PLOS ONE

RESEARCH ARTICLE

# A comparison of children's diet and movement behaviour patterns derived from three unsupervised multivariate methods

**Ninoshka J. D'Souza[1¤]*, Katherine Downing[1¤], Gavin Abbott[1¤], Liliana Orellana[2¤], Sandrine Lioret[3], Karen J. Campbell[1¤], Kylie D. Hesketh[1¤]**

**1** Institute of Physical Activity and Nutrition, School of Exercise and Nutrition Sciences, Deakin University, Geelong, Victoria, Australia, **2** Biostatistics Unit, Deakin University, Geelong, Victoria, Australia, **3** Research Center in Epidemiology and Biostatistics (CRESS), INSERM, INRAE, Université de Paris, Paris, France

¤ Current address: Faculty of Health, Deakin University, Geelong, Victoria, Australia
* njdsouza@deakin.edu.au





**Citation:** D'Souza NJ, Downing K, Abbott G, Orellana L, Lioret S, Campbell KJ, et al. (2021) A comparison of children's diet and movement behaviour patterns derived from three unsupervised multivariate methods. PLoS ONE 16(7): e0255203. https://doi.org/10.1371/journal.pone.0255203

**Editor:** Jane Anne Scott, Curtin University, AUSTRALIA

**Received:** February 24, 2021

**Accepted:** July 13, 2021

**Published:** July 27, 2021



**Data Availability Statement:** The datasets analysed for the current study are not publicly available due to ethical restrictions related to the consent given by participants at the time of study commencement. An ethically compliant dataset may be made available by the corresponding author on reasonable request and upon approval by the Deakin University Human Research Ethics Committee. Contact details: Human Research Ethics Office Deakin University 221 Burwood Hwy

## Abstract

### Background

Behavioural patterns are typically derived using unsupervised multivariate methods such as principal component analysis (PCA), latent profile analysis (LPA) and cluster analysis (CA). Comparability and congruence between the patterns derived from these methods has not been previously investigated, thus it's unclear whether patterns from studies using different methods are directly comparable. This study aimed to compare behavioural patterns derived across diet, physical activity, sedentary behaviour and sleep domains, using PCA, LPA and CA in a single dataset.

### Methods

Parent-report and accelerometry data from the second wave (2011/12; child age 6-8y, n = 432) of the HAPPY cohort study (Melbourne, Australia) were used to derive behavioural patterns using PCA, LPA and CA. Standardized variables assessing diet (intake of fruit, vegetable, sweet, and savoury discretionary items), physical activity (moderate- to vigorous-intensity physical activity [MVPA] from accelerometry, organised sport duration and outdoor playtime from parent report), sedentary behaviour (sedentary time from accelerometry, screen time, videogames and quiet playtime from parent report) and sleep (daily sleep duration) were included in the analyses. For each method, commonly used criteria for pattern retention were applied.

### Results

PCA produced four patterns whereas LPA and CA each generated three patterns. Despite the number and characterisation of the behavioural patterns derived being non-identical, each method identified a healthy, unhealthy and a mixed pattern. Three common underlying themes emerged across the methods for each type of pattern: (i) High fruit and vegetable intake and high outdoor play ("healthy"); (ii) poor diet (either low fruit and vegetable intake or

Burwood, VIC 3125 E-mail: research-ethics@deakin.edu.au.

**Funding:** The HAPPY study was funded by an Australian Research Council Discovery Grant (DP140100554). NJD is supported by a Deakin University Postgraduate Scholarship, KLD is supported by an Alfred Deakin Postdoctoral Research Fellowship, KDH is supported by an Australian Research Council Future Fellowship (FT130100637).

**Competing interests:** The authors have declared that no competing interests exist.

high discretionary food intake) and high sedentary behaviour ("unhealthy"); and (iii) high MVPA, poor diet (as defined above) and low sedentary time ("mixed").

## Conclusion

Within this sample, despite differences in the number of patterns derived by each method, a good degree of concordance across pattern characteristics was seen between the methods. Differences between patterns could be attributable to the underpinning statistical technique of each method. Therefore, acknowledging the differences between the methods and ensuring thorough documentation of the pattern derivation analyses is essential to inform comparison of patterns derived through a range of approaches across studies.

## Introduction

For children, diet and time spent in physical activity, sedentary behaviour, and sleep are key behaviours implicated in disease development [1–3]. Previous work has scrutinised the role of these behaviours individually; however, research regarding their synergistic or combined effects is an emerging area [3–5]. There is evidence an integrated approach that evaluates behavioural patterns will improve our understanding of complex health outcomes in children [3, 6–8].

Behavioural patterns are typically derived using unsupervised learning, a type of algorithm that does not involve a priori labelling or classification of the responses according to external criteria/guidelines [9, 10], but instead are data driven. The most common unsupervised learning methods used in the nutrition and physical activity field include principal component analysis (PCA), cluster analysis (CA), and latent class/profile analysis (LCA/LPA, for categorical and continuous input data respectively) [10]. A key distinction between the methods is their focus for grouping data (based on variables or individuals) and the resultant output type. PCA focuses on *variables*, identifying groups of highly correlated variables [11] using the covariance or correlation matrix [12] and transforming them into a smaller number of new, linearly uncorrelated variables known as principal components [13]. Each individual will have a 'score' for each of these principal components. Conversely, both CA and LCA/LPA focus on *individuals*, finding groups of individuals with similar characteristics [14] and assigning them into mutually exclusive clusters [13]. Individuals in different clusters have different patterns of the input variables. LCA/LPA also focus on individuals but assume that in the population there are "latent classes" or sub-types of individuals that have similar values of the input variables. Class membership of individuals is unknown but can be inferred from the input variables [15, 16]. All three methods can identify behaviours that are likely to co-occur in the same individual [10, 17]. Key differences across methods are further highlighted in Table 1.

A limited number of studies have compared behavioural patterns derived from multiple unsupervised learning methods within a single dataset. Two studies in adults compared dietary patterns, one [9] using PCA and CA and the other [18] using CA, PCA and index analysis (an investigator-driven method). They observed some similarity between pattern characteristics, despite different numbers of patterns being derived from the methods. They suggested that some direct comparisons are possible, however, the research question/objective should guide the choice of method for analyses. Based on the number of patterns obtained, they suggested those methods to be better if they provided more information, i.e., "patterns" from the sample, however, these conclusions are not generalizable. In another study [14], CA and LPA were

**Table 1. Brief summary of pattern derivation methods.**

| Methods | Principal component analysis | Cluster analysis | Latent profile analysis |
|---|---|---|---|
| Description | Data reduction technique that summarizes input variables into components which are linear combinations of the input variables. | Classification method that summarises input variables into clusters which are homogenous (comprising individuals with similar characteristics). | Classification method that summarizes input variables into latent profiles, where members are assigned probabilities of belonging to a profile. |
| Number of patterns | Initially equal to the number of input variables but number of patterns retained is based on criterion decided by the researcher (e.g., Horn's parallel analysis or eigenvalue cut-offs). | A number of criteria exist to decide clusters to be retained, which include the Calinksi-Harabasz statistic for K-means clustering. | Determined using statistical model-based criteria which include the BIC and aLMR test (used in the present study). Other common criteria include the AIC and the BLRT test. |
| Type of output variable | Continuous (component scores). | Categorical (cluster membership). | Categorical (after allocating each individual to their most probable profile membership). |

Abbreviations: BIC; Bayesian information criteria, aLMR; adjusted Lo-Mendel-Rubin test, AIC; Akaike information criteria, BLRT, bootstrap likelihood ratio test.

https://doi.org/10.1371/journal.pone.0255203.t001

used to derive different numbers of patterns of physical activity and sedentary behaviour in adolescents. The authors concluded that based on statistical technique, LPA overcomes certain limitations of CA and, therefore, may be a more reliable choice. Previous studies have been limited to patterns within behaviour domains; either diet or movement (physical activity/sedentary behaviour) [19]. To our knowledge, no studies have compared patterns derived from PCA, CA and LCA/LPA across lifestyle behaviours—diet, physical activity, sedentary behaviour, and sleep.

The various methods can produce different patterns (type/number) when applied to the same dataset [9, 14]. Differing results are possible even for the same approach due to the multiple decisions involved when deriving patterns, such as the definition/grouping/treatment of the initial variables, selection of the statistical criteria and/or subjective decisions made by the investigator [13, 14]. Lastly, interpretation of the output also influences the final patterns retained. Despite different solutions (i.e., number of and composition of patterns) being possible from different methods, under the assumption that a small number of behavioural patterns exist within the target population, it is of interest to explore if these methods identify at least a core set of behavioural patterns that are comparable. Previous studies have focused on identifying most suitable/ideal/common patterns in their study populations and only a few [14, 20] have discussed discrepancies that can arise from using different methods.

This paper aimed to investigate congruence between behavioural patterns across diet, physical activity, sedentary behaviour and sleep domains, derived using PCA, LPA and CA. This was assessed in a dataset of 6- to 8-year-old children.

## Methods

### Data source

Data were drawn from the Healthy Active Preschool and Primary Years (HAPPY) cohort. Details of the study are described in detail elsewhere [21]. In brief, the baseline sample (2008/09) included 1002 parents and their children aged 3-5y, recruited through preschools and childcare centres in Melbourne, Australia. Children were followed up at multiple time points (76% of the initial sample [n = 776] provided consent to be followed up). This study draws on data from the second wave in 2011/12 when children were aged 6-8y. In the second wave, 567 children participated, with 432 providing complete data who were included in this study. HAPPY was granted ethical approval from the Deakin University Human Research Ethics Committee (EC 291–2007), the Department of Education and Early Childhood Development

(2011_001008) and the Catholic Education Office (1714). At each time point, parents provided written informed consent for themselves and their child to participate.

## Measures and data management

**Diet.** Parents reported child dietary intake using a previously validated 15-item food frequency questionnaire [22]. Included items were those showing acceptable reliability (ICC > 0.6) and validity. Frequency of a range of discretionary foods eaten in the previous week was recorded using a 7-point Likert scale (0–6 or more times). Six sweet (spreads [peanut butter or Nutella], pre-sugared cereals, bakery items, lollies and snack bars, chocolate, ice-cream) and seven savoury (potato crisps or savoury biscuits, cheese and cheese spreads, pies and sausage rolls, pizza, hot chips or French fries, hot dogs and processed meats and takeaway foods) discretionary food items were summed and divided by seven to obtain daily intakes. The frequency of fruit and vegetables consumed in the past 24 hours (representing daily intake) were recorded using a 6-point Likert scale (0–5 or more times).

**Physical activity.** Physical activity was objectively measured using ActiGraph GT1M uniaxial accelerometers (Pensacola, FL, USA), worn for eight consecutive days. Accelerometers were worn at the hip, during waking hours and removed for water-based activities. Non-wear time > 20 minutes was classified as zero counts. Accelerometer data were considered valid if data were recorded for at least eight hours a day for four or more days, inclusive of at least one weekend day [23]. Accelerometer data were classified as moderate- to vigorous-intensity physical activity (MVPA) for counts >2296/min [23]. Additionally, parents reported information on total weekly duration of different organised sports their child engaged in, using a reliable survey [24]. These included swimming, gymnastics, dance, football, soccer, netball, basketball, and cricket, and also accounted for other sports not included in the list with an open choice 'other' category. Total weekly organised sport duration was obtained by summing the total weekly duration for each sport. This was divided by seven to derive daily equivalents. Parents also reported the time their child spent in outdoor play on a typical weekday and weekend day. This was weighted to obtain an average daily time in minutes, i.e., the sum of daily weekday time and daily weekend time averaged over five and two days respectively, divided by seven. Test retest reliability was acceptable for organised sport (ICC = 0.74) and outdoor play (ICC = 0.44) [24].

**Sedentary behaviour.** Sedentary behaviour was assessed by accelerometry (as described above) and parent-report. Accelerometer data of <100 counts per minute was classified as sedentary time [25] and reported as minutes per day. Parents reported the total number of hours their child usually spent in leisure-time sedentary behaviours during the week (Monday to Friday) and on weekends (Saturday and Sunday). Evidence suggests that children's screen time can be both active (e.g., videogames) and passive (e.g., television viewing) in terms of child engagement, and different modes can have differential effects on health [26]. These behaviours were therefore treated separately and categorized into screen time (television viewing and computer use excluding games), video game use (including computer games and handheld electronic games) and quiet playtime. When assessed for test-retest reliability these items showed low to moderate acceptability (ICC = 0.10 [quiet play], ICC = 0.44 [screen time] and ICC = 0.68 [video game use]) [21]. Daily duration for each of these categories was calculated by summing the weekday and weekend time duration and dividing by seven. The low reliability of these items was inferred to be due to true day-to-day variability in these behaviours assessed two weeks apart, rather than responses being inaccurate and unreliable themselves [24].

**Sleep.** Parents reported their child's usual nightly sleep duration in hours and minutes per night. This item showed good test retest reliability (ICC = 0.82) [24].

## Data analysis

Stata 15.0 (StataCorp, Texas, USA) was used for CA and PCA and Mplus 8.0 for LPA. All analyses included four dietary (intake of fruit, vegetables, sweet and savoury discretionary items), three physical activity (MVPA from accelerometry, organised sport duration, and outdoor playtime), four sedentary behaviour (sedentary time from accelerometry, screen time, videogames, and quiet playtime) and one sleep (daily sleep duration) variable. Using the residuals obtained from regressing accelerometry data on wear time, accelerometer variables (both MVPA and sedentary time) were adjusted for daily wear time (26). Input variables were converted to standardised scores (mean = 0, SD = 1) as they were measured using different scales.

**Cluster analysis.** Non-hierarchical K-means CA was performed. Since this method requires pre-specification of the number of clusters to be generated, a range of cluster solutions (2–10 clusters) were derived. The Calinski-Harabasz statistic, which measures improvement in some measures of fit between models, was then used to determine the optimal number of clusters [27]. Higher values for this statistic indicate better fit when comparing different number of clusters [27]. Post-assessment of the values from the Calinski-Harabasz statistic, the size of the clusters across the different cluster solutions and their interpretability and meaningfulness [13, 28] were considered to determine the final number of clusters. The Calinski-Harabasz statistic indicated a 2-cluster model as optimal; however, upon assessment of cluster size, it was found that two large and distinct clusters from the 3-cluster model were combined in this solution, with the remaining cluster unchanged across the 2- and 3- cluster models. Upon inspection, the 3-cluster model was considered more informative and hence selected.

**Latent profile analysis.** Since LPA requires pre-specification of the number of profiles (homogeneous subgroups), multiple profile models (2–10 profiles) were derived and then compared using two recommended model fit criteria; the Bayesian information criteria (BIC) and the adjusted Lo-Mendel-Rubin (aLMR) test. Models with lower values for the BIC indicate better fit [29]. The aLMR test compares fit between neighbouring profile models [16] and when the test does not reject the null-hypothesis the model with smaller number of classes is preferred [29]. BIC and aLMR model estimation criteria were inconsistent, suggesting 9-profile and 3-profile models, respectively, as optimal. Upon inspection, the 3-profile model was more interpretable and logical, hence selected. A general cut point of ±0.2 was used to interpret behaviours that were high/low in the patterns derived from LPA and CA.

**Principal component analysis.** The number of principal components to retain was determined using Horn's parallel analysis, where eigenvalues are compared to the mean of eigenvalues estimated from a Monte-Carlo simulated matrix of the same size. Horn's parallel analysis accounts for the sample bias in estimating the eigenvectors and eigenvalues and provides a more accurate and advanced approach than the Kaiser criterion which keeps principal components with eigenvalue greater than one [13, 30]. Varimax rotation was performed to obtain more interpretable component loadings and subsequently more contrasted patterns. When interpreting the derived principal components coefficients (loadings) all input variables were considered.

As these three statistical methods are potentially sensitive to outliers, the same analyses (results not shown) were conducted excluding outliers. Only minor differences (e.g., slightly different coefficients) in the pattern solutions were observed therefore results are reported on the full data set to maximise sample size and generalisability of results.

## Results

Participants (n = 432) with complete survey and accelerometer data were included in all analyses. Sample characteristics are presented in Table 2. Four distinct patterns were derived from

**Table 2. Characteristics of the sample included in the analysis (n = 432).**

| Characteristic | Mean (SD) | Median (IQR) | Range |
|---|---|---|---|
| Child age (years) | 7.6 (0.7) | 7.6 (7.0–8.2) | 5.4–9.1 |
| Sex [n (%)] | | - | - |
| Male | 244 (56.5) | | |
| Female | 188 (43.5) | | |
| Diet | | | |
| Fruit intake (times/day) | 2.3 (1.3) | 2.0 (1.0–3.0) | – 5.0 |
| Vegetable intake (times/day) | 2.9 (1.3) | 3.0 (2.0–4.0) | 0.0–5.0 |
| Sweet discretionary food intake (times/day) | 1.5 (0.7) | 1.4 (1.0–2.0) | 0.0–3.7 |
| Savoury discretionary food intake (times/day) | 1.2 (0.5) | 1.1 (0.8–1.6) | 0.0–2.6 |
| Physical activity | | | |
| Organised sport (mins/day) | 24.2 (20.0) | 21.4 (12.9–32.1) | 0.0–197.1 |
| Outdoor play (mins/day) | 143.0 (77.5) | 132.1 (94.3–180/0) | 6.4–617.1 |
| MVPA (mins/day)[a] | 106.6 (28.0) | 104.7 (86.1–124.1) | 42.9–190.0 |
| Sedentary behaviour | | | |
| Screen time (mins/day) | 93.0 (57.9) | 77.1 (47.1–128.6) | – 321.4 |
| Videogames (mins/day) | 23.2 (30.7) | 12.9 (0.0–34.3) | – 214.3 |
| Quiet play time (mins/day) | 52.5 (39.7) | 42.9 (25.7–68.6) | 0.0–342.9 |
| Sedentary time (mins/day)[a] | (41.7) | 362.3 (332.5–390.4) | 247.9–487.7 |
| Sleep (mins/day) | 623.3 (47.4) | 630 (600.0–660.0) | 450.0–720.0 |

Abbreviations: mins–minutes, IQR–interquartile range, MVPA–moderate- to vigorous-intensity physical activity, PA–physical activity, SB–sedentary behaviour, SD–standard deviation.

[a] Mean accelerometer wear time was 703 minutes per day.

https://doi.org/10.1371/journal.pone.0255203.t002

PCA, and three from CA and LPA. Despite these differences, each method identified a pattern that could be described as 'healthy'; that is, reflective of a healthy diet, high physical activity, low sedentary behaviour, and high sleep duration. Patterns with characteristics contrary to healthy behaviours were also found using each of the three methods and were described as 'unhealthy'. Lastly, patterns containing a mixture of healthy and unhealthy behaviours were identified using all three methods and were described as 'mixed'. Across LPA and CA, a large proportion of children (76%) were classified into the same pattern type (either unhealthy/unhealthy or mixed) whilst the remaining children were classified inconsistently. Patterns derived from each method are presented in Figs 1–3 and mean z-scores/component loadings are presented in S1 Table.

Cluster Analysis using K-means identified three clusters (see Fig 1). Cluster 1 (n = 133), labelled '**unhealthy**', was characterised by lowest fruit and vegetable intake, lowest overall physical activity (organized sport, outdoor play duration and MVPA), lowest sleep, highest sweet discretionary food intake and highest sedentary behaviour (screen, videogame, and sedentary time). Cluster 2 (n = 102), labelled '**active healthy eaters**' (healthy), was characterised by highest fruit and vegetable intake, highest outdoor play time, lowest discretionary food intake and lowest screen time. Cluster 3 (n = 197), labelled '**active sleepers, non-sedentary unhealthy eaters**' (mixed), was characterised by highest MVPA levels, highest sleep duration, low fruit intake and lowest videogame and sedentary time.

For LPA (see Fig 2), Profile 1 (n = 206) was labelled '**unhealthy**' and comprised of lowest fruit and vegetable intake, lowest outdoor play time and MVPA, and highest sedentary time. Profile 2 (n = 84), labelled '**active healthy eaters**' (healthy), was characterised by highest fruit and vegetable intake and highest outdoor play time. Profile 3 (n = 142), '**active non-sedentary**

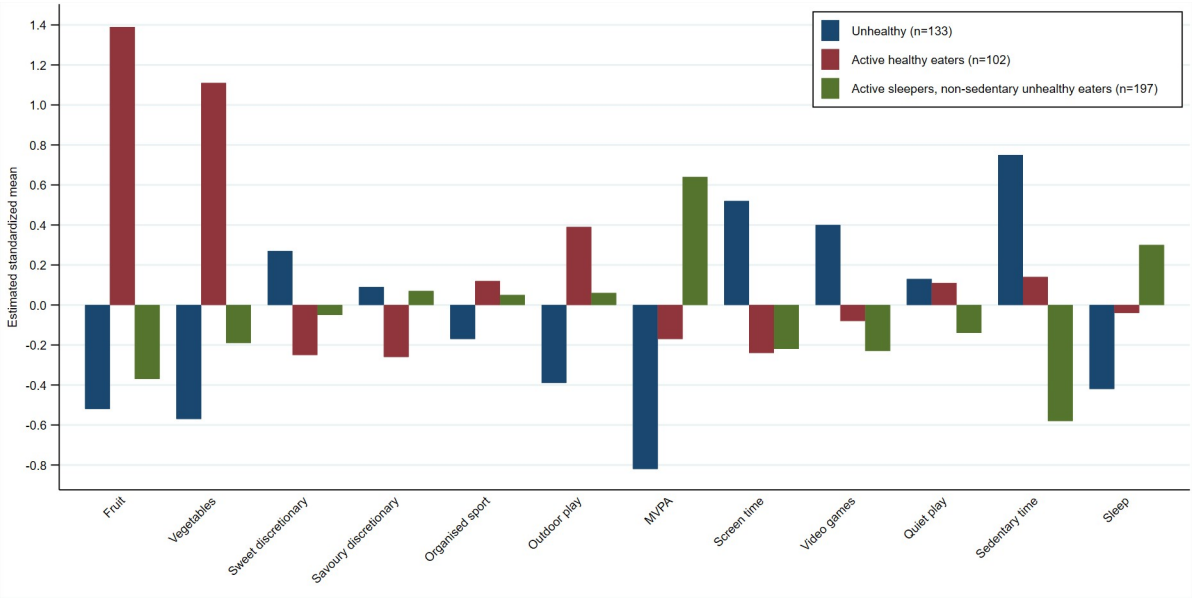


**Fig 1. Average value of the standardized input variables for each behavioural pattern derived using K-means cluster analysis.** Footnote: n; number of children in each pattern.

https://doi.org/10.1371/journal.pone.0255203.g001

**unhealthy eaters**' (mixed), was characterised by highest savoury discretionary food intake, highest MVPA, high outdoor play time, low fruit intake and lowest sedentary time.

Four principal components were retained according to Horn's parallel analysis, which explained 54% of the total variance. These four patterns (see Fig 3) were labelled: '**active sleepers, non-sedentary unhealthy eaters**' (mixed), (explained 16% of variance), characterised by intake of savoury discretionary food items, high MVPA, adequate sleep duration and low

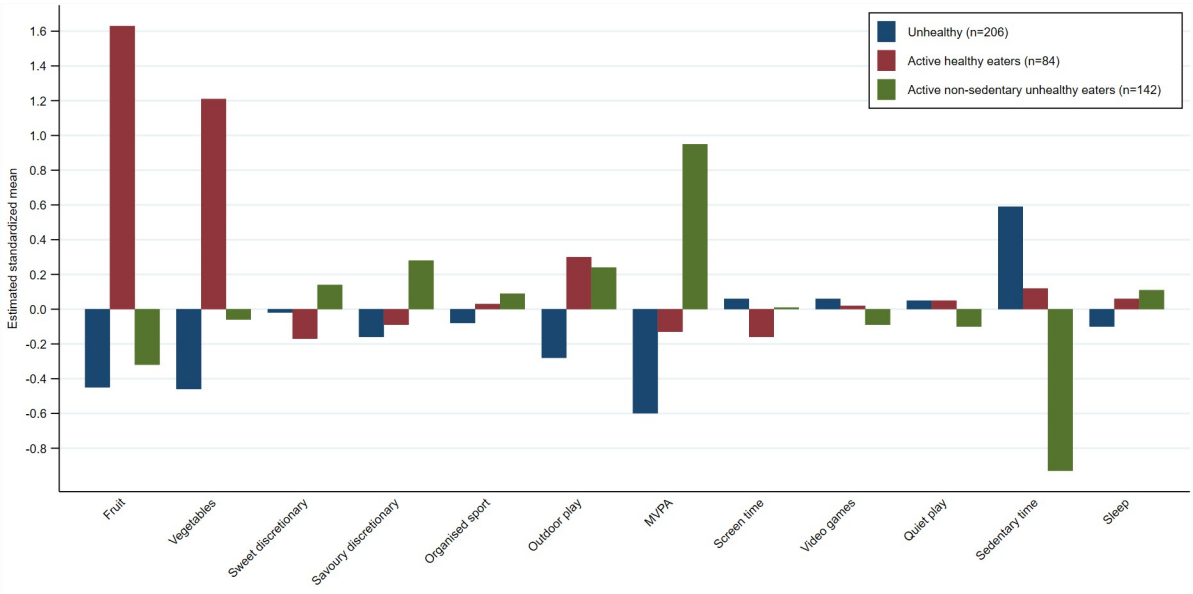


**Fig 2. Average value of the standardized input variables for the behavioural patterns derived using latent profile analysis.** Footnote: n; number of children in each pattern.

https://doi.org/10.1371/journal.pone.0255203.g002

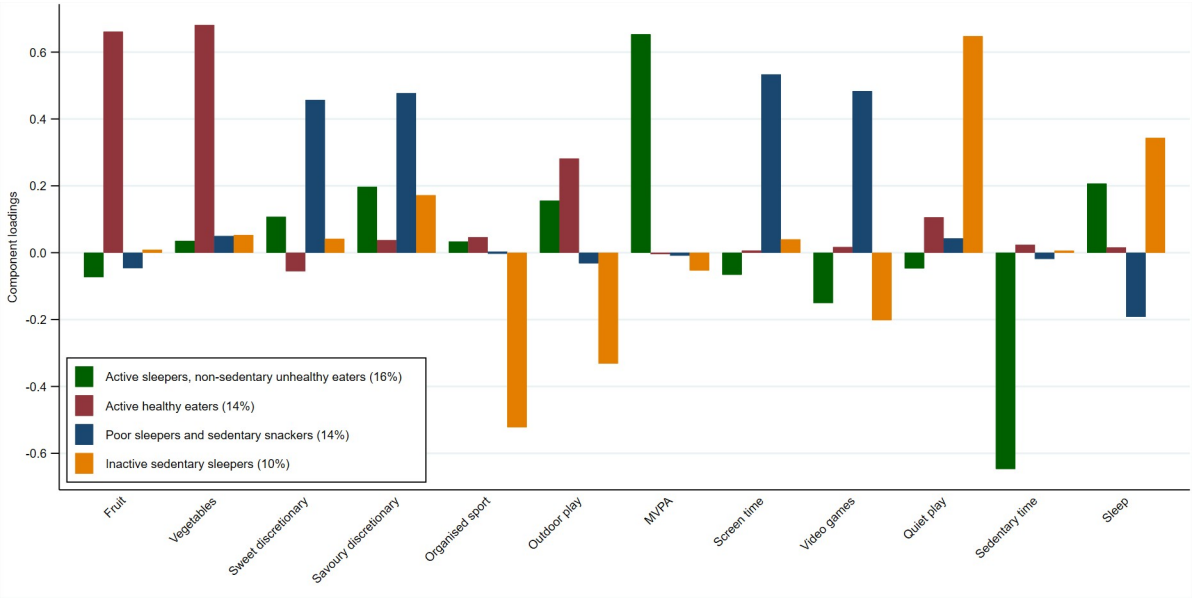


**Fig 3. Component loadings for each behavioural pattern derived using principal component analysis.** Footnote: The percentage after each pattern label is the percent of the total data variance explained by that pattern.

https://doi.org/10.1371/journal.pone.0255203.g003

sedentary time; '**active healthy eaters**' (healthy), (explained 14% of variance), comprising high fruit and vegetable consumption and outdoor play time; '**poor sleepers and sedentary snackers**' (unhealthy), (explained 14% of variance), comprised of high discretionary food intake, high screen and videogame time, and lowest sleep duration; and '**inactive sedentary sleepers**' (mixed), (explained 10% of variance), characterised by high quiet play time and sleep, and low organised sport and outdoor play time.

A three-component PCA model with varimax rotation (explaining 44% of the total variance) was additionally derived to allow direct comparison with the three patterns obtained from CA and LPA. The patterns identified were the same as the first three patterns from the four component PCA solution (results not presented).

## Discussion

To our knowledge, this was the first study to compare patterns derived from four behavioural domains (diet, physical activity, sedentary behaviour, and sleep) in children using three different analytic methods, PCA, LPA and CA. Patterns identified by different methods were not identical; however, some similar underlying themes emerged between the three methods indicating that such methods can identify 'a core set' of underlying patterns. The number of patterns identified were not uniform across all methods, however, each method identified a healthy, unhealthy, and mixed behavioural pattern. These core patterns were on the whole comparable to those reported in previous reviews [2, 19, 31] of behavioural patterns in children, where healthy, unhealthy, and mixed patterns have been identified despite variation in the specifics of such patterns and in the underlying behaviours analysed in different studies. It is, however, apparent that clear reporting of analytical techniques used and decisions made is important to enable assessment of comparability across studies given such decisions do influence the finer detail of the patterns derived.

Patterns derived across methods in the present study were comparable to previous studies. Healthy patterns characterised by outdoor play and healthy eating have similarly been reported

in three other studies [32–35], all using PCA. Similar unhealthy patterns comprising an unhealthy diet and high sedentary behaviour were reported in nine studies, five using CA [28, 36–39] and four using PCA [32, 34, 35, 40]. The mixed pattern, comprising high MVPA, high discretionary food intake and low sedentary time, was somewhat similar to the patterns derived by Seghers et al. [41]. However, the other mixed pattern from our study, characterised by high quiet play time, sleep and low organised sport and outdoor play time is unique. Few previous studies have included objectively assessed physical activity data in their analyses which may explain why this pattern has not been reported previously. While similar patterns can be identified in the literature, they have not been specifically observed within a single previous study. The discrepancies in pattern solutions possible can be attributed not only to the specific input variables used to assess behaviours in each study or to cultural differences between the study populations, but also to specific statistical criteria chosen and subjective decisions made when applying these methods, and also due to a broader range of behaviours included in this study. Future studies should consider detailed documentation of their methodological approaches to enable better comparisons of patterns across studies. Standardised reporting for these methodological approaches is also warranted.

When a three-pattern solution was compared across all methods, high concordance was observed with all methods identifying a healthy, unhealthy and a mixed pattern. As expected, patterns derived from LPA and CA were more similar than PCA. Additionally, most children were classified consistently into the same pattern type across LPA and CA, however, these comparisons are not as straightforward with PCA as children have scores for all patterns. PCA is a dimension-reduction method that looks for correlated variables, whereas LPA and CA look for similarities between individuals [11, 13]. One quarter of the sample was classified into different pattern types by CA and LPA; this could be due to the approach used by each method to assign pattern membership. CA uses a distance-based similarity index, whereas LPA estimates individual probabilities for each pattern derived and then assigns an individual into their most probable pattern. Despite underlying differences of how these methods derive patterns, the three approaches identified patterns that were promotive (healthy) and demotive (unhealthy, mixed) of health. This suggests there is reasonable overlap in the patterns derived from these different methods, with the choice of method more dependent on the preferred output type (categorical/continuous) or statistical technique (probabilistic/geometrical etc.). Nonetheless, given that the patterns derived were not identical (as expected, given the differing underlying algorithms used for pattern derivation by the different methods) and a quarter of the sample were classified into different pattern types inconsistently across two methods (CA and LPA), direct comparisons of patterns from studies using different methods must be interpreted with caution. These findings suggest that the method chosen to derive patterns may influence conclusions about associations of behavioural patterns with health outcomes. Studies assessing comparisons of associations with health outcomes using patterns derived from different methods are warranted to confirm this.

Each method has inherent strengths and limitations and a set of analyst decisions involved. Although within each method there is no absolute correct model, given the true underlying population patterns are unknown, it is essential authors justify decisions made and the final model chosen. For example, in our case, choosing the three instead of the four-pattern PCA model would not be incorrect, however, it explains a smaller proportion of the data total variance (44% versus 54%) and hence was deemed less ideal. Horn's parallel analysis provides an objective way to determine the number of principal components [30] but has not been frequently used or reported in previous studies using PCA, with the more inaccurate/biased Kaiser criteria typically used. Horn's parallel analysis generally outperforms approaches such as the Kaiser criteria, yet due to it being more computationally intensive, simpler methods such

as the Kaiser criteria had much higher adoption [42]. This practice has largely persisted despite today's computers being adequate for Horn's parallel analyses in most cases. Standard implementation of this technique across widely used statistical packages is also lacking [42]. The use of Horn's parallel analysis may change the number of components to be selected and thus may be useful for future investigators to consider, with the potential to provide greater consistency across studies. PCA provides scores for each sample member for all patterns derived and is useful for those studies requiring continuous variables for subsequent analysis [11]. Patterns derived from CA can vary widely due to the vast number of possibilities for the clustering algorithm and the parameter settings. Additionally, the heterogeneity of cluster members is ignored once clusters have been defined, as members within a cluster are considered homogenous. This may be disadvantageous as the internal variation may be extremely large, particularly when the number of clusters is small [18]. Large internal variation can introduce classification errors, thus impairing statistical power in subsequent analyses. Evaluating the best model fit for LPA was challenging as the preferred number of profiles based on the two criteria were not concordant. Although the BIC has been shown to outperform other model estimation criteria [16, 43], the model it suggested was not logical or interpretable. In comparison, the aLMR test provided a clear optimal number of profiles which were most interpretable in this study. Latent profile analysis is superior to cluster analysis in its ability to (a) provide individual probabilities for the different profiles identified, (b) account for missing data and (c) include health outcomes while deriving patterns, all while considering the probability of misclassification [20]. Although PCA produces variables/components being a data reduction method, and CA and LPA are classification methods which group individuals, they are nonetheless somewhat comparable as they indirectly provide similar information; variables from PCA inform us of behavioural patterns expressed by individuals, whereas groups of individuals from CA and LPA express behavioural patterns. It is important to consider the strengths and limitations of each method within the context of the individual study and the research question being asked to decide on the most appropriate method.

The study is not without some limitations and strengths. This study did not investigate an exhaustive list of statistical methods to derive patterns, rather opting for the most common methods reported in the literature. Within each method, several pattern solutions are possible and analyst decisions influence outcomes. Rather than investigating all possible alternatives, we utilised the most common approaches reported in the literature within each method. Limitations of the individual methods are described above, however, there are also some limitations common across all methods. These pattern derivation methods are sensitive to outliers and can be influenced by the distribution of the input variables [11], however, the fairly large sample in this study provided some protection against this. Future studies could also benefit from a larger sample size, as this study sample was limited by the availability of complete data for all 12 behaviours included in the pattern derivation analyses.

The final patterns obtained are dependent on decisions taken, given there is no single approach to derive patterns. Model estimation criteria may not always provide logical patterns (as seen for LPA in this study), therefore, future studies should consider balancing objective (model estimation criteria) and subjective decision making. This will ensure the patterns defined are valid based on the estimation criteria but also are pragmatic and logical using sound subjective decision-making. This overcomes some limitations of using a-priori methods which are predominantly analyst decision driven involving pre-classification of behaviours, thereby affecting final patterns derived. Most importantly, thorough documentation of all decisions taken will be crucial to assess comparability of their findings with other studies.

We acknowledge potential recall and social desirability bias introduced by using survey data [44, 45], in addition to the reliability of some items in the survey being low. Parent-report

data can limit measurement precision (explaining some of the discrepancies across methods) and also lead to over/under estimations of healthy/unhealthy behaviours respectively, due to social desirability bias, thereby potentially affecting the final patterns derived and subsequent associations. Given our primary aim was to compare patterns from different techniques in a single dataset, and since most previous studies used subjective measures, this is less concerning. The use of subjective and objective measures provides some balance between accuracy and context-rich information for some behaviours (physical activity and sedentary behaviour). Although objective measures help refine the accuracy of patterns obtained, using them alone cannot distinguish different activity types (context-based) within behavioural domains. The inclusion of quiet play time (in the sedentary behaviour domain) and sleep duration is novel, as these variables have not been frequently included in previous pattern derivation analyses. The diversity of the variables included across the behavioural domains to derive patterns is an additional strength of this study as it helps to better understand the interplay of these behaviours, ultimately valuable in health prevention efforts.

In summary, our findings indicate that while there are some similarities in patterns identified using different methods, there are also notable differences. The similarity in pattern characteristics across methods may help provide some confidence in the underlying patterns prevalent in a given dataset. However, the differences observed across patterns have potential to influence subsequent analyses of associations with various outcomes. Such differences are yet to be investigated, but they suggest that the choice of method may influence associations with health outcomes and might explain the discrepancy in findings reported across studies. Thus, comparison of findings across studies employing different pattern derivation methods should be done with caution. Whilst the goal of this study was not to recommend one method over the other, as each has strengths and weaknesses, future researchers should consider the choice of method based on their study objectives and subsequent analyses.

## Conclusion

Typically, health behaviours (e.g., diet, physical activity, sedentary behaviour, and sleep) are considered individually as predictors of health. However, these behaviours do not occur in isolation and are likely patterned as shown in the present study. Consequently, many studies are now deriving and reporting patterns of behaviour to help describe their synergistic influence. Multivariate methods such as PCA, LPA and CA are useful in identifying behavioural patterns in a given dataset. In this study, each of the three methods identified a core set of underlying patterns characteristic of a healthy, unhealthy and mixed pattern. The similarities provide greater confidence that the patterns are present in the target population. However, the patterns identified by the different methods were not identical. The differences can be attributable to the algorithms underpinning each method and highlight the importance of documenting not only the methods used, but the objective and subjective decisions taken in the analytic process. Overall, comparison of patterns at a broad level using different methods is possible. However, when comparing the finer details of pattern characteristics across studies utilising different methods, it is important to be mindful that the differences may be an artefact of the statistical techniques used rather than reflective of true differences in the samples.

## Supporting information

**S1 Table. Pattern characteristics for each method–component loadings (for PCA) and mean z-scores (for LPA/CA).** Abbreviations: CA–cluster analysis, LPA–latent profile analysis, mins–minutes, MVPA–moderate- to vigorous-intensity physical activity, PA–physical activity,

PCA–principal component analysis, SB–sedentary behaviour.
(DOCX)

## Acknowledgments

The authors thank the parents and children who participated in the HAPPY study and the researchers and staff involved for their hard work and efforts.

## Author Contributions

**Conceptualization:** Ninoshka J. D'Souza, Katherine Downing, Gavin Abbott, Liliana Orellana, Sandrine Lioret, Karen J. Campbell, Kylie D. Hesketh.

**Formal analysis:** Ninoshka J. D'Souza, Gavin Abbott.

**Funding acquisition:** Kylie D. Hesketh.

**Methodology:** Liliana Orellana.

**Project administration:** Kylie D. Hesketh.

**Supervision:** Katherine Downing, Gavin Abbott, Liliana Orellana, Sandrine Lioret, Karen J. Campbell, Kylie D. Hesketh.

**Writing – original draft:** Ninoshka J. D'Souza.

**Writing – review & editing:** Katherine Downing, Gavin Abbott, Liliana Orellana, Sandrine Lioret, Karen J. Campbell, Kylie D. Hesketh.

## References

1. Chaput JP, Saunders TJ, Carson V. Interactions between sleep, movement and other non-movement behaviours in the pathogenesis of childhood obesity. Obesity Reviews. 2017; 18:7–14. https://doi.org/10.1111/obr.12508 PMID: 28164448
2. Gubbels JS, van Assema P, Kremers SP. Physical Activity, Sedentary Behavior, and Dietary Patterns among Children. Curr Nutr Rep. 2013; 2(2):105–12. https://doi.org/10.1007/s13668-013-0042-6 PMID: 23638341
3. Mihrshahi S, Gow ML, Baur LA. Contemporary approaches to the prevention and management of paediatric obesity: an Australian focus. 2018. p. 267–74. https://doi.org/10.5694/mja18.00140 PMID: 30208819
4. Kovács E, Hunsberger M, Reisch L, Gwozdz W, Eiben G, De Bourdeaudhuij I, et al. Adherence to combined lifestyle factors and their contribution to obesity in the IDEFICS study. Obesity Reviews: An Official Journal Of The International Association For The Study Of Obesity. 2015; 16 Suppl 2:138–50. https://doi.org/10.1111/obr.12349 PMID: 26707023
5. Mayne SL, Virudachalam S, Fiks AG. Clustering of unhealthy behaviors in a nationally representative sample of U.S. children and adolescents. Preventive Medicine. 2020; 130.
6. Pronk NP, Anderson LH, Crain AL, Martinson BC, O'Connor PJ, Sherwood NE, et al. Meeting recommendations for multiple healthy lifestyle factors: Prevalence, clustering, and predictors among adolescent, adult, and senior health plan members. American Journal of Preventive Medicine. 2004; 27(2, Supplement):25–33.
7. Olds T, Sanders I, Maher C, Fraysse F, Bell L, Leslie E. Does compliance with healthy lifestyle behaviours cluster within individuals in Australian primary school-aged children? Child: Care, Health and Development. 2018; 44(1):117–23. https://doi.org/10.1111/cch.12497 PMID: 28736955
8. Parker KE, Salmon J, Brown HL, Villanueva K, Timperio A. Typologies of adolescent activity related health behaviours. Journal of Science and Medicine in Sport. 2019; 22(3):319–23. https://doi.org/10.1016/j.jsams.2018.08.015 PMID: 30190099
9. Hearty ÁP, Gibney MJ. Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults. British Journal of Nutrition. 2008; 101(4):598–608. https://doi.org/10.1017/S0007114508014128 PMID: 18577300

10. McAloney K, Graham H, Law C, Platt L. A scoping review of statistical approaches to the analysis of multiple health-related behaviours. Preventive Medicine. 2013; 56(6):365–71. https://doi.org/10.1016/j.ypmed.2013.03.002 PMID: 23518213

11. Newby PK, Tucker KL. Empirically Derived Eating Patterns Using Factor or Cluster Analysis: A Review. Nutrition Reviews. 2004; 62(5):177–203. https://doi.org/10.1301/nr.2004.may.177-203 PMID: 15212319

12. Hoffmann K, Schulze MB, Schienkiewitz A, Nöthlings U, Boeing H. Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. American Journal of Epidemiology. 2004; 159(10):935–44. https://doi.org/10.1093/aje/kwh134 PMID: 15128605

13. Gleason PM, Boushey CJ, Harris JE, Zoellner J. Publishing Nutrition Research: A Review of Multivariate Techniques—Part 3: Data Reduction Methods. Journal of the Academy of Nutrition and Dietetics. 2015; 115(7):1072–82. https://doi.org/10.1016/j.jand.2015.03.011 PMID: 25935571

14. Beets MW, Foley JT. Comparison of 3 Different Analytic Approaches for Determining Risk-Related Active and Sedentary Behavioral Patterns in Adolescents. Journal of Physical Activity & Health. 2010; 7(3):381–92. https://doi.org/10.1123/jpah.7.3.381 PMID: 20551496

15. Bucholz KK, Hesselbrock VM, Heath AC, Kramer JR, Schuckit MA. A latent class analysis of antisocial personality disorder symptom data from a multi-centre family study of alcoholism. Addiction. 2000; 95(4):553–67. https://doi.org/10.1046/j.1360-0443.2000.9545537.x PMID: 10829331

16. Nylund KL, Asparouhov T, Muthén BO. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. Structural Equation Modeling: A Multidisciplinary Journal. 2007; 14(4):535–69.

17. Boone-Heinonen J, Gordon-Larsen P, Adair LS. Obesogenic clusters: Multidimensional adolescent obesity-related behaviors in the U.S. Annals of Behavioral Medicine. 2008; 36(3):217–30. https://doi.org/10.1007/s12160-008-9074-3 PMID: 19067096

18. Reedy J, Wirfält E, Flood A, Mitrou PN, Krebs-Smith SM, Kipnis V, et al. Comparing 3 Dietary Pattern Methods—Cluster Analysis, Factor Analysis, and Index Analysis—With Colorectal Cancer Risk: The NIH–AARP Diet and Health Study. American Journal of Epidemiology. 2009; 171(4):479–87. https://doi.org/10.1093/aje/kwp393 PMID: 20026579

19. D'Souza NJ, Kuswara K, Zheng M, Leech R, Downing KL, Campbell KJ, et al. A systematic review of lifestyle patterns and their association with adiposity in children aged 5–12 years. Obesity Reviews. 2020; 21(8). https://doi.org/10.1111/obr.13029 PMID: 32297464

20. Magidson J, Vermunt JK. Latent class models for clustering: a comparison with K-means. Canadian Journal of Marketing Research. 2002; 20(1):37–44.

21. Hinkley T, Timperio A, Salmon J, Hesketh K. Does Preschool Physical Activity and Electronic Media Use Predict Later Social and Emotional Skills at 6 to 8 Years? A Cohort Study. Journal of Physical Activity & Health. 2017; 14(4):308–16. https://doi.org/10.1123/jpah.2015-0700 PMID: 28169562

22. Magarey A, Golley R, Spurrier N, Goodwin E, Ong F. Reliability and validity of the Children's Dietary Questionnaire; A new tool to measure children's dietary patterns. International Journal of Pediatric Obesity. 2009; 4(4):257–65. https://doi.org/10.3109/17477160902846161 PMID: 19922040

23. Abbott G, Hnatiuk J, Timperio A, Salmon J, Best K, Hesketh KD. Cross-sectional and Longitudinal Associations Between Parents' and Preschoolers' Physical Activity and Television Viewing: The HAPPY Study. Journal of Physical Activity and Health. 2016; 13(3):269–74. https://doi.org/10.1123/jpah.2015-0136 PMID: 26181513

24. Hinkley T, Salmon J, Okely AD, Crawford D, Hesketh K. The HAPPY Study: Development and reliability of a parent survey to assess correlates of preschool children's physical activity. Journal of Science and Medicine in Sport. 2012; 15(5):407–17. https://doi.org/10.1016/j.jsams.2011.12.009 PMID: 22480665

25. Carson V, Salmon J, Crawford D, Hinkley T, Hesketh KD. Longitudinal levels and bouts of objectively measured sedentary time among young Australian children in the HAPPY study. Journal of Science and Medicine in Sport. 2016; 19(3):232–6. https://doi.org/10.1016/j.jsams.2015.01.009 PMID: 25683731

26. Sweetser P, Johnson D, Ozdowska A, Wyeth P. Active versus passive screen time for young children. Australasian Journal of Early Childhood. 2012; 37(4):94–8.

27. Halpin B. Cluster Analysis Stopping Rules in Stata University of Limerick, Sociology Do; 2016.

28. Dumuid D, Olds T, Lewis LK, Martin-Fernández JA, Barreira T, Broyles S, et al. The adiposity of children is associated with their lifestyle behaviours: a cluster analysis of school-aged children from 12 nations. Pediatric Obesity. 2018; 13(2):111–9. https://doi.org/10.1111/ijpo.12196 PMID: 28027427

29. Magee CA, Caputi P, Iverson DC. Patterns of health behaviours predict obesity in Australian children. J Paediatr Child Health. 2013; 49(4):291–6. https://doi.org/10.1111/jpc.12163 PMID: 23574555

30. Dinno A. Implementing Horn's parallel analysis for principal component analysis and factor analysis. 2009. p. 291–8.

31. Leech RM, McNaughton SA, Timperio A. The clustering of diet, physical activity and sedentary behavior in children and adolescents: a review. Int J Behav Nutr Phys Act. 2014; 11:4. https://doi.org/10.1186/1479-5868-11-4 PMID: 24450617

32. Yannakoulia M, Ntalla I, Papoutsakis C, Farmaki A-E, Dedoussis GV. Consumption of Vegetables, Cooked Meals, and Eating Dinner is Negatively Associated with Overweight Status in Children. The Journal of Pediatrics. 2010; 157(5):815–20. https://doi.org/10.1016/j.jpeds.2010.04.077 PMID: 20955852

33. Lioret S, Campbell KJ, McNaughton SA, Cameron AJ, Salmon J, Abbott G, et al. Lifestyle Patterns Begin in Early Childhood, Persist and Are Socioeconomically Patterned, Confirming the Importance of Early Life Interventions. Nutrients. 2020; 12(3).

34. Lioret S, Touvier M, Lafay L, Volatier JL, Maire B. Dietary and physical activity patterns in French children are related to overweight and socioeconomic status. Journal of Nutrition. 2008; 138(1):101–7. https://doi.org/10.1093/jn/138.1.101 PMID: 18156411

35. Gubbels JS, Kremers SP, Goldbohm RA, Stafleu A, Thijs C. Energy balance-related behavioural patterns in 5-year-old children and the longitudinal association with weight status development in early childhood. Public Health Nutrition. 2012; 15(8):1402–10. https://doi.org/10.1017/S1368980011003089 PMID: 22124196

36. Cameron AJ, Crawford DA, Salmon J, Campbell K, McNaughton SA, Mishra GD, et al. Clustering of Obesity-Related Risk Behaviors in Children and Their Mothers. Annals of Epidemiology. 2011; 21(2):95–102. https://doi.org/10.1016/j.annepidem.2010.11.001 PMID: 21184950

37. Fernández-Alvira JM, De Bourdeaudhuij I, Singh AS, Vik FN, Manios Y, Kovacs E, et al. Clustering of energy balance-related behaviors and parental education in European children: The ENERGY-project. International Journal of Behavioral Nutrition and Physical Activity. 2013;10. https://doi.org/10.1186/1479-5868-10-5 PMID: 23320538

38. Leech RM, McNaughton SA, Timperio A. Clustering of children's obesity-related behaviours: Associations with sociodemographic indicators. European Journal of Clinical Nutrition. 2014; 68(5):623–8. https://doi.org/10.1038/ejcn.2013.295 PMID: 24424077

39. Leech RM, McNaughton SA, Timperio A. Clustering of diet, physical activity and sedentary behaviour among Australian children: Cross-sectional and longitudinal associations with overweight and obesity. International Journal of Obesity. 2015; 39(7):1079–85. https://doi.org/10.1038/ijo.2015.66 PMID: 25907316

40. Rodenburg G, Oenema A, Pasma M, Kremers SPJ, van de Mheen D. Clustering of food and activity preferences in primary school children. Appetite. 2013; 60(1):123–32. https://doi.org/10.1016/j.appet.2012.10.007 PMID: 23085278

41. Seghers J, Rutten C. Clustering of multiple lifestyle behaviours and its relationship with weight status and cardiorespiratory fitness in a sample of Flemish 11- to 12-year-olds. Public Health Nutrition. 2010; 13(11):1838–46. https://doi.org/10.1017/S1368980010000418 PMID: 20236562

42. Hayton J. Commentary on 'Exploring the Sensitivity of Horn's Parallel Analysis to the Distributional Form of Random Data'. 2009:389.

43. Tein J-Y, Coxe S, Cham H. Statistical Power to Detect the Correct Number of Classes in Latent Profile Analysis. Struct Equ Modeling. 2013; 20(4):640–57. https://doi.org/10.1080/10705511.2013.824781 PMID: 24489457

44. Reilly JJ, Penpraze V, Hislop J, Davies G, Grant S, Paton JY. Objective measurement of physical activity and sedentary behaviour: review with new data. Archives of Disease in Childhood. 2008; 93(7):614. https://doi.org/10.1136/adc.2007.133272 PMID: 18305072

45. Lubans DR, Hesketh K, Cliff DP, Barnett LM, Salmon J, Dollman J, et al. A systematic review of the validity and reliability of sedentary behaviour measures used with children and adolescents. Obesity Reviews. 2011; 12(10):781–99. https://doi.org/10.1111/j.1467-789X.2011.00896.x PMID: 21676153