person on that factor's underlying continuum. These scores can be used for further analysis, such as the identification of variables that are associated with the factors.

Both factor analysis and principle component analysis are appropriate for the analysis of behaviors measured at a continuous or categorical measurement scale. For the latter situation, a categorical version of principle component analysis (CATPCA) has been developed (Linting et al., 2007).

With regard to risk factors measured on a binary scale (e.g., present or absent), we also came across a few studies using prevalence odds ratios (PORs) to investigate the associations between life-style risk behaviors (Alamian and Paradis, 2009; Faeh et al., 2006; Poortinga, 2007; Schuit et al., 2002). Using PORs, an association between risk behaviors is identified when the combination of risk behaviors exceeds the expected prevalence of the combination of these risk behaviors. The expected prevalence is calculated using the individual probabilities of each risk behavior based on their occurrence in the study population. When the number of behaviors increases, the computation of PORs becomes very time consuming (e.g., with seven risk behavior variables, 128 possible associations have to be explored), and factor analysis would be a more appropriate technique for analyzing the structure of associations between behaviors.

Examples from literature

Van Nieuwenhuizen et al. (2009) used exploratory factor analysis to investigate the grouping of health-compromising behaviors and delinquent behavior in a Dutch sample. Three factors were found in an adult sample. A factor named *Health* comprised behaviors such as eating breakfast, adequate fruit and vegetable intake, adequate physical exercise, and non-smoking behavior. An *Alcohol* factor contained behavior related to alcohol consumption and unsafe sex. A latent variable named *Delinquency* included physical and verbal aggression, delinquency behavior, drug abuse, and ignoring a red light in traffic. Factor scores representing these latent variables can subsequently be used as outcomes to assess determinants of these groups of behaviors (see, for example, Dusseldorp et al., in press). Principal component analysis was used in a study by Lippke et al. (2012) to investigate whether behaviors are interrelated. The authors found that health-promoting behaviors (such as nutrition and exercising) loaded on one factor, whereas health-risk behaviors (smoking, alcohol consumption) loaded on another.

Table 1
Item-response probabilities within each of the three clusters: Risk Behavior Survey, The Netherlands, 2005–2006 (N = 3975).

	Cluster 1 (N = 2283)	Cluster 2 (N = 1189)	Cluster 3 (N = 503)
Cluster size	.537	.322	.142
Alcohol			
Does not adhere to norm	.184	.455	.215
Adheres to norm	.816	.545	.786
Smoking			
Does not adhere to norm	.069	.735	.123
Adheres to norm	.931	.265	.877
Illegal drug use			
Does not adhere to norm	.003	.247	.045
Adheres to norm	.997	.753	.955
Fruit intake			
Does not adhere to norm	.871	.900	.475
Adheres to norm	.129	.100	.525
Eating breakfast			
Does not adhere to norm	.209	.484	.009
Adheres to norm	.791	.517	.991
Vegetable intake			
Does not adhere to norm	.917	.910	.686
Adheres to norm	.083	.090	.314
Physical activity			
Does not adhere to norm	.139	.067	.013
Adheres to norm	.861	.933	.987

Illustrative example

To illustrate the use of cluster analysis and factor analysis and the inferences that can be drawn from these techniques, we used data from a representative sample (N = 4395) of the Dutch population aged 12 to 40 years (Van Nieuwenhuizen et al., 2009). Data contained (among other things) information about various health-compromising behaviors, namely alcohol consumption, smoking, illegal drug use, physical inactivity, skipping breakfast, and not eating fruit and vegetables. Norm scores were calculated for each of these seven behaviors according to separate guidelines for age groups, to assess whether respondents adhered to the norm of healthy behavior (for an overview of the norms, see Appendix A[3]). This resulted in seven variables in two categories, namely 'adheres to the norm' and 'does not adhere to the norm'. Analyses were based on 3975 respondents who had completed data on these seven variables.

Latent cluster analysis was conducted using LatentGold (Vermunt and Magidson, 2005). LatentGold can handle all types of variables (i.e., categorical and continuous) and offers high flexibility to the user. One- to four-cluster models were estimated, and the best fitting model was assessed using the L^2 statistic and the associated p-value. Among models for which this p-value was greater than 0.05 (indicating adequate fit), the one with the smallest number of parameters was selected. Using this criterion, the best model was given by a three-cluster model (p = .65, $Npar$ = 23). As can be seen in Table 1, one cluster was composed of respondents who showed a high probability of adherence to the norm of alcohol consumption, smoking, drug use, eating breakfast, and adequate physical activity, and was characterized by a very high probability (>.85) of no adherence to the norms of fruit and vegetable intake. A second cluster contained respondents who were characterized by high probabilities of smoking and no adherence to the norms of fruit and vegetable intake, and a moderate probability of no adherence to the alcohol norm. This second cluster can be considered to be the unhealthiest cluster, as the probabilities of all the 'no

[2] It should be noted that cluster analysis can also be used to identify groups of behaviors. The convention in cluster analysis is to convert the data in an $n \times n$ dissimilarity matrix (see paragraph 2.1). However, the data can also be reversed into a $p \times p$ similarity matrix, where correlations are used as a similarity measure. Although there are (a few) studies using cluster analysis of variables (for example, Bender et al., 2005), we are unaware of studies in the area of multiple risk behaviors that use cluster analysis to identify groups of behaviors.

[3] Included as supplementary file.

Table 2
Result of principal component analysis[a]: Risk Behavior Survey, The Netherlands, 2005–2006 (N = 3975).

	Factor 1	Factor 2	Factor 3
Smoking	.697		
Illegal drug use	.668		
Alcohol	.608		
Breakfast	.432	.410	
Vegetable intake		.698	
Fruit intake		.664	
Physical activity			.955

[a] Varimax rotated principal component analysis. Loadings below .4 are suppressed.

adherence to the norm' categories were higher or almost equal to those in the other two clusters (Table 1). Respondents showing the healthiest behaviors were grouped into a third cluster.

Principal component analysis with Varimax rotation was performed using SPSS Statistics 20.0 to investigate whether groups of behaviors were present in the data. Using both the scree criterion (Cattell, 1966) and Kaiser's criterion (Kaiser, 1960), a three factor solution was found, explaining about 53% of the variance in the data. As can be seen in Table 2, behaviors related to alcohol use, smoking, and illegal drug use were grouped into one factor. Another factor contained behavior related to fruit and vegetable intake. Breakfast behavior was related to both factors. Physical inactivity was not related to the other behaviors, and was defined as a third factor.

A first setup to a guideline

Our example, in which we used both cluster analysis and factor analysis to analyze the same dataset, illustrates that the techniques lead to different results. Using cluster analysis, we identified three clusters of individuals sharing the same (un)health(y) behaviors: a poor diet cluster, a high risk cluster, and a low risk cluster. Noticeably, the physical exercise norm and illegal drug use norm did not differentiate individuals, since item-response probabilities for not adhering to these norms were below 0.25 within each of the three clusters. With factor analysis, three groups of interrelated behaviors were identified: an addictive behavior factor, a healthy eating behavior factor, and a physical activity factor. Higher scores on the first factor meant abstinence or restraint from health-risk behaviors, while the healthy eating factor suggested the active engagement in health-promoting activities. Clearly, physical activity did not share a common source with the other observed variables. In other words, there is no relationship between (not) adhering to the physical activity norm and the other behavioral norms. Adhering to the drug use norm was part of the addictive behavior factor, and therefore shared the same underlying common source as smoking, alcohol use, and eating breakfast.

The results from this illustrative example and the theory explained above give ample evidence that the choice between cluster analysis and factor analysis is an important one. A short summary with typical research questions and inferences that can be drawn with each technique is provided in Table 3. In addition, implications for policy and interventions in multiple risk behavior research are presented; these implications will be discussed in the next section.

Discussion

Cluster analysis and factor analysis are both dimension reduction techniques, and can be used to analyze the co-occurrence of lifestyle-risk behaviors. However, each technique comes with its own unique goal, which is often not clear to the applied researcher. In this paper, we have tried to provide the reader with a clarification of terminology, a clear differentiation between the two techniques, research questions they can answer, and inferences that can be drawn.

Our results from cluster analysis are consistent with a study by de Vries et al. (2008), who also found three clusters defined as a healthy, an unhealthy, and a poor nutrition cluster. However, in their study physical activity appeared to be more distinctive among individuals. This disagreement in finding could be due to the fact that in our example data almost all respondents adhered to the physical activity norm, while the adherence rate for physical activity in the sample of de Vries et al. (2008) was much lower. The high risk cluster of our solution was also apparent in a study among women by Hagoel et al. (2002), who found a cluster characterized by (among other variables not used in our example) smoking behavior and unhealthy diet.

Comparable results from our example using factor analysis were found by Lippke et al. (2012), who found a health-risk behavior factor (smoking, drinking) and a health-promoting behavior factor (nutrition). Contrary to our example, however, physical exercise was also part of this latter health-promoting behavior factor. This could be explained by a difference in measuring physical exercise. In the study by Lippke et al. (2012), engagement in regular physical exercise (i.e., at least four times per week for at least 30 min each day at moderate or high intensity) was classified in five stages, ranging from no engagement in physical exercise and no intention of doing so in the next six months to currently exercising, but for less than six months. We, however, defined being physically active as adhering to the norm (yes/no) of exercising at least five days a week for at least 30 min each day at a moderate or high intensity.

Using both methods, we have reaffirmed that risk behaviors do not occur in isolation. More importantly, it has been made explicitly clear that the use of factor analysis techniques could increase our knowledge of possible associations between health-risk behaviors, whereas the application of cluster analysis techniques serves to focus on particular clusters of individuals showing the same behavioral pattern. An important fundamental difference between the two techniques is represented in the underlying model. Cluster analysis is an intrinsic discrete model: individuals belong to one and only one cluster. In factor analysis, on the other hand, a factor score is represented on a continuum. The distinction with respect to the underlying model is an important one, and can affect the choice between cluster and factor analysis.

Cluster analysis techniques cluster groups of individuals with homogeneous behavioral patterns. Health-promoting campaigns can be

Table 3
Comparison of cluster analysis and factor analysis concerning type of research questions, inferences, and policy/intervention.

	Cluster analysis	Factor analysis
Research questions	- Which subgroups of individuals share the same behaviors? - How can individuals be classified according to similarities in behaviors?	- What kind of behaviors group together? - Which behaviors are interrelated?
Inferences	- Cluster assignment can be used as a predictor variable of other behaviors. - Cluster assignment can be used as a nominal outcome variable to identify person characteristics that are shared by individuals from the same cluster (e.g., using discriminant analysis).	- A factor score can be used as a continuous outcome measure to identify variables that are associated with that factor (e.g., using multiple regression). - A factor of interrelated behaviors can be used to shed light on an underlying common source.
Policy/intervention	- Health-promoting campaigns can be targeted at clusters of individuals showing the same behavioral pattern, and could therefore be aimed at behaviors from different domains. - Demographic variables that are related to the clusters can be used for segmentation in health-campaigns.	- Intervention strategies can be targeted at behaviors sharing the same underlying source: transfer of new acquired knowledge, attitudes or skills can be induced more easily between behaviors of the same factor, than between behaviors of different factors.

targeted at different groups of individuals showing the same behavioral pattern, and could therefore be aimed at behaviors from different domains (e.g., smoking and fruit intake; cluster two in our example). Findings by O'Halloran et al. (2001) suggest that individuals with certain behavioral patterns may differ in their response to interventions. Demographic variables, such as age and gender, can also be used for segmentation in health-promoting campaigns (Reedy et al., 2005). Results from studies using factor analysis techniques, on the other hand, concentrate on the finding of an unknown common source that can explain the association between variables. Knowledge about this source may help intervention strategies to focus on behaviors sharing the same underlying source. In this way, transfer of new acquired knowledge, attitudes or skills can be induced between behaviors sharing the same underlying component (referring to our example: between the behaviors from the health-promoting factor or the health-risk factor). This type of transfer can be referred to as near transfer, and will occur more easily than far transfer (transfer to behaviors from a dissimilar domain). Accordingly, if an intervention succeeds in changing a particular behavior (for example, alcohol consumption), related behaviors (e.g., smoking behavior) may also change (Peters et al., 2013). As explained in this article, comparison between studies on multiple health behaviors is hampered, in part because of an inconsistency of terminology, statistical approaches, and inferences drawn from results. This article serves as a guideline to a universal understanding: a systematic approach could help to integrate information from various studies, such that the effectiveness of multiple behavior change interventions can be enhanced. A short summary is provided to guide the reader in choosing the most suitable analysis technique to meet his or her needs. We hope to have raised awareness among researchers in multiple behavior research of the importance of thinking carefully about their research question and the aim of their research, thereby enabling them to choose the appropriate analysis technique.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ypmed.2014.07.007.

Conflict of interest statement

The authors declare that they have no conflicts of interest.

Acknowledgments

We thank Prof. dr. S. Van Buuren for his comments on earlier drafts of this manuscript and L. Ouwehand for her assistance.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, B.F. (Eds.), Second International Symposium on Information Theory. Academiai Kiado, Budapest, pp. 267–281.

Alamian, A., Paradis, G., 2009. Clustering of chronic disease behavioral risk factors in Canadian children and adolescents. Prev. Med. 48 (5), 493–499.

Aldenderfer, M.S. & Blashfield, R.K. 1984, Cluster Analysis, 07-044 edn, Sage University Paper series on Qualitative Applications in the Social Sciences, Beverly Hills.

Bailey, R.L., Gutschall, M.D., Mitchel, D.C., Miller, C.K., Lawrence, F.R., Smiciklas-Wright, H., 2006. Comparative strategies for using cluster analysis to assess dietary patterns. J Am Diet Assoc 106, 1194–1200.

Bender, C.M., Ergyn, F.S., Rosenzweig, M.Q., Cohen, S.M., Sereika, S.M., 2005. Symptom clusters in breast cancer across 3 phases of the disease. Cancer Nurs. 28 (3), 219–225.

Carlerby, H., Englund, E., Viitasara, E., Knutsson, A., Gadin, K.G., 2012. Risk behaviour, parental background, and wealth: a cluster analysis among Swedish boys and girls in the HBSC study. Scand. J. Public Health 40 (4), 368–376.

Cattell, R.B., 1966. The scree test for the number of factors. Multivar. Behav. Res. 1 (2), 245–276.

Conry, M., Morgan, K., Curry, P., et al., 2011. The clustering of health behaviours in Ireland and their relationship with mental health, self-rated health and quality of life. BMC Public Health 11 (1), 692.

de Vries, H., van 't Riet, J., Spigt, M., et al., 2008. Clusters of lifestyle behaviors: results from the Dutch SMILE study. Prev. Med. 46, 203–208.

Dodd, L.J., Al-Nakeeb, Y., Nevill, A., Forshaw, M.J., 2010. Lifestyle risk factors of students: a cluster analytical approach. Prev. Med. 51, 73–77.

Dusseldorp, E., Klein velderman, M., Paulussen, T.G.W.M., Junger, M., Van Nieuwenhuijzen, M., Reijneveld, S.A., 2014. Targets for primary prevention: Cultural, social and interpersonal factors associated with co-occurring health-related behaviours. Psychology and Health. http://dx.doi.org/10.1080/08870446.2013.879137 (in press).

Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Cluster Analysis, 5th ed. John Wiley & Sons Ltd., West Sussex, UK.

Faeh, D., Viswanathan, B., Chiolero, A., Warren, W., Bovet, P., 2006. Clustering of smoking, alcohol drinking and cannabis use in adolescents in a rapidly developing country. BMC Public Health 6 (169).

Flannery, W.P., Sneed, C.D., Marsh, P., 2003. Toward an empirical taxonomy of suicide ideation: a cluster analysis of the youth risk behavior survey. Suicide Life Threat. Behav. 33 (4), 365–372.

Hagoel, L., Ore, L., Neter, E., Silman, Z., Rennert, G., 2002. Clustering women's health behaviors. Health Educ. Behav. 29 (2), 170–182.

Heroux, M., Janssen, I., Lee, D., Sui, X., Hebert, J.R., Blair, S.N., 2012. Clustering of Unhealthy Behaviors in the Aerobics Center Longitudinal Study. Prev. Sci. 13, 183–195.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441.

Kaiser, H.F., 1960. The application of electronic computers to factor analysis. Educ. Psychol. Meas. 20, 141–151.

Laska, M.N., Pasch, K.E., Lust, K., Story, M., Ehlinger, E., 2009. Latent class analysis of lifestyle characteristics and health risk behaviors among college youth. Prev Sci 10, 376–386.

Linting, M., Meulman, J.J., Groenen, P.J.F., van der Kooij, A.J., 2007. Nonlinear principal components analysis: introduction and application. Psychol. Methods 12, 336–358.

Lippke, S., Nigg, C.R., Maddock, J.E., 2012. Health-promoting and health-risk behaviours: theory-driven analyses of multiple health behaviour change in three international samples. Int. J. Behav. Med. 19 (1), 1–13.

Nigg, C.R., Allegrante, J.P., Ory, M., 2002. Theory-comparison and multiple behavior research: common themes advancing health behavior research. Health Educ Research 17 (5), 670–679.

O'Halloran, P., Lazovich, D., Patterson, R.E., et al., 2001. Effect of health lifestyle pattern on dietary change. Am. J. Health Promot. 16 (1), 27–33.

Pearson, K., 1901. On lines and planes of closest fit to system of points in space. Phil. Mag. 2 (6), 559–572.

Peters, L.W.H., Ten Dam, G.T.M., Kocken, P.L., Buijs, G.J., Dusseldorp, E., Paulussen, T.G.W. M., 2013. Effect of transfer-oriented curriculum on multiple behaviors in The Netherlands. Health Promot. Int. 19 (Online).

Poortinga, W., 2007. The prevalence and clustering of four major lifestyle risk factors in an English adult population. Prev. Med. 44, 124–128.

Prochaska, J.O., 2008. Multiple Health Behavior Research represents the future of preventive medicine. Preventive Medicine 46, 281–285.

Pronk, N.P., Anderson, L.H., Crain, A.L., Martinson, B.C., O'Connor, P.J., Sherwood, N.E., Whitebird, R.R., 2004. Meeting recommendations for multiple healthy lifestyle factors: prevalence, clustering, and predictors among adolescent, adult, and senior health plan members. Am J Prev Med 27 (28), 25–33.

Reedy, J., Haines, P.S., Campbell, M.K., 2005. The influence of health behavior clusters on dietary change. Prev. Med. 41, 268–275.

Schuit, A., Van Loon, A.J.M., Tijhuis, M., Ocké, M.C., 2002. Clustering of life style risk factors in a general adult population. Prev. Med. 35, 219–224.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Van Nieuwenhuizen, M., Junger, M., Klein Velderman, M., et al., 2009. Clustering of health-compromising behavior and delinquency in adolescents and adults in the Dutch population. Prev. Med. 48, 572–578.

Velicer, W.F., Jackson, D.N., 1990. Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. Multivar. Behav. Res. 25 (1), 1–28.

Vermunt, J.K., Magidson, J., 2002. Latent class cluster analysis. In: Hagenaars, J., McCutcheon, A. (Eds.), Applied Latent Class Analysis. Cambridge University Press, Cambridge, pp. 89–106.

Vermunt, J.K., Magidson, J., 2005. Latent Gold 4.0 User's Guide. Statistical Innovations, Belmont, Massachusetts.

Zwick, W.R., Velicer, W.F., 1982. Factors influencing four rules for determining the number of components to retain. Multivar. Behav. Res. 17 (2), 253–269.