


APPLICATIONS OF LATENT STRUCTURE ANALYSIS TO SAMPLE SURVEYS

Zacharias Panagiotis Tsalavoutas
Tsakiroglou



PANTEION UNIVERSITY OF SOCIAL AND POLITICAL SCIENCES



SCHOOL OF ECONOMIC AND PUBLIC ADMINISTRATION SCIENCES

DEPT. OF ECONOMIC AND REGIONAL DEVELOPMENT

Postgraduate Program of Applied Economics and Management

Applications of Latent Structure Analysis to Sample Surveys

Postgraduate Thesis

Tsalavoutas Tsakiroglou Zacharias Panagiotis

Supervising Professor

Clive Richardson

Athens, Greece, 2022

Tsalavoutas – Tsakiroglou ©2022

Examining Committee:

Richardson Clive, Panteion University Economic and Regional Development Professor
(Supervisor)

Palaskas Theodosios, Panteion University Economic and Regional Development Professor

Degiannakis Stavros, Panteion University Economic and Regional Development Professor

All rights reserved.

No part of this postgraduate thesis may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, for profit. Reproduction, storing in a retrieval system and transmission of any part of this postgraduate thesis is permitted for research and educational purposes, with the sole requirement being to always report the origin of the postgraduate thesis, as well as this message. Potential inquiries about using said thesis for profitable purposes should be directed at the author.

The acceptance of submitting this postgraduate thesis by Panteion University does not directly or otherwise imply the acceptance or embrace of the author's opinions.

Abstract

Statisticians have been using various techniques in order to overcome the single most prominent problem of statistics; dealing with the vast data that can be associated with a group, or a population. The even greater complication of trying to handle entire populations can be summarized as a two-fold problem of the time and money required to firstly collect statistical information about a populace, organize the data, export findings and finally publish the findings. In order to circumvent that, statisticians often use targeted or random groups of people. However, by doing that, other questions are raised; what if the groups are correlated by a currently unknown factor, and ergo are no longer representative of the populace? This is where Latent Structure Analysis (LSA) steps in, which tries to find latent factors which connect such groups and then tries to interpret how the latent factors affect these groups. This research drew samples from HBSC's surveys on adolescent teens in Greece from 2006, 2010 and 2014. Samples of Latent Structure Analyses are implemented, particularly on questions 48 and 65 of the relevant survey questionnaire, separately. The surveys focus on the impact of the economic crisis on the adolescents' lives, with data drawn from times before, during and after the crisis, for comparison's sake, and to observe the times before and after the crisis as reference points. The LSAs were performed using an easily accessible, free to download – with AGPL licencing – software, R Studio, which is a statistical package used in a variety of statistical studies and other applications. The LSAs performed on each question separately, but from the same sample showed the different latent classes that can be extracted from one such analysis that focuses on a single factor. Different factors that are taken into consideration on the time of pursuing a Structure Analysis play a major role in what can be considered a Latent Class as an outcome. The following findings can be used to strengthen one's perception of the Latent Structure Analysis, and provide a clear depiction of which aspects of a sample should be taken under consideration when performing LSAs on statistical samples. Lastly, it should be noted that all relevant tables and outputs are knowingly included in the body of this paper, as it is aimed to be a "how-to" guide to prospective researchers, since at the time of writing, no such guide exists in the bibliography.

Keywords: Sample, latent structure, population, analysis, factor, adolescent.

Περίληψη – Abstract

Οι Στατιστικοί χρησιμοποιούν σωρεία τεχνικών για να υπερπηδήσουν το κυριότερο πρόβλημα της στατιστικής: την επεξεργασία τεραστίων όγκων δεδομένων που μπορεί να έχουν αντληθεί από μια ομάδα ή έναν πληθυσμό. Η ακόμα μεγαλύτερη επιπλοκή της επεξεργασίας δεδομένων ολόκληρων πληθυσμών μπορεί να περιληφθεί σε ένα πρόβλημα δύο συνιστωμένων, του χρήματος και του χρόνου που απαιτούνται, πρωτίστως για την συλλογή των δεδομένων, και έπειτα για την οργάνωση αυτών, την εξαγωγή ευρημάτων και τελικώς την κοινοποίηση αυτών. Για να παρακάμψουν αυτό το σκόπελο, συχνά οι Στατιστικοί κάνουν χρήση τυχαίων δειγμάτων ομάδων ανθρώπων. Όμως, κάνοντας χρήση αυτών των δειγμάτων, εγείρονται άλλα ερωτήματα: τί γίνεται στην περίπτωση που οι δειγματικές ομάδες αυτές συσχετίζονται με κάποιον, προς το παρόν, άγνωστο τρόπο, οπότε και πλέον δεν είναι αντιπροσωπευτικές το πληθυσμού; Σε αυτό το ερώτημα επιχειρεί να απαντήσει η Ανάλυση Λανθανουσών Δομών, ή Latent Structure Analysis (LSA), όπου προσπαθεί να βρεί λανθάνοντες παράγοντες (factors) οι οποίοι συνδέουν τέτοιες ομάδες δείγματος, και στην συνέχεια προσπαθεί να εξηγήσει πώς οι λανθάνοντες παράγοντες επηρεάζουν τις ομάδες αυτές. Η παρούσα έρευνα, άντλησε δείγματα από τα αποτελέσματα των ερωτηματολογίων που διέθεσε ο HBSC σε εφήβους της Ελλάδας τις χρονιές 2006, 2010 και 2014. Εφαρμόστηκαν τεχνικές Ανάλυσης Λανθανουσών Δομών, συγκεκριμένα όσον αφορά τα δεδομένα των ερωτήσεων 48 και 65 του αντίστοιχου ερωτηματολογίου, ξεχωριστά. Οι έρευνες μέσω ερωτηματολογίων επικεντρώνονται στον αντίκτυπο που έχει η οικονομική κρίση στην ζωή των εφήβων, με δεδομένα αντλούμενα σε περιόδους πριν, κατά την διάρκεια και μετά την οικονομική κρίση που χτύπησε την Ελλάδα το 2008, με σκοπό την σύγκριση, και με χρήση των περιόδων πριν και μετά την Κρίση ως σημεία αναφοράς. Τα LSA εφαρμόστηκαν με την χρήση ενός λογισμικού εύκολου στην πρόσβαση, το οποίο διατίθεται δωρεάν – και με άδεια χρήσης AGPL, το επονομαζόμενο “R Studio”, το οποίο είναι μία σουίτα στατιστικής, το οποίο χρησιμοποιείται ευρέως σε διάφορους τομείς στατιστικών μελετών αλλά και άλλων εφαρμογών. Επιπροσθέτως, η εφαρμογή των LSA σε κάθε ερώτηση ξεχωριστά αλλά χρησιμοποιώντας στοιχεία από το ίδιο δείγμα, έδειξε τις διαφορετικές λανθάνουσες κλάσεις (ομάδες) που μπορούν να παρατηρηθούν μέσω μίας τέτοιας ανάλυσης που επικεντρώνεται σε έναν μοναδικό παράγοντα. Επιλέγοντας διαφορετικούς παράγοντες την ώρα της εφαρμογής ενός LSA αλλάζουν το αποτέλεσμα των λανθανουσών κλάσεων. Τα παρακάτω αποτελέσματα της έρευνας αποσκοπούν στην ενίσχυση της αντίληψης της έννοιας της Ανάλυσης Λανθανουσών Δομών, καθώς και αποσκοπούν στην παροχή μιας καθαρής οπτικής για την απόφαση του ερευνητή, ως προς το ποια χαρακτηριστικά και παράγοντες ενός δείγματος θα πρέπει να λάβει υπόψιν κατά την εκτέλεση μιας LSA. Επισημαίνεται, ότι εκουσίως συμπεριλαμβάνονται όλες οι εκτυπώσεις της σουίτας “R Studio”, ούτως ώστε οι ερευνητές να έχουν ένα σημείο αναφοράς για την δική τους έρευνα, καθότι κάτι παρόμοιο εκλείπει από την βιβλιογραφία την στιγμή της συγγραφής.

Λέξεις κλειδιά: Δείγμα, λανθάνουσα δομή, πληθυσμός, ανάλυση, παράγοντας, έφηβοι.

This thesis is dedicated to my mother, who supported me – and put up with my caffeine-fuelled person,

To my professors and friends, who offered me knowledge openhandedly, as well as support throughout my years, academic and not.

To Katia, who's been here for me day and night throughout the difficult times of going through the pandemic while researching my paper. Maybe we are worlds apart, yet always next to each other.

And to my late father, who gave me my most useful tool, trust in my self.

“Trust in yourself and your determination, and you can achieve anything”

Table of Contents

	<i>Introduction</i>	<i>1</i>
<i>1</i>	<i>Literature Review</i>	<i>4</i>
1.1	Historical Reference	4
1.2	Factor Analysis; The obvious connections	5
1.3	Latent Class Analysis; Finding the “unseen”	7
1.4	Examples of implementation and comparison of FA and LCA; or, history is nice and fun, but what are FA and LCA really about?	9
<i>2</i>	<i>Method</i>	<i>12</i>
2.1	Data Collection	12
2.2	Internal & external validity	14
<i>3</i>	<i>Findings.....</i>	<i>63</i>
<i>4</i>	<i>Final discussion</i>	<i>66</i>
<i>5</i>	<i>References.....</i>	<i>68</i>

Introduction

Statisticians, since the early conception of statistical science, have been battling a great problem, that is two-fold; firstly, obtaining the relevant data for their research from a population, and secondly, actually processing and utilizing the said data to export answers on the original research. Starting a statistical research, e.g. the percentage of people who prefer salty snacks, in a population, begins with a single, simple or complex question; how would one get an answer from every person within a population? Upon starting to materialize this research, one is sure to stumble on multiple problems in this early stage of the research, namely the time needed to question each person within a population, the money that has to be spent – since taking such a project upon oneself is impossible, if one wants to complete the research within a reasonable time-frame – as in many cases, people's choices are fickle, and might change from time to time, resulting in inaccurate data, if one is to take too long to gather the answers from a population. Moreover, prospective researchers will face additional problems, such as lack of funding; no individual, company or organization would fund a simple statistical research such as the preference of salty over sweet snacks, if it were to require thousands of dollars, euros etc. and a timeframe of over five years. As such, statisticians have been battling to find methods of extracting data as close as it can get to the actual truth, using smaller groups, rather than the whole population. Naturally, it is exponentially easier, cheaper and faster to gather answers from a few hundred or thousands of people, than it is for a whole country, never mind on bigger scale.

A multitude of methods have been devised through the ages, in order to tackle the problem of choosing the right group or groups, which will be able to accurately represent the population. These methods range from simple random selection, to stratifications of groups, clustering groups, or even using advanced mathematical formulas, in order for the researchers to find suitable groups in which to perform the statistical research. However, since a complete analysis of these methods is beyond the scope of this thesis, useful links can be found in the references, for the reader who wishes to learn more on the details, advantages and disadvantages of the different grouping methods and theories.

After the selection of a grouping method, or even a combination of methods, researchers start collecting the required data, in the form of surveys, questionnaires, etc. by whichever means they have concluded is appropriate, both in terms of efficiency and of course budget and time limitations. When it comes to process such data, though, researchers commence performing statistical analyses, or in layman's terms, "crunching the data". During this stage, researchers have to be inquisitive towards the groups that they have concluded to use, and raise the first question; Do these groups correlate with one another? Is the correlation immediately apparent, or "hidden", the so-called latent correlation? How do they correlate, and how does this affect the data? To answer this question, Lazarsfeld was the first to talk about the "Latent Structure Analysis" in 1968.

Latent Structure Analysis (in literature LSA) is a family of statistical models, with which researchers try to find latent correlations between seemingly uncorrelated groups in statistical research. Early models of LSA were Factor Analysis, and Latent Class Analysis (known in literature as LCA), the first of which focuses on unobservable factors which can affect the data, while the latter focuses on a latent class, the "unseen" correlation between groups, which can affect the data. Later models include hybrids of both Factor Analysis and LCA, as well as more complex models, which are made possible both in theory and calculations through the use of econometric theories, such as time series theorems, but such models span beyond the scope of this introductory paper.

Given the extent of bibliography dedicated to the aforementioned models, there seems to be a shortage of a true practical example on how to implement a form of LSA on a given set of statistical observations (data set). This paper aims to add to the vast bibliography on LSA such a practical example, using existing data sets from HBSC, who surveyed adolescents in 2006, 2010 and 2014, in order to find a link between adolescent aberrant behaviour, such as consumption of alcohol, tobacco and drugs in Greece, before, during and after the Great Economic Depression of 2008. The data presented here can provide a reference guide, of sorts, on how the science of Statistics reached the point that it is nowadays, as well as a "how-to" in implementing Latent Structure Analyses on existing data sets, how can a prospective researcher decide on which model to implement, as well as caveats that one should be wary of when implementing such models. The differentiating factor from other

papers touching similar subjects is the aim to provide to prospective researchers, as well as readers not fully versed in Statistical theories, a robust historical reference on the family of Latent Structure Analyses, while giving examples of real-world data and scenarios, all while putting an effort to explain relevant terms without expecting readers to be familiar with jargon of the statistical field.

1 Literature Review

1.1 Historical Reference

Latent Structure Analysis is a family of statistical models, which can be implemented on data sets, in order for the researcher to either deduce or observe the potential correlation between groups of variables within the data set.

But why is LSA important? Researchers were having difficulty in explaining how groups were affecting each other, or how certain factors could affect the results of a statistical study. Researchers would also face difficulties in measuring, as well as integrating in their study, and finally interpreting the data of intangible factors, such as certain social attributes of people, intelligence, authoritarianism. How would one come to measure authoritarianism within a company's subdivision, or compare the level of authoritarianism between two countries? There is no clear scale of measurement for such variables, and there surely is no definite unit of measuring the depression of a person. So how could one tackle this problem? These were the thoughts that ultimately made their way to become the "latent factors" that always lie inside a population, a sample, or between variables, that inadvertently cause deviations in the data, of what researchers were expecting to receive, and what was the result of an analysis.

As a result of that, researchers firstly took upon themselves the quest to find what the factors could be, through which some variables are correlated, while others appear to be completely unaffected. This was the beginning of what is now known as Factor Analysis (henceforth FA). Factor analysis was firstly minted as a complete theory by Charles Spearman in 1904.

1.2 Factor Analysis; The obvious connections

Factor Analysis, firstly minted in Spearman's paper, although rough around the edges in comparison to modern tools, was the first to capture the idea behind having "unseen forces" which alter the data enough, so as we (the researcher) observe different results from what was expected to be the outcome of a statistical study. Spearman's paper was the first to capture the lynchpin, the central idea of LSA as we know today. Not only did he provide a theoretical background for the Factor Analysis theory, but had also had to come up with his own – at the time – algebraic method, as well as a computational method, in order to support his theory. Never before had one come up with a similar idea, even more so with their own method of extracting the relevant answers from a data set.

However, Spearman's paper was not flawless. He was so focused on only a few latent variables, namely a general term of a person's ability – one person naturally having better dexterity or aptitude at something – or the person's intelligence, that his study was practically narrow sighted, or in a big percentage blinded, by these two latent variables. As a result, his theory could not be fully supported in this primal stage, since as in the progress of the science of Statistics has made apparent, there are many more latent variables than those two. It would have to take nearly thirty years for another contributor to further develop Spearman's theory, and pick up from where he left; L. L. Thurstone. Thurstone, while writing his *Primary Mental Abilities* (Thurstone, 1938), made use of Spearman's "General Intelligence" theory, and pushed the theory further, by making a model coming toe-to-toe with Spearman's model, the "Theory of Primary Mental Abilities", which theorized that mental abilities were not a singular trait, but actually multifaceted. This theory provided Thurstone with the much needed ground to devise the "Multi-Factor Analysis Theory", which, while based on the single latent variable of Spearman, was taking Spearman's theory a step further towards the right direction and supporting that the latent variables were multiple. Thurstone wrote the "Multiple Factor Analysis" theory in 1947, setting the base of modern Multi-Factor Analysis as it is known and used today. Most of the progress on Factor Analysis, as well as the inception of the theory, as it has been apparent so far in the so-far historical reference, was made prominently by psychologists,

bar Hotelling, who was a statistician; Spearman (1904), Thurstone (1938, 1947) and Hotelling (1933), with the latter being mostly considered as the “black sheep” of the theory contributors since his participation is often omitted in literature, as his theory on Principal Component Analysis was of little contribution to the theory, as the theory approaches the same problem with a fundamentally different approach than how FA does. It wasn't until the 1970s that another statistician would incorporate the theory into the science of Statistics. “Factor Analysis as a Statistical Method” written by Lawley and Maxwell in 1971, was the first attempt to solidify FA as a statistical method, which was perceived as a highly controversial theory when it came to statisticians, who traditionally preferred the Principal Component Analysis, over the at-the-time subjective Factor or Multi-Factor Analysis. Lawley and Maxwell were successful in providing solid ground for the theory to be incorporated into the statistical science, although not at their current time. Further development on the Factor Analysis theory would be made, but in much later years, and with less controversy over the subject. In recent years, Factor Analysis has been split into two major categories, namely Confirmatory Factor Analysis (CFA) and Exploratory Factor Analysis (EFA). In Confirmatory Factor Analysis, the researcher hypothesizes the existence of latent variables or latent functions on which the main variables (x_{ij}) are dependent upon. Plainly put, the researcher has deduced either from general research on the subject, previous knowledge, logic, or by simply hypothesizing that the main variables of the statistical analysis are dependent on latent variables or latent functions of variables, and, as the name suggests, performs the CFA in order to confirm their theory, as well as identify the affected variables and finally discover the latent variables or functions, and prove their existence firstly and secondly, their correlation. On the other hand, Exploratory Factor Analysis hypothesises that the researcher has no knowledge of latent variables or latent functions affecting the main variables (x_{ij}), and is implementing EFA on the data at hand, in order to uncover such variables, and draw conclusions on which variables are dependent on latent variables, and discover *possible* latent functions. In layman's terms, EFA provides possible answers, not definitive ones, in most cases, such as solving the equation $x^2 - 4 = 0$, where the *possible* answers are $x_1 = 2$ and $x_2 = -2$, but there is no *definitive* answer on whether x is equal to two or minus two.

Furthermore, EFA cannot only be run once on a given data set, as it firstly requires the researcher to determine how many variables (x_{ij}) are dependent on a factor (f), which in itself has to be run several times in order to deduce the number of variables dependent on a single factor, then run again to deduce the number of variances dependent on two factors, and so on. On each run, the model has to be adjusted for a different number of variables (x_{ij}) that are supposed to be dependent on a latent factor (f), but always bearing the same load, meaning the same force of dependence to the latent factor. If the researcher wants to hypothesize a different load, a different number of variances, or a different number of factors, the model has to be adjusted accordingly, and run again, until the researcher determines that the number of factors, variances and load per variance is correct, through performing a statistical check of hypothesis – setting H_0 the assumption that the answer is correct, and H_1 the alternative, and “check” the validity of the answer, or with the aid of various Goodness of Fit tests (Nylund et al., 2007). While a literature review of Goodness of Fit tests is beyond the scope of this paper, readers keen on this subject will be provided with relevant literature at the end of the paper.

It is apparent that EFA is quite a lot more tedious of a procedure, when compared to CFA, which can be run a couple of times, since the researcher only confirms their “suspicion” on the number of latent factors affecting variables. Furthermore, CFA provides more solid conclusions in comparison to EFA, which, as shown above, can provide *possible* answers through Goodness of Fit tests. Moreover, there is an apparent procedural similarity of EFA with Principal Components Analysis, in the form that the researcher is trying to determine factors (and components) in each analysis, without prior knowledge of their number or even existence. CFA is quite unlike Principal Components Analysis, since it either confirms or rejects a prior assumption of factors.

1.3 Latent Class Analysis; Finding the “unseen”

While many consider Paul Lazarsfeld the “father” of Latent Structures, it would be a great discrepancy on the side of literature to not mention C.S Peirce, Goodman and Kruskal, as well as Henry, in this subject. C.S. Peirce, in his works published in 1884 was the first to introduce the world to what we now refer to as a structure model, which would allow Peirce

to observe and closely monitor a relationship formed between two variables which are only able to take two values, e.g. yes/no, true/false, or numeric values, called dichotomous variables. Peirce's goal was to inspect the relationship of dichotomous variables through this structured model, in order to try and measure the success of predicting the outcome of such variables. While probably missing the point of the importance of latent structures, Lazarsfeld, Goodman and Kruskal were able to perceive it. Goodman and Kruskal, amongst their series of joint papers (Goodman, & Kruskal, 1959) provided a definition of latent classes, as well as an algebraic approach to latent class models. Paul Lazarsfeld and Neil Henry, in their 1968 work "Latent Structure Analysis" were the first to completely capture the idea of the Latent Class Analysis model, the definition of the model, and relevant algebraic solutions and computations (Lazarsfeld & Henry, 1968). Therein lies the reason with which most literature reviews consider Lazarsfeld – and usually mistakenly omitted Henry, the fathers of Latent Structure models and the Latent Class Analysis model. However, Lazarsfeld and Henry focused on highlighting the differences of LCA in comparison to FA, rather than explore the similarities amongst the two models. While providing a robust and comprehensive presentation of the Latent Class Analysis, and how useful the model could be in social sciences, they failed to provide a reliable method with which to obtain parameter estimates. Parameter estimates are what we now know and name coefficients, or plainly put, the rate of change that a variable – latent or otherwise – will impose upon the result. Not showing the similarities of LCA and FA led to the common belief that LCA and FA were two opposing models, and were treated for decades. Not providing a reliable and easy method for parameter estimate obtaining was a major barrier for the method to be used more widely by other researchers, for nearly a decade. Goodman's works in 1974 were the ones to establish a firm position for LCA in the researchers' arsenal, after he provided an easy-to-follow, and implement, method, which obtained maximum likelihood estimates of latent class parameters. This was the deciding factor in LCA gaining a strong foothold amongst the theories and models used by statisticians, and in turn, this made statisticians more confident in exploring the potential of LCA in their studies.

Further exploration of the potential that LCA has as a statistical model was made by A. K Formann, in his papers “Linear logistic latent class analysis”(1982) and “Constrained latent class models: Theory and applications” (1985), S.J. Haberman’s “Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations” (1974) and “Analysis of qualitative data: Volume 2. New developments” (1979) and finally J.A.Hagenaars’ “Categorical causal modeling: Latent class analysis and directed log-linear models with latent variables” (1998). The aforementioned researchers were able to give LCA a much broader field of application, after they managed to integrate LCA in a framework of log-linear models, which made LCA a much more useful model, that can be applied in a broad field of sciences, from behavioural sciences to geographic statistical studies. As this became the case, it is apparent that more and more scientists were open to using LCA as a serious model from the early 1980s onwards, where before this time, it was considered an experimental model not used often in serious work.

1.4 Examples of implementation and comparison of FA and LCA; or, history is nice and fun, but what are FA and LCA really about?

As previously mentioned, Factor Analysis and Latent Class Analysis are nowadays considered closely related brethren within the Latent Structure Analysis family of models. These models are used by researchers who are trying to find connections, either amongst their recorded variables, or between their recorded variables and other “unseen” forces that might be in play.

The Factor Analysis model is most commonly used in junction with surveys, whose applications vary from business and marketing research – e.g. how brand loyal are buyers of jeans, social/behavioural research – e.g. how lonely do people in cities feel versus people living in the countryside, economic research etc. In reality though, it is a model that is applicable in a vast array of sciences and scenarios. FA is a statistical model that can be implemented in any set of observed variables, and tries to describe variability amongst said

variables, in terms of other variables, which are unobserved – and usually ones that cannot be measured, called factors. Usually, the unobserved variables are lower in number than the observed variables. FA tries to find the responses which these unobserved variables impose on the observed variables, with the potential of reducing the work needed to perform a statistical analysis in a smaller number of unobserved (latent) variables, rather than the greater number of observed, since some of the latter will have little to no contribution to the end results (Bandalos, 2017). Plainly put, Factor Analysis tries to find latent variables that are actually significant to the dataset, and makes the researcher's workload lighter in the process, since the former has fewer variables, and consequently fewer values, in which they will implement the statistical analysis that they wanted to. So, in essence, Factor Analysis finds the underlying variables that truly affect the observed variables, which can translate to unmeasured variables in the researcher's dataset that can prove truly important – e.g. the weight that the intelligence of a person bears in a psychosocial research, or the weight that a car with sub-par service can play in a household's brand loyalty for a certain set of products. A crucial step in preparing a Factor Analysis, is to determine the number of factors for which the Factor Analysis will be constructed around. When it comes to the number of factors that the researcher has to set, there are a plethora of theories with their own unique criteria for the former to choose from. Older methods include the Kaiser rule and the Cattell Scree Plot. The Kaiser rule, or Kaiser criterion, presented by Henry Kaiser in 1960, instructed that the researcher could disregard all components of the dataset, whose eigenvalue is less than 1 (Kaiser, 1960). However, in later re-examinations of this criterion, researchers have concluded that the criterion is highly prone to errors, as it has a tendency to present more factors than what they had to be (Bandalos and Boehm-Kaufman, 2008). The Cattell Scree Plot illustrates the components as the X axis, and their eigenvalues as the Y axis. As one observes the progression of the components from the earliest towards the last – moving to the right, eigenvalues have a declining tendency. Cattell supports that once the eigenvalues cease their decline, the plotted curve will look like a shape reminiscent of an elbow. Cattell proposes to discard all components following the one corresponding to the “point” of this “elbow” in the curve (Cattell, 1966). However, a practical rule of thumb amongst statisticians is to avoid this criterion, as in most cases, the point of the elbow is not that apparent, and can lead to subjective results.

Other, more complex criteria are used today, which include Velicer's MAP test (Velicer, 1976) and other variations of it from later researchers (Zwick and Velicer, 1986; Warne and Larsen, 2014; Ruscio and Roche, 2012), as well as Horn's Parallel Analysis (Horn, 1965). However, both of these, as well as their predecessors, have received their share of criticism, as they do appear to have their respective deficiencies. Velicer's MAP test, albeit relatively dependable and robust in execution, relies on a Principal Component Analysis being run before running the Factor Analysis, which does extend the work required to implement the test, but delivers satisfactory results (Garrido et al, 2012, Warne and Larsen, 2014). On the other hand, Horn's Parallel Analysis has been criticized for being easily affected by the type of correlation coefficients, as well as the sample size (Tran & Formann, 2009).

Now let's shift our focus to the other "sibling" in the family of Latent Structures, the Latent Class Analysis model. This model is suitable for use in small data sets, as well as large ones, and is not constrained by the sample size, as was the case with Factor Analysis.

2 Method

2.1 Data Collection

The data used in this paper have been acquired from two sources, the HBSC surveys of WHO, as well as a sample that was processed and firstly used in “Adolescents in Greece in Time of Economic Crisis” (Kokkevi et al., 2017). HBSC is the acronym for Health Behaviour in School-aged Children, a collaborative cross-national study quadrennially, backed up by the World Health Organization. The study includes data that was mined from the participating countries from North America and Europe, and is always conducted in cooperation with the European Regional Office of W.H.O.. The study aims to research the health behaviours, as well as the health in general, of children and young adolescents, and how they are affected by various external factors and stimuli. This study is considered by many as essential for developing new policies, practices and programmes considering the promotion of healthy lifestyle or improving the health of children and young adolescents, or improving the existing ones. The ideology behind this study is to identify the level in which the health of children and young adolescents is currently at, as well as the level of their general well being in relations to their physical situation, as much as their emotional and social well-being, and observe how the external stimuli or forces can impact the health of children and young adolescents. The key word in this study, therefore, is context. The context in which children and young adolescents form their habits, which ultimately affect their health. This context is what we call society and social circles (Currie et al., 2010).

The HBSC survey is administered inside classrooms to each and every student in a probability sample of school classes corresponding to the target age group, in the form of questionnaires. The children and young adolescents then fill and complete the questionnaires, which are then collected. The data from the self completing questionnaires from all the participating countries are then compiled into one international data file. As stated before, the survey takes place every four years, and for each survey a survey method protocol, called the International Protocol, is comprised by study members. Ergo, a quadrennial International Protocol is published, following each survey and its results, for reasons of transparency of the processes which took place when the survey took place, and for posterity and historical reasons as

well. The HBSC survey is comprised by three sets of questions; mandatory, which all participating countries must include, pre-constructed “packages” of questions that each individual country can opt to include in its survey, as well as questions that are specific to the country’s circumstances, and can be considered important for that country’s unique characteristics. As mentioned before, the questions which comprise the questionnaire are styled in such a manner so as to expose the health related behaviours of children and young adolescents, as well as the indicators of the groups’ health. Moreover, the survey unveils the everyday circumstances which affect the health children and young adolescents, such as economic variables, how the health choices of their parents affect them, how their peers affect them, and even how technology affects their health, either in positive or negative manners. The collection and digitalization of these questionnaires happens in each participant country, and the collection of all international data within a single file has many advantages. Namely, results can be compared amongst countries, and conclusions drawn between them, or even study groups of this survey can compare the results of a single country against the median of other countries, etc. It is apparent that such a survey is invaluable to researchers of health in said countries, as well as many other fields of research. Future researchers interested in more details concerning the process of collecting the data of this survey can refer to the HBSC Study Protocol by Currie C. et al, also mentioned in the References section of this paper. The second source of this paper’s data comes from “Adolescents in Greece in Time of Economic Crisis” (Kokkevi et al., 2017), a paper which focuses on children and young adolescents which reside in Greece, and focuses on the effect that the economic crisis of the time affected the age group’s health and health choices. The paper is not a different source of data per se, but more of a focus point on the HBSC’s data, hence actually a second-hand source of sorts. The paper focuses on the data coming from and concerning Greece, from the years 2006, 2010 and 2014, and focusing on children and young adolescents aged 11, 13 and 15 years old. According to the paper, the sample of these ages and years was drawn from stratified probability samples drawn from a group of 3500 to 4200 students, who completed the questionnaire in their classrooms and kept the survey anonymous.

2.2 Internal & external validity

Concerning the validity of the sampling method for the data that will be dealt with in this paper, we again have to turn our focus on both the HBSC and Kokkevi's et al paper. The HBSC study provides the hard facts and numbers as they are processed and digitized from the anonymous questionnaires of children and young adolescents from all participating countries. Furthermore, the HBSC is scrutinized as far as the process of collecting and digitizing the questionnaires from organizations and/or government agencies of each and every participating country. Moreover, the results are checked with the aid of W.H.O., in order for the resulting numbers to be accurate and undeniable. Finally, as mentioned previously, for each iteration of the study, the study group responsible for that years' study issues a Study Protocol, which describes in detail the procedures with which the group acquires the data. This protocol breaks down the full length of the questionnaire; what questions were used, which questions are considered mandatory for that years' study and why, which questions were included in the optional packs and why, and finally a lengthy analysis on the questions which are unique to each country – and which unique factors that are found in that particular country it mitigates (Currie et al.,2010). On the other hand, Kokkevi's et al paper does pick samples, since it focuses on the populace of children and young adolescents of Greece, rather than all the countries participating in the HBSC study. The paper regards the classroom as the primal sampling unit. This means that the unit of measuring a percentage of a populace for one country is considered this one class of students. However, deviations can occur from such a decision, as all school classes are not equal; the number of students per class can vary, as well as the structure of the class – such as the ration of female and male students. Moreover, we have to consider which type of school can be considered as the sample. The paper discusses that Greece has a variety of schools, namely private, public, contemporary and technical schools. This complicates the model of the sample, and the best way to mitigate this complication was to extract a sample from the “populace” of Greece. As reported in the paper, the participants of the study were students who were present on the days in which the surveys took place, and were reported as follows; in 2006 the participants were 3690, of which 47.3% were male, in 2010 the participants were 4899, of which 48.5% were male, and in 2014 there were 4389. However, for this particular year corrections had

to be made, as not all questionnaires were meeting the inclusion standards. The questionnaires which were excluded from the sample were picked according to the common rules for all countries placed by the HBSC for data cleansing. Plainly put, people sometimes fail to complete integral questions on questionnaires, either by absentmindedness, failure to properly read questions or guidelines, or even by accident (Einola & Alvesson, 2021). The actual reasons behind this kind of data error are beyond the scope of this paper, however what is important is to keep in mind that most of the time, in this kind of surveys, there exists a percentage of error – or plainly put, incomplete or wrongly completed surveys. In this particular situation, questionnaires were excluded from the final sample for reasons of failure to complete the gender identifying question, and for completing the majority of questions using the extreme response (Kokkevi et al., 2017). The inclusion of such incomplete questionnaires and ones who were deliberately answered with the extreme responses would degrade any outcome that would be based off of this sample. As such, the abovementioned questionnaires were excluded from the study – what is referred to as data cleansing – so as to keep this sample and the corresponding values as close to the populace as possible. Finally, after this process, the sample of 2014 was comprised of 4141 students, of which 49.8% were male. It is important to note at this point that the entirety of the study and its processes, as described in Kokkevi et al paper, has received approval from the Ministry of Education of Greece, and in particular from the Institute of Educational Policy, concerning the ethics of this study (Kokkevi et al., 2017).

For the sake of simplicity, this paper will only focus on two questions of the study, namely questions No. 48 and No.65. After all, this paper aims at being an easy to comprehend how to guide on Latent Structure Analysis, rather than an extensive study on children's health. The above-mentioned questions refer to two distinct subjects, and asked the following, translated from Greek. The original questions in Greek are available in the Appendix of this paper.

48. In general, in which aspects has the Internet been useful to you (tick all the options which are true for you)

- | | |
|--|---|
| <input type="checkbox"/> School Work | <input type="checkbox"/> Used for informative purposes |
| <input type="checkbox"/> Keeping in touch with friends and relatives | <input type="checkbox"/> Used to create or join new teams or social movements like..... |
| <input type="checkbox"/> Used to fight loneliness | <input type="checkbox"/> Other (Please Specify) |
| <input type="checkbox"/> Used to make new friends in real life | <input type="checkbox"/> The Internet has not been useful to me |
| <input type="checkbox"/> Used to acquire new skills like... | |

65. Have there been any negative effects or trouble in the following aspects of your life, as a result of your behaviour relating to computer games?

- A. Trouble at work, in sports training or school (e.g. bad grades)
- B. Trouble with family/partner or friends (e.g. fights)
- C. Money problems (e.g. debts)
- D. Negligence towards other forms of entertainment
- E. Negligence towards friends/partner
- F. Health problems (e.g. lack of sleep, malnutrition/bad eating habits)

After the questions are set, let's see how the data are formed depending on the answers. First off, the gender of the person answering the questionnaire is marked as **"Q1"**, since it's the first question of the questionnaire. Another relevant question towards the internet is question 20, which asks in what age was the first time the children used the Internet as – **"Q20"**. Question 48 is a question with multiple choices as answers. As such, the multiple choices have to be marked separately. Each choice is a binary variable, which means that can only carry one of two values – 1 for chosen and 0 for not chosen – and is marked as **"Q48a1"** through **"Q48a8"**. The last variable ("The Internet has not been useful to me") was already excluded from Kokkevi et al paper, for simplicity reasons, as it is of little use. Kokkevi et al have included in their data files a few more important variables. **"Age"** is the variable which shows the age of the child answering the questionnaire. **"Ed"** is the variable that signifies the educational level of the child's parents, and finally **"IAT"** is a scale of

measuring the internet addiction of children and the range goes from 1 through 4, with 4 being the worst condition (i.e. “internet addiction”) and 1 being the least bad case. Kokkevi et al note in their notes that the fourth step of the scale, namely the worst category of internet addiction, only includes a small section of the participants and could be joined with step 3 as one single step. It is also important to note that the variables **Q20**, **Ed** and **IAT** have missing values in some cases.

Concerning Question 65, there are a few things to be noted as well. Firstly, this question is related to another, question 53.

53. How often do you play computer games?

- A. Every day
- B. 2 -3 times per week
- C. Once a week
- D. Once a month
- E. Less than once a month
- F. Never

After answering this question, the guidelines commanded the participants to either proceed to the next question, or, if the answer on question 53 was either “Less than once a month” or “Never”, they should skip ahead to question 66. Due to this clause, participants who answer question 65 play computer games at least once a month. The reason behind this is to eliminate answers of infrequent players, whose life is not impacted by computer games, as it is a rare occurrence rather than a habit, hobby or daily/weekly occurrence. The infrequent players would degrade the result of question 65, and as such are refrained from answering the question. As such, the number of participants in questions 48 and 65 are different – $n_{48}=1892$ and $n_{65}=1337$, since there were 555 children who were excluded from answering question 65, as they were identified as infrequent computer game players.

Other variables of question 65 such as **Q1**, **Q20**, **Age**, **Ed**, and **IAT** follow the same principles as stated for question 48.

In this paper, the chosen computer program used to perform the necessary calculations, graphs and analyses is R Studio (Version 1.4.1103, J.J. Allaire), which is actually the program R but with an Integrated Development Environment (IDE). Plainly put, it is the widely known statistical software R, which is accepted and used extensively in statistical studies, with the same, open software tools and underlying code of R, but with the

Tsalavoutas – Tsakiroglou ©2022

additional option of visually representing various commands and functions. It is the same as using R, but easier to work with, as it can be tasked to do anything R does, with buttons and menus instead of having to write code. However, R Studio provides the same tools, same code as R, plus the ability to edit the commands that are input through buttons and menus on the Console Panel, a window which is exactly as the main Console window in R. The decision to use R Studio was a decision made purely for ease of access. What is meant by that is that R Studio is the same software package as R, but laid in a more user friendly way, ergo potential prospective researchers will have an easier time working with R Studio. Rather than having to research and learn how to use a command line interface and the proper syntax required for R, R Studio – in general – provides the same functionality but with graphical user interface – buttons, windows, drop-down menus, while also making first-time users familiar with R's command line system, as R Studio displays a command line console as well. Lastly, this paper provides a few important commands to be input manually in the console, so as to simplify the import of data to R Studio.

This is the point where the Latent Structure Analysis example begins. What is given in the data files Q48 and Q65 are the answers of children and young adolescents, concerning the purposes they have the Internet for in question 48, and the possible problems they may face as an outcome of playing videogames on the computer. At first glance, the questions seem irrelevant with each other. However, the answers come from the same group of children to both questions, mainly, as it was noted that 1337 answers were given to question 65, whereas 1892 answers were given to question 48. Although 555 children and young adolescents have not answered question 65, since it did not apply to them, the rest have participated in both questions. It is apparent that there is some correlation between users of the Internet and participants who have problems in their life due to playing videogames, as both the Internet and computer games require the use of a computer. Keeping in mind that the participants in this study are children and young adolescents, it is easily understood that in these ages children get carried away easily, and can lose their focus from studying, in favour of playing. As such is the usual situation with children and young adolescents, the hypothesis is that there are latent groups amongst the children, which get carried away when using the internet, and instead of focusing on the task at hand, they play computer games, and that behaviour results in various degrees of problematic outcomes, in the children's lives.

Starting off the statistical analysis, two files containing the data of the aforementioned questions 48 and 65 were kindly provided by the researchers of “Adolescents in Greece in Time of Economic Crisis” paper, as mentioned earlier, in the files Q48.sav and Q65.sav . The file type in which they were provided was “.sav”, the file type that is used to save data in SPSS, another statistical software by IBM. The first step to importing the data was to ensure that R Studio can read SPSS data. This can be done using the following command.

```
install.packages("foreign")
```

After entering this command in the console space, R Studio will fetch all relevant code it needs to read data from SPSS save files. Next, the data files were imported through clicking **File → Import Dataset→ From SPSS** and choosing the desired files in the file browsing window that opens. Now the data have been read successfully and are ready to be processed.

As mentioned previously, there are two prominent ways around Latent Structure Analysis, looking for latent factors – Factor Analysis, or latent groups – Latent Class Analysis. In both questions, the data is dichotomous, and can only take a 0/1, or Yes/No, answer, so it is apparent that **Latent Class Analysis** is the appropriate analysis for this type of data.

After choosing the correct analysis, it is important for the prospective researcher to inquire what it is they want to find. What is being examined through the LCA is, in practical terms, the probability that within the students, there are latent groups of students which use the Internet, and have problems with videogame addiction, in various degrees. In order to be easier to refer to the data, the following command is useful.

```
attach(Q48)
```

```
attach(Q65)
```

Using this command points the program to the data sets for question 48 and 65, and there is need no more to point the program towards the datasets every time they are needed. Before getting to the actual LCA though, the data needs to be grouped in a way to be referred to easily, when it is needed to be recalled in the various commands needed for the LCA.

The first 25 entries of the Q48 data file are presented as follows, headers and row numbering are added for ease of use.

Applications of Latent Structure Analysis To Sample Surveys

Table 1: The first 25 entries in the Q48 data file

	Q1	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32
1	3608	0	12	0	1	0	0	0	0	0	0	0	2	0
2	2446	0	12	0	1	0	0	0	0	0	0	0	1	0
3	3873	0	14	1	1	1	0	0	0	0	0	0	1	1
4	3443	1	8	1	1	0	1	0	1	0	0	0	1	1
5	2903	0	13	1	1	0	0	0	0	0	0	0	1	#N UL L!
6	2843	1	10	1	1	0	0	0	1	0	0	0	2	1
7	4257	0	13	0	1	0	0	0	0	0	0	0	2	0
8	4152	0	13	1	1	1	0	0	1	0	0	0	1	1
9	3125	0	13	0	1	0	1	0	0	0	0	0	1	1
10	4452	1	#N UL L!	1	0	0	0	0	0	0	0	0	1	1
11	3609	0	#N UL L!	1	0	0	0	0	0	0	0	0	1	#N UL L!
12	2956	0	13	1	0	0	0	0	1	0	0	0	1	0
13	3554	0	16	1	0	0	0	0	1	0	0	0	2	#N UL L!
14	3113	0	10	1	1	0	0	0	1	0	0	0	1	1
15	3921	0	12	1	1	0	0	0	0	0	0	0	1	1
16	4075	0	10	0	1	0	0	0	0	0	0	0	1	0
17	3930	0	8	1	0	1	0	0	1	0	0	0	1	0
18	4239	0	13	0	1	0	0	0	0	0	0	0	2	0
19	3948	0	14	1	1	0	0	0	0	0	0	0	2	1
20	2594	0	11	1	0	1	1	0	0	0	0	0	2	0
21	4038	0	8	1	0	0	0	0	1	0	0	0	1	1
22	4344	0	13	1	0	0	1	0	1	0	0	0	2	0
23	3524	0	12	1	1	0	0	0	0	0	0	0	1	0
24	3325	0	10	0	1	1	1	0	0	0	0	0	1	1
25	3701	0	13	1	1	0	0	0	0	0	0	0	2	#N UL L!

It is apparent from a quick look on this sample table that Q1 contains the gender data – where 0 are girls and 1 are boys, Q20 the age where the child first used the Internet. It is also apparent that there are missing values in Q20, Ed and IAT. However, these columns contain

data that is irrelevant to the LCA of this paper, and have to be removed from the data set with which the LCA is performed. Instead of deleting the data though, R Studio can instead take data from the original file and group them differently, which is what is used in the following step.

In order to group together the abovementioned columns containing the corresponding data, the following command is input into the console.

```
df48<-data.frame(id_TOTAL,Q48a1,Q48a2,Q48a3,Q48a4,Q48a5,Q48a6,Q48a7)
```

This command makes a data frame, or put more simply, a table containing all the values of the columns Q48a1 to Q48a7 in order, as they are put in the original data file. The reason for doing this is to have all the data relevant to the LCA in one single data frame, enabling easier access at the following commands needed to perform the LCA. A visual representation of the new data frame for Q48 follows, where again the numbering column and headers are added for ease of use, while the column id_TOTAL remains in the data frame, as it will be needed later. The table is saved in memory as df48.

Table 2: Visual Representation of dataframe df48

	id_TO TAL	Q4 8a1	Q4 8a2	Q4 8a3	Q4 8a4	Q4 8a5	Q4 8a6	Q4 8a7	Q4 8a8
1	3608	0	1	0	0	0	0	0	0
2	2446	0	1	0	0	0	0	0	0
3	3873	1	1	1	0	0	0	0	0
4	3443	1	1	0	1	0	1	0	0
5	2903	1	1	0	0	0	0	0	0
6	2843	1	1	0	0	0	1	0	0
7	4257	0	1	0	0	0	0	0	0
8	4152	1	1	1	0	0	1	0	0
9	3125	0	1	0	1	0	0	0	0
1	4452	1	0	0	0	0	0	0	0

0									
1 1	3609	1	0	0	0	0	0	0	0
1 2	2956	1	0	0	0	0	1	0	0
1 3	3554	1	0	0	0	0	1	0	0
1 4	3113	1	1	0	0	0	1	0	0
1 5	3921	1	1	0	0	0	0	0	0
1 6	4075	0	1	0	0	0	0	0	0
1 7	3930	1	0	1	0	0	1	0	0
1 8	4239	0	1	0	0	0	0	0	0
1 9	3948	1	1	0	0	0	0	0	0
2 0	2594	1	0	1	1	0	0	0	0
2 1	4038	1	0	0	0	0	1	0	0
2 2	4344	1	0	0	1	0	1	0	0
2 3	3524	1	1	0	0	0	0	0	0
2	3325	0	1	1	1	0	0	0	0

4									
2	3701	1	1	0	0	0	0	0	0
5									

The same process is followed for Q65, with the following command.

```
df65 <- data.frame(id_TOTAL,Q65a,Q65b,Q65c,Q65d,Q65e,Q65f)
```

As before, a table is created, named **df65**, containing the values for the answer choices of Q65a through Q65f, whereas the first 25 entries of df65 are shown as follows.

Table 3: Visual Representation of the data frame df65

	id_TOTAL	Q65a	Q65b	Q65c	Q65d	Q65e	Q65f
1	3608	0	1	1	0	0	0
2	2446	1	0	0	0	0	0
3	3443	0	0	0	0	0	0
4	2903	0	0	0	0	0	0
5	2843	0	1	0	1	0	1
6	4257	0	0	0	0	0	0
7	4152	0	0	0	0	0	0
8	3125	0	0	0	0	0	0
9	2956	0	0	0	0	0	0
10	3113	0	0	0	0	0	0
11	4075	0	1	0	1	1	0
12	3930	0	0	0	1	0	0
13	4239	1	1	0	0	0	1
14	4344	0	1	0	0	1	0

15	3524	0	0	0	0	1	0
16	3478	1	1	0	0	1	0
17	2490	0	0	0	0	0	0
18	3504	0	0	0	1	0	0
19	3279	0	1	0	1	1	0
20	4376	0	1	1	0	0	0
21	3567	0	0	0	0	0	0
22	3903	0	0	0	0	0	0
23	3441	1	1	1	1	1	1
24	2962	0	0	0	0	0	0
25	3219	0	0	0	0	0	0

The formation of the data frames is not complete yet, though. The basis of this analysis is to define a correlation between students who use the internet and students who have experienced problems from excessively playing video games. However, Q48a8 contains a group of people who have answered that the Internet is of no use to them. This would needlessly degrade the analysis outcome, so the data linked to those students has to be excluded from the analysis. In order for this to be done, the following command has to be input in the console.

```
df48_new<-subset(df48, Q48a8!=1)
```

The **subset** command creates a new table, which has the data of the old table, but has removed the corresponding rows (unique student) to students who have answered “I find no use for the Internet”. The new data frame for question 48, now called **df48_new** has now removed the rows of the abovementioned students, along with their unique ID number in **id_TOTAL**.

Both data sets are now ready to be processed. However, as mentioned previously, **n₄₈=1892** and **n₆₅=1337**, since there were 555 students who were deterred from answering question 65 by their choice in question 53 (“Less than once a month” in the question on how often the

child played computer games). So what this means is that there will be around 555 rows with data for question 48, in which there will be no data available for question 65. Furthermore, the rows could be less than 555, since in the previous step the students who didn't find the Internet useful were purposefully excluded. This unknown set of rows will have to be removed from the table, but will be dealt with further down the steps of the LCA. At present, the information available suggests that the two data frames are not of the same size. The merging of these tables though can be done by comparing the **id_TOTAL** for each row, and merge the IDs which are present in both tables. This ensures that students who answered both questions, and did not answer "I find no use for the Internet" are present in one data frame, with their data aligned to each student. In order to perform this step, the following command is input in the console.

```
df4865<-merge(df48_new, df65, by = "id_TOTAL")
```

The **merge** command merges the two data sets, but not blindly. Instead, the condition **by** = is used, which tells the program to merge the two data frames by searching the column **id_TOTAL** in both data frames – which is the reason this column was left in both **q48_new** and **q65**. Simply put, this condition searches the **id_TOTAL** column of **q48** and **q65**, finds the corresponding values to both sets and joins only the rows which correspond to the values of **id_TOTAL** in both data frames. The resulting data frame **df4865** contains the data of the responses of students who

- have *some* kind of use for the Internet,
- *while* they also play computer games, and
- *might* be facing problems in various aspects of their life due to computer games.

The above three sentences are the criteria for participating in this LCA, and the data frame **q4865** is the data frame that includes all the students who meet these criteria. The first 25 entries of the visual representation of this data frame is shown below, with the addition of headers and row numbering. Now all of the relevant data is included in one single place, easily accessible and also easy to recall in the following steps.

Table 4: Visual Representation of the data frame df4865

Applications of Latent Structure Analysis To Sample Surveys

	id_TOTAL	Q48a1	Q48a2	Q48a3	Q48a4	Q48a5	Q48a6	Q48a7	Q48a8	Q65a	Q65b	Q65c	Q65d	Q65e	Q65f
1	3608	0	1	0	0	0	0	0	0	0	1	1	0	0	0
2	2446	0	1	0	0	0	0	0	0	1	0	0	0	0	0
3	3873	1	1	1	0	0	0	0	0	0	0	0	0	0	0
4	3443	1	1	0	1	0	1	0	0	0	0	0	0	0	0
5	2903	1	1	0	0	0	0	0	0	0	1	0	1	0	1
6	2843	1	1	0	0	0	1	0	0	0	0	0	0	0	0
7	4257	0	1	0	0	0	0	0	0	0	0	0	0	0	0
8	4152	1	1	1	0	0	1	0	0	0	0	0	0	0	0
9	3125	0	1	0	1	0	0	0	0	0	0	0	0	0	0
10	4452	1	0	0	0	0	0	0	0	0	0	0	0	0	0
11	3609	1	0	0	0	0	0	0	0	0	1	0	1	1	0
12	2956	1	0	0	0	0	1	0	0	0	0	0	1	0	0
13	3554	1	0	0	0	0	1	0	0	1	1	0	0	0	1
14	3113	1	1	0	0	0	1	0	0	0	1	0	0	1	0
15	3921	1	1	0	0	0	0	0	0	0	0	0	0	1	0
16	4075	0	1	0	0	0	0	0	0	1	1	0	0	1	0
17	3930	1	0	1	0	0	1	0	0	0	0	0	0	0	0
18	4239	0	1	0	0	0	0	0	0	0	0	0	1	0	0
19	3948	1	1	0	0	0	0	0	0	0	1	0	1	1	0
20	2594	1	0	1	1	0	0	0	0	0	1	1	0	0	0
21	4038	1	0	0	0	0	1	0	0	0	0	0	0	0	0
22	4344	1	0	0	1	0	1	0	0	0	0	0	0	0	0
23	3524	1	1	0	0	0	0	0	0	1	1	1	1	1	1
24	3325	0	1	1	1	0	0	0	0	0	0	0	0	0	0
25	3701	1	1	0	0	0	0	0	0	0	0	0	0	0	0

The Latent Class Analysis, as well as several other analyses akin to LCA, such as the Latent Profile Analysis, can be achieved with multiple sets of commands. These sets of commands are combined in pre-coded “packages” in R, and as an extension of that, in R Studio. The most commonly used package is **poLCA**, an abbreviation for Polytomous variable Latent Class Analysis (Linzer & Lewis, 2011). This package of code includes the calculations required to perform a Latent Class Analysis in, as the name suggests, polytomous variables. Since the data frame is comprised of polytomous data – as multiple choices in a question are by nature polytomous variables – it is the most suited package for use in this analysis. However, as perfect as it is for this use scenario, poLCA needs the variables to be

set as 1 for No, 2 for Yes etc. While it accepts binary data, or more easily comprehensible Yes/No answers, it cannot comprehend data in the usual binary format, that is, zero for “No” and 1 for “Yes”. It is a limitation of the package, that is also present in other packages as well, while other software packages such as SPSS do not have such unusual limitations. Other programs aside, the data frame that was created in the last stage, **df4865**, is comprised only by zeroes and ones, as it was exported from – and probably used in – SPSS. Ergo, the data frame has to be transformed once again, in order to be accepted by the code package. Much akin to setting a variable as a function containing other variables, is the process that is followed in this case also. Instead of trying to recode the data and changing “No” to 1 and “Yes” to 2, it is much more efficient instead to alter the whole frame at once. The following command utilizes this principle.

```
df4865<-df4865+1
```

This command creates a new data frame, with the values of the old one, but increased by 1. Effectively, it eliminates the 0 values which were the numerical representation of “No” and makes the “No” representation as 1, and the “Yes” representation as 2. The first 25 rows of this data frame are represented in the table below.

Table 5: Visual Representation of the data frame df4865 (new)

	id_TOTa L	Q48a1	Q48a2	Q48a3	Q48a4	Q48a5	Q48a6	Q48a7	Q48a8	Q65a	Q65b	Q65c	Q65d	Q65e	Q65f
1	360	1	2	1	1	1	1	1	1	1	2	2	1	1	1
	8														
2	244	1	2	1	1	1	1	1	1	2	1	1	1	1	1
	6														
3	387	2	2	2	1	1	1	1	1	1	1	1	1	1	1
	3														
4	344	2	2	1	2	1	2	1	1	1	1	1	1	1	1
	3														
5	290	2	2	1	1	1	1	1	1	1	2	1	2	1	2
	3														
6	284	2	2	1	1	1	2	1	1	1	1	1	1	1	1
	3														
7	425	1	2	1	1	1	1	1	1	1	1	1	1	1	1
	7														
8	415	2	2	2	1	1	2	1	1	1	1	1	1	1	1
	2														
9	312	1	2	1	2	1	1	1	1	1	1	1	1	1	1
	5														
10	445	2	1	1	1	1	1	1	1	1	1	1	1	1	1
	2														
11	360	2	1	1	1	1	1	1	1	1	2	1	2	2	1
	9														
12	295	2	1	1	1	1	2	1	1	1	1	1	2	1	1
	6														
13	355	2	1	1	1	1	2	1	1	2	2	1	1	1	2
	4														
14	311	2	2	1	1	1	2	1	1	1	2	1	1	2	1
	3														
15	392	2	2	1	1	1	1	1	1	1	1	1	1	2	1
	1														
16	407	1	2	1	1	1	1	1	1	2	2	1	1	2	1
	5														
17	393	2	1	2	1	1	2	1	1	1	1	1	1	1	1
	0														
18	423	1	2	1	1	1	1	1	1	1	1	1	2	1	1
	9														
19	394	2	2	1	1	1	1	1	1	1	2	1	2	2	1
	8														

Applications of Latent Structure Analysis To Sample Surveys

2 0	259 4	2	1	2	2	1	1	1	1	1	2	2	1	1	1
2 1	403 8	2	1	1	1	1	2	1	1	1	1	1	1	1	1
2 2	434 4	2	1	1	2	1	2	1	1	1	1	1	1	1	1
2 3	352 4	2	2	1	1	1	1	1	1	2	2	2	2	2	2
2 4	332 5	1	2	2	2	1	1	1	1	1	1	1	1	1	1
2 5	370 1	2	2	1	1	1	1	1	1	1	1	1	1	1	1

Now the data is in a format which the code package can understand and process without trouble. In order to use the poLCA package, the following commands have to be input into the console. The first one is as follows.

```
f1 <- as.formula(cbind(Q48a1, Q48a2, Q48a3, Q48a4, Q48a5, Q48a6, Q48a7, Q65a, Q65b, Q65c, Q65d, Q65e, Q65f)~1)
```

This command is a facilitating command, which creates a formula, **f1**. This formula selects the columns Q48a1 to Q65f from the data frame. This command is the equivalent of showing the program where it should pick the data from for the analysis. Afterwards, the actual package of code containing the command for starting the analysis, poLCA, has to be installed and called from the “library” of the program. The complete commands required to install and recall poLCA are input to the console, as follows.

```
install.packages("LCA")
```

```
library(poLCA)
```

Now the code for running an LCA analysis is loaded, and can be activated using the following command.

```
LCA2 <- poLCA(f1, data = df4865, nclass = 2, graphs = TRUE, na.rm = TRUE)
```

The command starts a latent class analysis on the previously selected columns in the formula, which data it draws from the data frame that was constructed last, the **df4865** where the zeroes were replaced by 1 and the ones were replaced with 2. The command also instructs the code to perform an LCA for a number of 2 hypothetical classes and to draw the relevant bar plot. Lastly, the **na.rm** function tells the program to remove any rows which

might have cells which contain NA (Not Available) values, and is used as a precaution to expunge any data that might be incomplete and affect the resulting outcome.

The complete printout of this procedure in the console of R Studio can be seen beneath. Note that items in blue bold lettering follow the colour palette of R and R Studio console input, when the user inputs commands – also apparent from the “greater than” symbol, which indicates the start of a new line, waiting for input of a new command.

```
> library(haven)
> Q48 <- read_sav("~/Desktop/Applications of Latent Structure Analysis to Sample
Surveys/DATA/Q48.sav")
> View(Q48).
> attach(Q48)
> df48<-data.frame(id_TOTAL,Q48a1,Q48a2,Q48a3,Q48a4,Q48a5,Q48a6,Q48a7)
> library(haven)
> Q65 <- read_sav("~/Desktop/Applications of Latent Structure Analysis to Sample
Surveys/DATA/Q65.sav")
> View(Q65)
> attach(Q65)
> df65 <- data.frame(id_TOTAL,Q65a,Q65b,Q65c,Q65d,Q65e,Q65f)
> df48_new<-subset(df48, Q48a8!=1)
> df4865<-merge(df48_new, df65, by = "id_TOTAL")
> View(df4865)
> df4865<-df4865+1
> fl <- as.formula(cbind(Q48a1, Q48a2, Q48a3, Q48a4, Q48a5, Q48a6, Q48a7, Q65a,
Q65b, Q65c, Q65d, Q65e, Q65f)~1)
> LCA2 <- poLCA(fl, data = df4865, nclass = 2, graphs = TRUE, na.rm = TRUE)
```

Conditional item response (column) probabilities,
by outcome variable, for each class (row)

\$Q48a1

Pr(1) Pr(2)

class 1: 0.2390 0.7610

class 2: 0.4363 0.5637

\$Q48a2

Pr(1) Pr(2)

class 1: 0.2602 0.7398

class 2: 0.3132 0.6868

\$Q48a3

Pr(1) Pr(2)

class 1: 0.8918 0.1082

class 2: 0.6693 0.3307

\$Q48a4

Pr(1) Pr(2)

class 1: 0.8159 0.1841

class 2: 0.7352 0.2648

\$Q48a5

Pr(1) Pr(2)

class 1: 0.8931 0.1069

class 2: 0.8696 0.1304

\$Q48a6

Pr(1) Pr(2)

class 1: 0.5110 0.4890

class 2: 0.5978 0.4022

\$Q48a7

Pr(1) Pr(2)

class 1: 0.9547 0.0453

class 2: 0.8905 0.1095

\$Q65a

Pr(1) Pr(2)

class 1: 0.9619 0.0381

class 2: 0.4965 0.5035

\$Q65b

Pr(1) Pr(2)

class 1: 0.941 0.059

class 2: 0.426 0.574

\$Q65c

Pr(1) Pr(2)

class 1: 0.9781 0.0219

class 2: 0.7442 0.2558

\$Q65d

Pr(1) Pr(2)

class 1: 0.9080 0.0920

class 2: 0.4792 0.5208

\$Q65e

Pr(1) Pr(2)

class 1: 0.9840 0.0160

class 2: 0.6293 0.3707

\$Q65f

Pr(1) Pr(2)

class 1: 0.9627 0.0373

class 2: 0.5937 0.4063

Estimated class population shares

0.6916 0.3084

Predicted class memberships (by modal posterior prob.)

0.7058 0.2942

=====

Fit for 2 latent classes:

=====

number of observations: 1244

number of estimated parameters: 27

residual degrees of freedom: 1217

maximum log-likelihood: -7186.075

AIC(2): 14426.15

BIC(2): 14564.55

G²(2): 2172.553 (Likelihood ratio/deviance statistic)

X²(2): 11724.09 (Chi-square goodness of fit)

*****End of output*****

The plot created for the LCA with a number of classes set to 2 (**nclass = 2**) is the following.

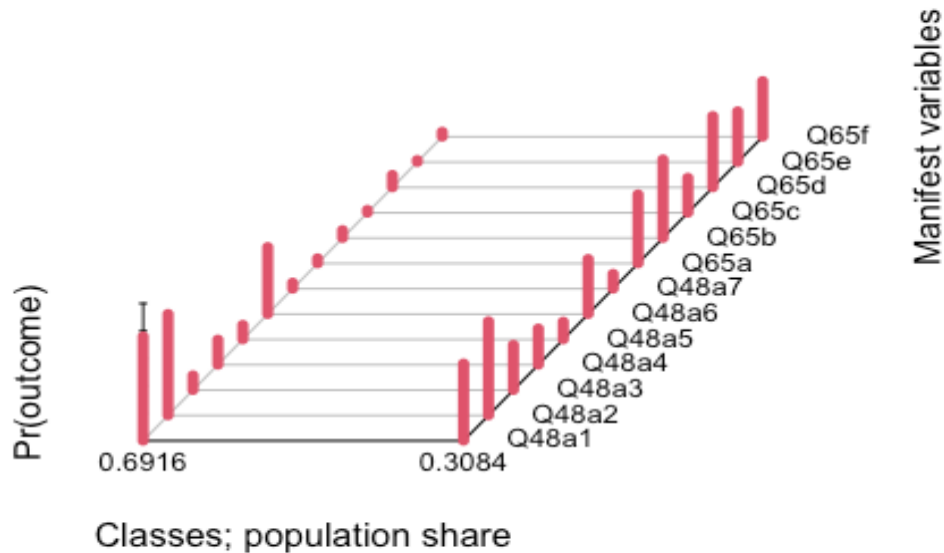


Figure 1 LCA for *nclass = 2*

Continuing the analysis, the `poLCA` command has to be run again for more hypothetical classes. Starting with `nclass = 3`, the command has to be edited each time it is run, as follows.

```
LCA3 <- poLCA(f1, data = df4865, nclass = 3, graphs = TRUE, na.rm = TRUE)
```

The same has to be run for `nclass = 4`, `nclass = 5`, `nclass = 6` etc, for example

```
LCA4 <- poLCA(f1, data = df4865, nclass = 4, graphs = TRUE, na.rm = TRUE)
```

```
LCA5 <- poLCA(f1, data = df4865, nclass = 5, graphs = TRUE, na.rm = TRUE)
```

```
LCA6 <- poLCA(f1, data = df4865, nclass = 6, graphs = TRUE, na.rm = TRUE)
```

While this command has to be run each time for each consecutive change in class number, the previous steps of organizing the data frames, they are already ready for all repeats of `poLCA`.

The printouts of the console of R Studio for each consecutive run of poLCA are as follows.

```
> LCA3 <- poLCA(f1, data = df4865, nclass = 3, graphs = TRUE, na.rm = TRUE)
```

Conditional item response (column) probabilities,
by outcome variable, for each class (row)

\$Q48a1

Pr(1) Pr(2)

class 1: 0.0498 0.9502

class 2: 0.2656 0.7344

class 3: 0.5523 0.4477

\$Q48a2

Pr(1) Pr(2)

class 1: 0.0168 0.9832

class 2: 0.2961 0.7039

class 3: 0.4022 0.5978

\$Q48a3

Pr(1) Pr(2)

class 1: 0.6292 0.3708

class 2: 0.9173 0.0827

class 3: 0.7202 0.2798

\$Q48a4

Pr(1) Pr(2)

class 1: 0.5936 0.4064

class 2: 0.8530 0.1470

class 3: 0.7700 0.2300

\$Q48a5

Pr(1) Pr(2)

class 1: 0.7334 0.2666

class 2: 0.9189 0.0811

class 3: 0.9063 0.0937

\$Q48a6

Pr(1) Pr(2)

class 1: 0.1198 0.8802

class 2: 0.5685 0.4315

class 3: 0.7420 0.2580

\$Q48a7

Pr(1) Pr(2)

class 1: 0.7684 0.2316

class 2: 0.9783 0.0217

class 3: 0.9392 0.0608

\$Q65a

Pr(1) Pr(2)

class 1: 0.7472 0.2528

class 2: 0.9762 0.0238

class 3: 0.4754 0.5246

\$Q65b

Pr(1) Pr(2)

class 1: 0.6942 0.3058

class 2: 0.9541 0.0459

class 3: 0.4157 0.5843

\$Q65c

Pr(1) Pr(2)

class 1: 0.9875 0.0125

class 2: 0.9764 0.0236

class 3: 0.6769 0.3231

\$Q65d

Pr(1) Pr(2)

class 1: 0.6343 0.3657

class 2: 0.9295 0.0705

class 3: 0.4900 0.5100

\$Q65e

Pr(1) Pr(2)

class 1: 0.8914 0.1086

class 2: 0.9862 0.0138

class 3: 0.5871 0.4129

\$Q65f

Pr(1) Pr(2)

class 1: 0.7888 0.2112

class 2: 0.9762 0.0238

class 3: 0.5741 0.4259

Estimated class population shares

0.1617 0.597 0.2413

Predicted class memberships (by modal posterior prob.)

0.1399 0.635 0.2251

=====
Fit for 3 latent classes:

=====
number of observations: 1244

number of estimated parameters: 41

residual degrees of freedom: 1203

maximum log-likelihood: -7089.595

AIC(3): 14261.19

BIC(3): 14471.36

$G^2(3)$: 1979.593 (Likelihood ratio/deviance statistic)

$X^2(3)$: 10181.12 (Chi-square goodness of fit)

*****End of output*****

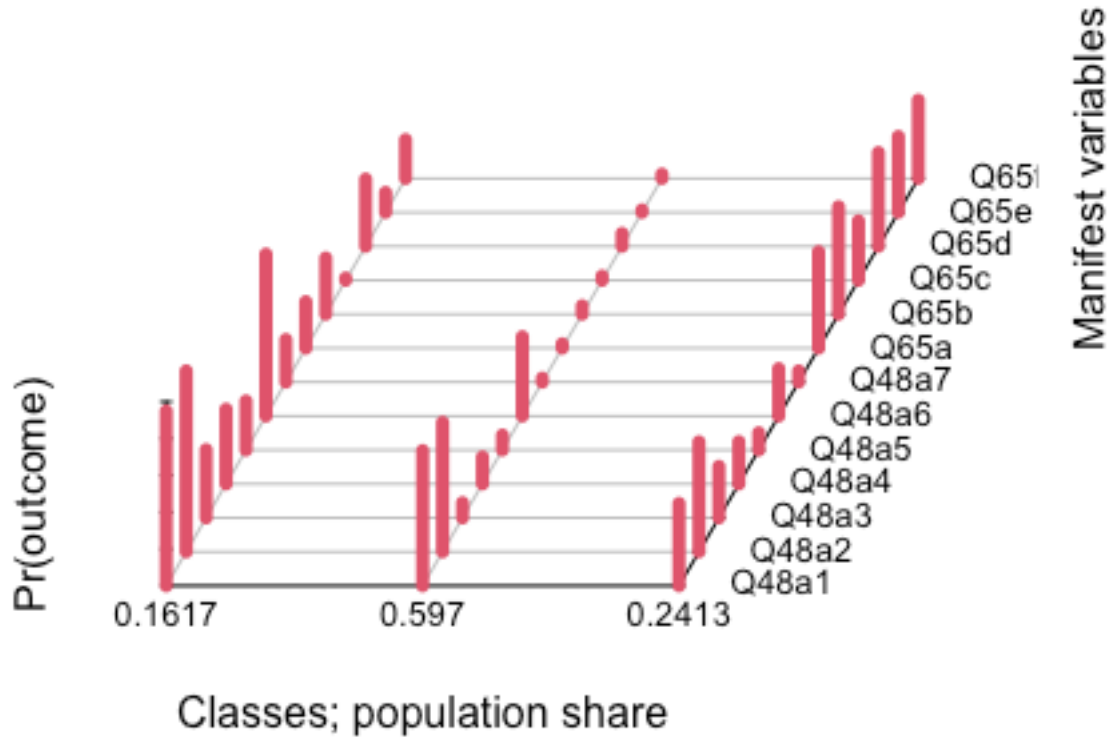


Figure 2 LCA for $n_{class} = 3$

```
> LCA4 <- poLCA(f1, data = df4865, nclass = 4, graphs = TRUE, na.rm = TRUE)
```

Conditional item response (column) probabilities,

by outcome variable, for each class (row)

```
$Q48a1
```

```
Pr(1) Pr(2)
```

```
class 1: 0.6412 0.3588
```

```
class 2: 0.2615 0.7385
```

```
class 3: 0.1329 0.8671
```

class 4: 0.0000 1.0000

\$Q48a2

Pr(1) Pr(2)

class 1: 0.4118 0.5882

class 2: 0.2844 0.7156

class 3: 0.0000 1.0000

class 4: 0.1176 0.8824

\$Q48a3

Pr(1) Pr(2)

class 1: 0.7288 0.2712

class 2: 0.9128 0.0872

class 3: 0.4331 0.5669

class 4: 0.6806 0.3194

\$Q48a4

Pr(1) Pr(2)

class 1: 0.7642 0.2358

class 2: 0.8416 0.1584

class 3: 0.2799 0.7201

class 4: 0.7603 0.2397

\$Q48a5

Pr(1) Pr(2)

class 1: 0.9131 0.0869

class 2: 0.9120 0.0880

class 3: 0.4648 0.5352

class 4: 0.8556 0.1444

\$Q48a6

Pr(1) Pr(2)

class 1: 0.7809 0.2191

class 2: 0.5526 0.4474

class 3: 0.1158 0.8842

class 4: 0.2258 0.7742

\$Q48a7

Pr(1) Pr(2)

class 1: 0.9351 0.0649

class 2: 0.9766 0.0234

class 3: 0.2869 0.7131

class 4: 0.9462 0.0538

\$Q65a

Pr(1) Pr(2)

class 1: 0.4395 0.5605

class 2: 0.9721 0.0279

class 3: 0.6064 0.3936

class 4: 0.7659 0.2341

\$Q65b

Pr(1) Pr(2)

class 1: 0.4304 0.5696

class 2: 0.9541 0.0459

class 3: 0.7782 0.2218

class 4: 0.5395 0.4605

\$Q65c

Pr(1) Pr(2)

class 1: 0.6265 0.3735

class 2: 0.9760 0.0240

class 3: 0.9760 0.0240

class 4: 0.9952 0.0048

\$Q65d

Pr(1) Pr(2)

class 1: 0.5071 0.4929

class 2: 0.9325 0.0675

class 3: 0.7442 0.2558

class 4: 0.4828 0.5172

\$Q65e

Pr(1) Pr(2)

class 1: 0.5934 0.4066

class 2: 0.9888 0.0112

class 3: 0.9572 0.0428

class 4: 0.7604 0.2396

\$Q65f

Pr(1) Pr(2)

class 1: 0.5686 0.4314

class 2: 0.9758 0.0242

class 3: 0.8940 0.1060

class 4: 0.6868 0.3132

Estimated class population shares

Applications of Latent Structure Analysis To Sample Surveys

0.2078 0.6157 0.042 0.1344

Predicted class memberships (by modal posterior prob.)

0.1897 0.6519 0.0362 0.1222

=====
Fit for 4 latent classes:

=====
number of observations: 1244

number of estimated parameters: 55

residual degrees of freedom: 1189

maximum log-likelihood: -7059.271

AIC(4): 14228.54

BIC(4): 14510.48

$G^2(4)$: 1918.945 (Likelihood ratio/deviance statistic)

$X^2(4)$: 9477.614 (Chi-square goodness of fit)

ALERT: iterations finished, MAXIMUM LIKELIHOOD NOT FOUND



Figure 3 LCA for $n_{\text{class}} = 4$

The last repetition of the LCA returns an alert, reading that the maximum likelihood could not be found. This means that having 4 classes in which the students might be grouped is not a good fit for this set of data. Repeating the analysis for $n_{\text{class}} = 5$ prints out the following.

```
> LCA5 <- poLCA(f1, data = df4865, nclass = 5, graphs = TRUE, na.rm = TRUE)
```

Conditional item response (column) probabilities,

by outcome variable, for each class (row)

```
$Q48a1
```

```
Pr(1) Pr(2)
```

```
class 1: 0.3293 0.6707
```

class 2: 0.2534 0.7466

class 3: 0.0000 1.0000

class 4: 0.6834 0.3166

class 5: 0.1292 0.8708

\$Q48a2

Pr(1) Pr(2)

class 1: 0.2651 0.7349

class 2: 0.2835 0.7165

class 3: 0.1364 0.8636

class 4: 0.4236 0.5764

class 5: 0.0000 1.0000

\$Q48a3

Pr(1) Pr(2)

class 1: 0.6558 0.3442

class 2: 0.9194 0.0806

class 3: 0.7129 0.2871

class 4: 0.7489 0.2511

class 5: 0.4381 0.5619

\$Q48a4

Pr(1) Pr(2)

class 1: 0.5684 0.4316

class 2: 0.8444 0.1556

class 3: 0.7860 0.2140

class 4: 0.7869 0.2131

class 5: 0.2819 0.7181

\$Q48a5

Pr(1) Pr(2)

class 1: 0.8947 0.1053

class 2: 0.9156 0.0844

class 3: 0.8583 0.1417

class 4: 0.9113 0.0887

class 5: 0.4846 0.5154

\$Q48a6

Pr(1) Pr(2)

class 1: 0.6717 0.3283

class 2: 0.5510 0.4490

class 3: 0.2510 0.7490

class 4: 0.7925 0.2075

class 5: 0.1120 0.8880

\$Q48a7

Pr(1) Pr(2)

class 1: 0.9547 0.0453

class 2: 0.9773 0.0227

class 3: 0.9522 0.0478

class 4: 0.9368 0.0632

class 5: 0.3213 0.6787

\$Q65a

Pr(1) Pr(2)

class 1: 0.0000 1.0000

class 2: 0.9814 0.0186

class 3: 0.8027 0.1973

class 4: 0.5355 0.4645

class 5: 0.5992 0.4008

\$Q65b

Pr(1) Pr(2)

class 1: 0.0000 1.0000

class 2: 0.9689 0.0311

class 3: 0.5847 0.4153

class 4: 0.5217 0.4783

class 5: 0.7761 0.2239

\$Q65c

Pr(1) Pr(2)

class 1: 0.3605 0.6395

class 2: 0.9812 0.0188

class 3: 0.9936 0.0064

class 4: 0.6885 0.3115

class 5: 0.9726 0.0274

\$Q65d

Pr(1) Pr(2)

class 1: 0.0000 1.0000

class 2: 0.9429 0.0571

class 3: 0.5297 0.4703

class 4: 0.6122 0.3878

class 5: 0.7312 0.2688

\$Q65e

Pr(1) Pr(2)

class 1: 0.0000 1.0000

class 2: 0.9936 0.0064

class 3: 0.7900 0.2100

class 4: 0.7043 0.2957

class 5: 0.9597 0.0403

\$Q65f

Pr(1) Pr(2)

class 1: 0.2525 0.7475

class 2: 0.9851 0.0149

class 3: 0.7113 0.2887

class 4: 0.6486 0.3514

class 5: 0.8850 0.1150

Estimated class population shares

0.0253 0.5643 0.1567 0.2088 0.0449

Predicted class memberships (by modal posterior prob.)

0.0281 0.5924 0.1559 0.1857 0.0378

=====
Fit for 5 latent classes:

=====
number of observations: 1244

number of estimated parameters: 69

residual degrees of freedom: 1175

maximum log-likelihood: -7020.336

AIC(5): 14178.67

BIC(5): 14532.37

$G^2(5)$: 1841.074 (Likelihood ratio/deviance statistic)

$X^2(5)$: 8405.01 (Chi-square goodness of fit)

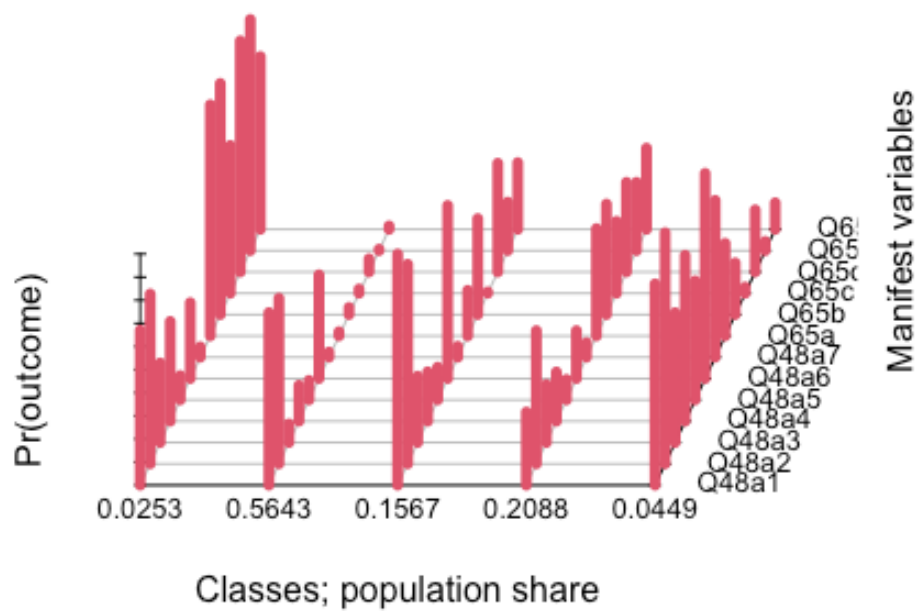


Figure 4 LCA for $n_{class} = 5$

For nclass = 6

```
> LCA6 <- poLCA(f1, data = df4865, nclass = 6, graphs = TRUE, na.rm = TRUE)
```

Conditional item response (column) probabilities,
by outcome variable, for each class (row)

\$Q48a1

Pr(1) Pr(2)

class 1: 1.0000 0.0000

class 2: 0.3277 0.6723

class 3: 0.0400 0.9600

class 4: 0.1657 0.8343

class 5: 0.1313 0.8687

class 6: 0.6965 0.3035

\$Q48a2

Pr(1) Pr(2)

class 1: 0.0000 1.0000

class 2: 0.2602 0.7398

class 3: 0.1478 0.8522

class 4: 0.3067 0.6933

class 5: 0.0000 1.0000

class 6: 0.4686 0.5314

\$Q48a3

Pr(1) Pr(2)

class 1: 1.0000 0.0000

class 2: 0.6598 0.3402

class 3: 0.7271 0.2729

class 4: 0.9066 0.0934

class 5: 0.4390 0.5610

class 6: 0.7198 0.2802

\$Q48a4

Pr(1) Pr(2)

class 1: 0.9631 0.0369

class 2: 0.5684 0.4316

class 3: 0.8131 0.1869

class 4: 0.8309 0.1691

class 5: 0.2843 0.7157

class 6: 0.7616 0.2384

\$Q48a5

Pr(1) Pr(2)

class 1: 0.9747 0.0253

class 2: 0.8960 0.1040

class 3: 0.8517 0.1483

class 4: 0.9080 0.0920

class 5: 0.5208 0.4792

class 6: 0.9145 0.0855

\$Q48a6

Pr(1) Pr(2)

class 1: 1.0000 0.0000

class 2: 0.6739 0.3261

class 3: 0.3211 0.6789

class 4: 0.4903 0.5097

class 5: 0.1067 0.8933

class 6: 0.7869 0.2131

\$Q48a7

Pr(1) Pr(2)

class 1: 0.9866 0.0134

class 2: 0.9574 0.0426

class 3: 0.9474 0.0526

class 4: 0.9767 0.0233

class 5: 0.3747 0.6253

class 6: 0.9322 0.0678

\$Q65a

Pr(1) Pr(2)

class 1: 0.9494 0.0506

class 2: 0.0000 1.0000

class 3: 0.7831 0.2169

class 4: 0.9777 0.0223

class 5: 0.6016 0.3984

class 6: 0.5021 0.4979

\$Q65b

Pr(1) Pr(2)

class 1: 0.9014 0.0986

class 2: 0.0000 1.0000

class 3: 0.4956 0.5044

class 4: 0.9723 0.0277

class 5: 0.7803 0.2197

class 6: 0.5218 0.4782

\$Q65c

Pr(1) Pr(2)

class 1: 0.9270 0.0730

class 2: 0.3460 0.6540

class 3: 0.9962 0.0038

class 4: 0.9825 0.0175

class 5: 0.9709 0.0291

class 6: 0.6584 0.3416

\$Q65d

Pr(1) Pr(2)

class 1: 0.9366 0.0634

class 2: 0.0000 1.0000

class 3: 0.5034 0.4966

class 4: 0.9329 0.0671

class 5: 0.7188 0.2812

class 6: 0.5996 0.4004

\$Q65e

Pr(1) Pr(2)

class 1: 0.9593 0.0407

Applications of Latent Structure Analysis To Sample Surveys

class 2: 0.0000 1.0000

class 3: 0.7447 0.2553

class 4: 0.9935 0.0065

class 5: 0.9605 0.0395

class 6: 0.6962 0.3038

\$Q65f

Pr(1) Pr(2)

class 1: 1.0000 0.0000

class 2: 0.2322 0.7678

class 3: 0.6956 0.3044

class 4: 0.9757 0.0243

class 5: 0.8728 0.1272

class 6: 0.6249 0.3751

Estimated class population shares

0.0648 0.0244 0.1481 0.532 0.0491 0.1817

Predicted class memberships (by modal posterior prob.)

0.0828 0.0265 0.1262 0.5603 0.0426 0.1616

=====
Fit for 6 latent classes:

=====
number of observations: 1244

number of estimated parameters: 83

residual degrees of freedom: 1161

maximum log-likelihood: -6991.891

AIC(6): 14149.78

BIC(6): 14575.25

$G^2(6)$: 1784.185 (Likelihood ratio/deviance statistic)

$X^2(6)$: 7801.602 (Chi-square goodness of fit)

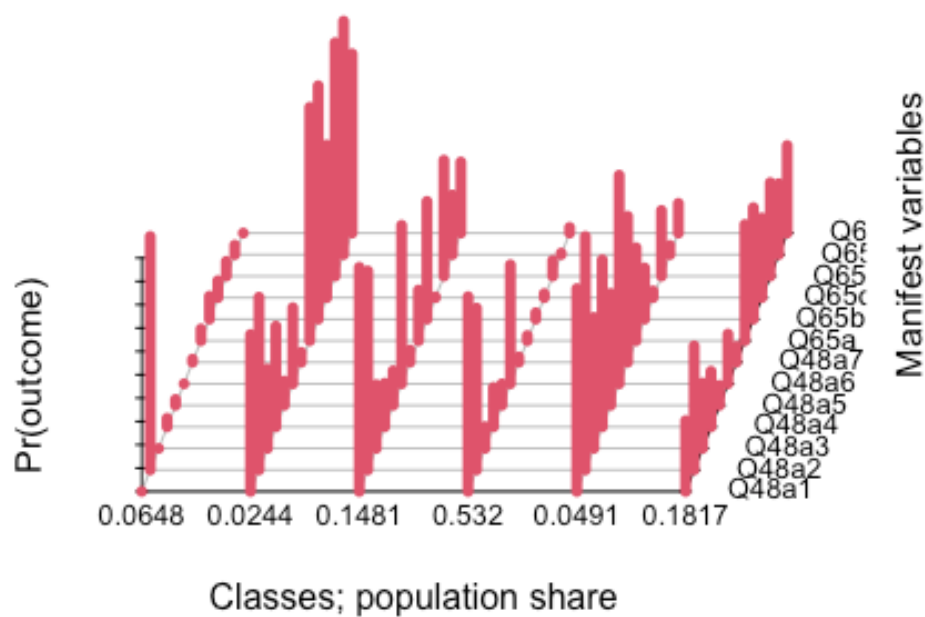


Figure 5 LCA for $n_{class} = 6$

For $n_{class} = 7$

```
> LCA7 <- poLCA(f1, data = df4865, nclass = 7, graphs = TRUE, na.rm = TRUE)
```

Conditional item response (column) probabilities,

by outcome variable, for each class (row)

\$Q48a1

Pr(1) Pr(2)

class 1: 0.0883 0.9117

class 2: 0.1931 0.8069

class 3: 0.8097 0.1903

class 4: 0.1495 0.8505

class 5: 0.6349 0.3651

class 6: 0.6522 0.3478

class 7: 0.0947 0.9053

\$Q48a2

Pr(1) Pr(2)

class 1: 0.0000 1.0000

class 2: 1.0000 0.0000

class 3: 0.4603 0.5397

class 4: 0.0000 1.0000

class 5: 0.4119 0.5881

class 6: 0.0000 1.0000

class 7: 0.1965 0.8035

\$Q48a3

Pr(1) Pr(2)

class 1: 0.8813 0.1187

class 2: 0.8917 0.1083

class 3: 0.6257 0.3743

class 4: 0.3860 0.6140

class 5: 0.7470 0.2530

class 6: 0.9972 0.0028

class 7: 0.6875 0.3125

\$Q48a4

Pr(1) Pr(2)

class 1: 0.7554 0.2446

class 2: 0.9108 0.0892

class 3: 0.6481 0.3519

class 4: 0.2719 0.7281

class 5: 0.7581 0.2419

class 6: 0.9367 0.0633

class 7: 0.8035 0.1965

\$Q48a5

Pr(1) Pr(2)

class 1: 0.8884 0.1116

class 2: 0.9124 0.0876

class 3: 0.9292 0.0708

class 4: 0.4549 0.5451

class 5: 0.9239 0.0761

class 6: 0.9498 0.0502

class 7: 0.8376 0.1624

\$Q48a6

Pr(1) Pr(2)

class 1: 0.3039 0.6961

class 2: 0.6285 0.3715

class 3: 0.7898 0.2102

class 4: 0.1179 0.8821

class 5: 0.7808 0.2192

class 6: 0.9213 0.0787

class 7: 0.3754 0.6246

\$Q48a7

Pr(1) Pr(2)

class 1: 0.9744 0.0256

class 2: 0.9622 0.0378

class 3: 1.0000 0.0000

class 4: 0.1797 0.8203

class 5: 0.9236 0.0764

class 6: 0.9880 0.0120

class 7: 0.9215 0.0785

\$Q65a

Pr(1) Pr(2)

class 1: 0.9642 0.0358

class 2: 0.9542 0.0458

class 3: 0.8073 0.1927

class 4: 0.5461 0.4539

class 5: 0.2175 0.7825

class 6: 0.9642 0.0358

class 7: 0.7312 0.2688

\$Q65b

Pr(1) Pr(2)

class 1: 0.9668 0.0332

class 2: 0.9898 0.0102

class 3: 0.4327 0.5673

class 4: 0.7548 0.2452

class 5: 0.4646 0.5354

class 6: 0.9183 0.0817

class 7: 0.3824 0.6176

\$Q65c

Pr(1) Pr(2)

class 1: 0.9831 0.0169

class 2: 0.9637 0.0363

class 3: 0.6787 0.3213

class 4: 0.9805 0.0195

class 5: 0.5184 0.4816

class 6: 0.9674 0.0326

class 7: 1.0000 0.0000

\$Q65d

Pr(1) Pr(2)

class 1: 0.8945 0.1055

class 2: 0.9234 0.0766

class 3: 1.0000 0.0000

class 4: 0.7121 0.2879

class 5: 0.2969 0.7031

class 6: 0.9268 0.0732

class 7: 0.4690 0.5310

\$Q65e

Pr(1) Pr(2)

class 1: 0.9968 0.0032

class 2: 0.9845 0.0155

class 3: 0.8630 0.1370

class 4: 0.9498 0.0502

class 5: 0.4812 0.5188

class 6: 0.9638 0.0362

class 7: 0.6728 0.3272

\$Q65f

Pr(1) Pr(2)

class 1: 0.9567 0.0433

class 2: 0.9663 0.0337

class 3: 0.6721 0.3279

class 4: 0.8879 0.1121

class 5: 0.5111 0.4889

class 6: 1.0000 0.0000

class 7: 0.6481 0.3519

Estimated class population shares

0.3312 0.1714 0.0647 0.0349 0.1165 0.1422 0.139

Predicted class memberships (by modal posterior prob.)

0.385 0.1801 0.0611 0.0322 0.1037 0.1093 0.1286

=====

Fit for 7 latent classes:

=====

number of observations: 1244

number of estimated parameters: 97

residual degrees of freedom: 1147

maximum log-likelihood: -6977.424

AIC(7): 14148.85

BIC(7): 14646.08

$G^2(7)$: 1755.25 (Likelihood ratio/deviance statistic)

$X^2(7)$: 10155.29 (Chi-square goodness of fit)

ALERT: iterations finished, MAXIMUM LIKELIHOOD NOT FOUND

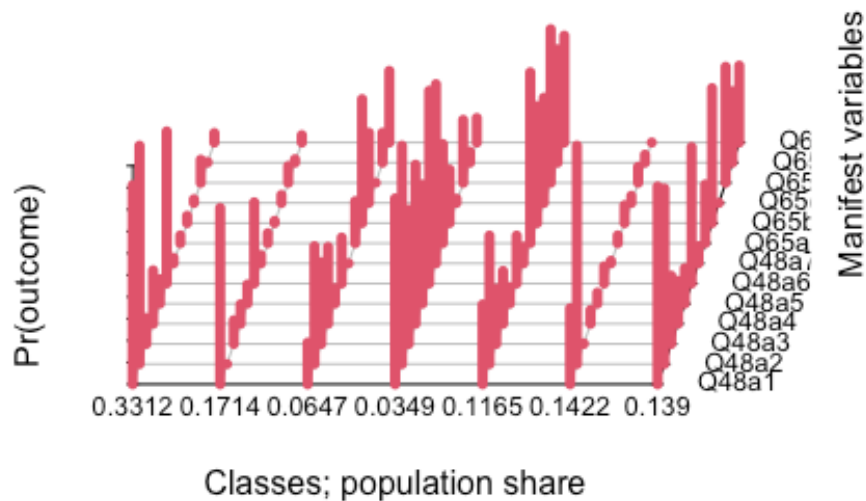


Figure 6 LCA for nclass = 7

Every consecutive iteration of the LCA after $n_{class} = 6$ returns the same alert, signifying that there cannot be any more classes than 6, in which the students might be grouped by. It is the same “tell” that signifies the end of the latent class analysis for this data frame.

3 Findings

The statistical analysis begins by appropriately organizing the data, as previously mentioned. It is imperative to expunge the unneeded values before commencing the calculations, as failure to do so results in corrupt results that will not represent the reality. This train of thought follows closely the famous proverb within the science of IT; “junk in – junk out”, meaning that if the input data are of low quality – inappropriate for the task at hand, then the result will be equally “junk”, and far off the real situation. For this reason, there was extensive thought given to the data, in order to be appropriate for the analysis. Meanwhile, the aim was to expunge as little observations as possible, in order to have a sample as close to the population as possible.

After organizing the data and expunging the values which would negatively impact the outcome, the analysis commenced. The analysis was run for 12 scenarios, where in each scenario only the number of classes would change. In particular, the analysis was run for classes ranging from $n_{class} = 2$ until $n_{class} = 12$. While executing the analysis for $n_{class} = 4$, an alert popped up after the analysis was finished, as can be observed in the previous section of the paper. This alert signifies that the hypothesised number of classes for this run of the analysis is inappropriate for the set of data. Plainly put, the participants are not meant to be split into 4 groups – it is simply not a good fit. The same simple signifying alert was shown for $n_{class} = 7$ and above. Ergo, the classes above 6 are not a good fit, and are rejected. The results for classes 2 until 7 are mentioned above, including $n_{class} = 4$ and $n_{class} = 7$, for reasons of posterity, and to show why a 4 latent class and 7 latent class models are rejected. The remaining repetitions until $n_{class} = 12$ were omitted, since the corresponding results would be irrelevant. Turning the focus on the feasible classes, it is apparent that the students could be divided in either 2, 3, 5, or 6 distinct classes, in which they would be categorized according to their use of the Internet and the level in which they face problems in aspects of their life due to computer games. Before determining which number of classes is more appropriate, the results of the analyses have to be discussed. What is presented in the console printouts of the previous section are calculations, specifically on participants who answered e.g. the first choice of question 48 and what is the probability that these people will belong in one or the others of the hypothesized classes. Simply put, what is

calculated is how likely it is for people who answered the first option of question 48 to belong in the first class, how likely it is that they belong to the second class, etc. The results are presented as decimal numbers ranging from 0 to 1, where 0 means that these people will surely not be included in that particular class, whereas 1 is a definite inclusion of said people to that class. Continuing, the probability results are presented in a linear form, firstly as estimated shares of the sample group to the classes, and afterwards as the predicted shares that each class holds over the sample group. It is important to note, that after expunging the irrelevant observations in the first steps of the preparation of data, the number of participants has dropped only by a little bit. The observations of question 65 should be considered the starting point, ergo $n=1337$, since only the 1337 participants answered both questions. In the LC analysis it is shown that the number of observations – participants, presented as rows of data – has been reduced to 1244. What this means is that the sample of participants is still large enough to be representative of the population. After all, that is the main purpose of taking a sample from a population; easier to keep track of data for fewer observations, fewer observations needed to perform calculations, less expensive in both time and money to perform studies in samples rather than whole populations, all while keeping the sample as diverse as possible, and as close to representing the whole of the population as possible.

In order to decide on which number of classes is more appropriate, though, the attention has to move to the last part of the analysis, the “Fit for x latent classes”. As stated earlier, a tell-tale of whether a number of classes is a good fit or not is the indication of the alert at the end of the analysis. This however, in this example of LCA, leaves the numbers 2, 3, 5 and 6 as the possible number of classes that the observations can be clustered. This is where the Goodness of Fit tests step in, and the code package for LCA in R and R Studio calculates the most “popular”, or widely accepted tests, automatically. The following criteria are automatically calculated; AIC, BIC, G^2 and X^2 . The first two criteria, Akaike’s Information Criterion – abbreviated AIC – (Akaike, 1987) and Bayesian Information Criteria – abbrev. BIC – (Schwartz, 1978) are criteria based on the information available from the observations. BIC in particular is considered to be a good indicator for choosing the best size of classes, but often can be misleading, as at times BIC can indicate one number of classes to be appropriate, but other criteria such as X^2 , which is considered a more solid indicator, can indicate another (Hagenaars & McCutcheon, 2002). By putting the BIC

Tsalavoutas – Tsakiroglou ©2022

values in a table, it is relatively easy to determine the best fit of the number of classes for this sample, as according to theory, the smallest number of BIC is the indicator of best fit (Nylund et al., 2007).

Classes	BIC
2	14564.55
3	14471.36
5	14532.37
6	14575.25

It is apparent that the BIC indicator points towards the 5 classes, as it is relatively the smallest number amongst the results of BIC for each repetition of the analysis.

Turning the attention to the other criteria of goodness of fit, namely the G^2 and X^2 , time and again it is discussed that these goodness of fit criteria are again based on the smaller value – the smaller the value the better the fit. It has to be noted though that through G^2 and X^2 only the efficiency of the model can be measured, meaning that classes with bigger G^2 and X^2 values are not to be rejected, but are rather considered as inefficient for that particular set of data and therefore the one with the smallest value is the most efficient (Nylund et al., 2007). While a complete analysis on the ways to determine AIC, BIC, G^2 and X^2 are beyond the scope of this paper, the use of theory concerning these goodness of fit tests is, and therefore a conclusive answer on what number of classes should the group be clustered as can be drawn. Using a table again, the values can be observed more easily.

Classes	G^2	X^2
2	2.172.553	11724.09
3	1.979.593	10181.12
5	1.841.074	8405.01
6	1.784.185	7.801.602

It is apparent from the table that the most efficient number of classes appears to be $n_{class} = 6$, while the BIC suggested that the 5 classes were better fitting. As mentioned earlier, BIC

is only an indicator, while statisticians accept the Chi Squared Goodness of Fit test as the dominating test to find the most appropriate number of classes in LCA. Ergo, the most appropriate number of classes is 6.

4 Final discussion

While the data available are not so vast in order to conclude on which the classes are, or plainly put to identify each class, it is apparent that the number of classes that the students should be divided to is 6. What this sample Latent Class aimed to achieve is to showcase the process of a standard LCA using a statistical package, in this case the graphical front-end of R, R Studio. The aim of the analysis was to inquire whether the sample of students could be divided into groups, and in extension the population of students, in x classes, based on whether there was a correlation between children who use the internet and children who frequently play computer games. This aim was achieved, and it was proved that the most appropriate number of classes in which the students could be divided by, based on the criterion of using the internet and frequently playing video games.

While the findings of this thesis are not different from other similar findings in numerous other papers – after all the Latent Class Analysis is a well-discussed subject, the importance of this paper lies elsewhere. This thesis aimed to become an easy to comprehend introduction to the Latent Structure Analyses, defining the basic “members” of the Latent Structure Analyses “family”, namely Factor Analysis and Latent Class Analysis, as well as a step by step how to guide on choosing the right “member” when dealing with sample surveys, such as the HBSC survey on Greek children and young adolescents. Furthermore, it aimed to provide a guide on how to perform the chosen analysis using the most basic statistical package for computers, which happens to be one of the few statistical packages that is free to download and use – something that is directly opposed to other statistical packages like SPSS, Mplus etc, which require the prospective researcher to pay in order to download and use them. So far in the statistical bibliography, there has not been another paper which tries to combine the above subjects, in an easy to comprehend manner. While papers – and online guides – which try to explain or introduce a person to the Latent Structure family and the way to perform a Latent Class Analysis do exist, they do one or the other, covering partially the subject.

This is the reason why this thesis is important, as it is unique in its character and the way it becomes a “one-stop guide” on the subject, containing as much information possible, while keeping the terms as easy to understand as possible, in order to be easy to comprehend towards people unfamiliar with the terms used in the statistics field. This paper did not try to “water down” the subjects by removing parts – either from the procedure or theory – but rather tries to put all relevant information on the subject as plain as possible.

Having achieved that goal or not is ultimately decided by the readers and prospective researchers. However, the most important subject that has to be said is that the Science of Statistics and its bibliography needs more theses that are easy to understand for people not well-versed in statistical jargon or thinking in terms of statistics. The bibliography needs easy to understand guides on how to achieve the required analyses, why one should perform this analysis instead of the other, and how to perform them using statistical packages that are freely available, rather than paywalled ones. Without the existence of such papers, most people not relevant with the field of statistics will continue to perceive Statistics as an impossible task, a mythical being that is so perplexed and difficult to research, that they dare not come close. Without such easy to comprehend, complete guides, people are deterred from venturing into the world of statistics, and it is quite frankly a shame, as the world of Statistics is a wondrous one, and should be a welcome place to “newcomers”.

5 References

- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317–332.
- Bandalos, D.L. (2017). *Measurement Theory and Applications for the Social Sciences*. The Guilford Press.
- Bandalos, D.L.; Boehm-Kaufman, M.R. (2008). "Four common misconceptions in exploratory factor analysis". In Lance, Charles E.; Vandenberg, Robert J. (eds.). *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences*. Taylor & Francis. pp. 61–87 ISBN 978-0-8058-6237-9.
- Cattell, R. (1966). "The scree test for the number of factors". *Multivariate Behavioral Research*. 1 (2): 245–76. doi:10.1207/s15327906mbr0102_10
- Currie C, Griebler R, Inchley J, Theunissen A, Molcho M, Samdal O, Dür W & (eds.) (2010). *Health Behaviour in School-aged Children (HBSC) Study Protocol: Background, Methodology and Mandatory Items for the 2009/10 Survey*. Edinburgh: CAHRU & Vienna: LBIHPR.
- Einola, K., & Alvesson, M. (2021). Behind the Numbers: Questioning Questionnaires. *Journal of Management Inquiry*, 30(1), 102–114. <https://doi.org/10.1177/1056492620938139>
- Formann A.K., (1982), Linear logistic latent class analysis. *Biometrical Journal*, 24: 171-190
- Formann A.K., (1985), Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology*, Volume 38, Issue 1, p. 87-111
- Garrido, L. E., Abad, F. J., Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods* 18(4): 454-474. doi:10.1037/a0030005
- Goodman L.A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I - A modified latent structure approach. *American Journal of Sociology*, 79: 1179-1259.
- Goodman L.A., (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61: 215-231.
- Goodman, L. A., Kruskal, W. H. (1959). Measures of association for cross classifications. II: Further discussion and references. *Journal of the American Statistical Association*, 54(285), 123–163. <https://doi.org/10.1080/01621459.1959.10501503>
- Haberman, S.J., (1974). Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations. *The Annals of Statistics*, 2: 911-924.
- Haberman, S.J. (1979) *Analysis of qualitative data: Volume 2. New developments*. Academic Press, New York.
- Hagenaars J.A., (1988), Latent structure models with direct effects between indicators: Local dependence models. *Sociological Methods & Research*, 16: 379-405.
- Hagenaars J.A. (1993), *Loglinear models with latent variables*. Sage, Thousand Oaks, CA.

- Hagenaars J.A., (1998) Categorical causal modeling: Latent class analysis and directed log-linear models with latent variables. *Sociological Methods & Research*, 26:436-486.
- Hagenaars, J., & McCutcheon, A. (2002). *Applied latent class analysis models*. New York: Cambridge University Press.
- Horn, J. L. (1965). "A rationale and test for the number of factors in factor analysis". *Psychometrika*. 30 (2): 179–185. doi:10.1007/BF02289447
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441. <https://doi.org/10.1037/h0071325>
- Kaiser, H. F. (1960). "The Application of Electronic Computers to Factor Analysis". *Educational and Psychological Measurement*. 20 (1): 141–151. doi:10.1177/001316446002000116.
- Kokkevi, A., Stavrou, M., Kanavou, E., Fotiou, A., Richardson, C. (2017). Adolescents in Greece in time of economic crisis. *Child Indicators Research*, 11(3), 945–962. <https://doi.org/10.1007/s12187-017-9458-7>
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton, Mifflin.
- Linzer, D. A., Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10). <https://doi.org/10.18637/jss.v042.i10>
- Ruscio, John; Roche, B. (2012). "Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure". *Psychological Assessment*. 24 (2): 282–292. doi:10.1037/a0025697
- Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.
- Spearman, C. E. (1904). 'General intelligence' objectively determined and measured. *American Journal of Psychology*, 15, 201-293.
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.
- Thurstone, L. L. (1947). *Multiple-Factor Analysis*. Chicago: University of Chicago Press.
- Tran, U. S., & Formann, A. K. (2009). Performance of parallel analysis in the number of components to retain". *Psychological Bulletin*. 99 (3): 432–442. doi:10.1037//0033-2909.99.3.432
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3), 321–327. <https://doi.org/10.1007/bf02293557>
- Zwick, William R.; Velicer, Wayne F. (1986). "Comparison of five rules for determining the number of components to retain". *Psychological Bulletin*. 99 (3): 432– 442. doi:10.1037//0033-2909.99.3.432
- Warne, R. T.; Larsen, R. (2014). "Evaluating a proposed modification of the Guttman rule for determining the number of factors in an exploratory factor analysis". *Psychological Test and Assessment Modeling*. 56: 104–123.

Appendix

Questions 48 and 53 in Greek

48. Γενικά, σε ποια πεδία υπήρξε ΧΡΗΣΙΜΟ το διαδίκτυο για σένα; (Σημείωσε ότι ισχύει)

<input type="radio"/> δουλειά του σχολείου	<input type="radio"/> απόκτηση νέων δεξιοτήτων, όπως _____
<input type="radio"/> διατήρηση επαφής με φίλους και συγγενείς	<input type="radio"/> συνεχής ενημέρωση πάνω στις εξελίξεις
<input type="radio"/> καταπολέμηση μοναξιάς	<input type="radio"/> δημιουργία ή συμμετοχή σε νέες ομάδες ή κοινωνικά κινήματα, όπως _____
<input type="radio"/> απόκτηση νέων φίλων στην πραγματική ζωή	<input type="radio"/> Άλλο (Παρακαλώ διευκρίνισε: _____)
	<input type="radio"/> Το διαδίκτυο δεν έχει υπάρξει χρήσιμο σε εμένα

Αρχείο Q48.sav (μορφή SPSS), n =

1892 Q1 Φύλο

Q20 Ηλικία 1^{ης} χρήσης του διαδικτύου

Q48a1 έως Q48a8 – οι πρώτες 8 υποερωτήσεις. Το «1» σημαίνει ότι σημειώθηκε, το «0» 'οχι. Η τελευταία («δεν έχει υπάρξει χρήσιμο») δεν περιλαμβάνεται στο αρχείο.

Age Ηλικία του παιδιού

Ed Μορφωτικό επίπεδο των γονέων

IAT Κλίμακα που μετράει την υπερβολική χρήση του διαδικτύου, σε 4 κατηγορίες 1...4.

Το 4 δείχνει τη χειρότερη κατάσταση («εξάρτηση από το διαδίκτυο») αλλά αφορά λίγα παιδιά και αν θέλουμε μπορεί να μπει μαζί με το 3 ως μία κατηγορία.

Υπάρχουν κενά (missing values) στις μεταβλητές Q20, Ed και IAT.

53. Πόσο συχνά παίζεις παιχνίδια στον υπολογιστή;

- ☐ κάθε μέρα
- ☐ 2-3 φορές την εβδομάδα
- ☐ μια φορά την εβδομάδα
- ☐ μια φορά το μήνα
- ☐ λιγότερο από μια φορά το μήνα
- ☐ ποτέ

Αν απάντησες «ΠΟΤΕ» ή «ΛΙΓΟΤΕΡΟ ΑΠΟ ΜΙΑ ΦΟΡΑ ΤΟ ΜΗΝΑ» ΠΡΟΧΩΡΗΣΕ ΣΤΗΝ ΕΡΩΤΗΣΗ 66

65. Έχουν ποτέ προκύψει αρνητικές συνέπειες ή προβλήματα στους ακόλουθους τομείς, ως αποτέλεσμα της συμπεριφοράς σου που σχετίζεται με τα παιχνίδια στον υπολογιστή;

		Ναι	Όχι
A	Προβλήματα στη δουλειά, στην προπόνηση ή στο σχολείο (πχ. κακοί βαθμοί).	<input type="radio"/>	<input type="radio"/>
B	Προβλήματα με την οικογένεια / τον σύντροφο ή με φίλους (πχ. τσακωμοί).	<input type="radio"/>	<input type="radio"/>
Γ	Οικονομικά προβλήματα (πχ. χρέη).	<input type="radio"/>	<input type="radio"/>
Δ	Παραμέληση άλλων δραστηριοτήτων αναψυχής	<input type="radio"/>	<input type="radio"/>
Ε	Παραμέληση φίλων / συντρόφου	<input type="radio"/>	<input type="radio"/>
ΣΤ	Προβλήματα υγείας (πχ. έλλειψη ύπνου, κακή διατροφή)	<input type="radio"/>	<input type="radio"/>

Αρχείο Q65.sav, n = 1337 – είναι μόνο τα παιδιά που απάντησαν στην Ερ. 65 (λόγω κατάλληλης απάντησης στην Ερ. 53).

Οι μεταβλητές Q1, Q20, Age, Ed, IAT όπως και στο άλλο αρχείο