Università degli Studi di Milano Bicocca
Dipartimento di Statistica
Dottorato di Ricerca in Statistica
XXIII *ciclo* SECS-S/01

Mixture factor model for hierarchical data structure and applications to the Italian educational school system

Daniele Riggi

Tutor: **Prof. Salvatore Ingrassia**

Co-tutor: **Prof. Jeroen K. Vermunt**

Coordinator: **Prof. Giorgio Vittadini**

To my parents, because they always support me in each choice of my life

Acknowledgement

Nome
Data

Contents

List of Figures

List of Tables

Introduction

In the past years, the educational context has increased its importance. Reading ability has became a fundamental dimension that should be taken into account in maximizing success in daily life and realizing own potential. A proof in favor of this is the fact that a well literate population is essential to a nation social growth and economic development. For these reasons, the International Association for the Evaluation of Educational Achievement (IEA) realized the Progress in International Reading Literacy Study (PIRLS) survey. In this way, countries could take wise and reasonable decisions on the reading education context. This survey provides data for an internationally comparison between students and country on the reading achievement in the primary school. The subjects are fourth grade students of 40 different countries. The choice of the fourth grade students depends on the fact that this age is an important transition point in children development as readers, because *"most of them should have learned to read, and are now reading to learn"*.

In our work, attention has been focused on the survey PIRLS 2006. The IEA conducts comparative studies between student achievement. With this survey, IEA wants to provide information to make possible the comparison between different educational policies and practices around the world. For almost 50 years, the IEA has carried out studies on different topics, such as mathematics, science, civics, technology, and reading. With studies as the Program for International Student Assessment(PISA), Trends in International Mathematics and Science Study (TIMMS), International Study of Computer and Information Literacy (ICILS) and PIRLS, IEA gave significant contributions to increase the knowledge of the educational process. Other aspects investigated by these surveys are the cross-national analysis of the education systems, school organizational and instructional practices. They want to measure the trends in the student ability. The comparison of students between and within countries is a useful tool to analyze and individuate the differences inside and outside the countries.

PIRLS 2006 is the second edition of IEA international studies on reading literacy. PIRLS is responsible for monitoring international trends on reading achievement in primary school each 5-years. The previous release of PIRLS in 2001 was conducted in 35 countries around the world. PIRLS 2001 was based on a newly developed framework. It described interactions between two reading purposes (literary and informative) and a range of four comprehension processes. The assessment was based on a variety of "authentic" texts taken from children reading materials. PIRLS 2006 was conducted in 40 countries, including Belgium with 2 education systems and Canada with 5 provinces (45 participants in total). The two purposes of PIRLS 2006 were the literacy experience, and the attainment or use of the information. The four processes related to comprehension were:

- the focus on and retrieve explicitly stated information,
- to make straightforward inferences,
- to interpret and to integrate ideas and information
- the examination/evaluation of content, language and textual elements.

The main aim of these surveys is the countries comparison on the bases of the measured student level achievement in literacy, mathematics and science. They want to analyze how the fourth-grade students develop reading literacy skills, behaviors and attitudes at home and school. The home and the school experiences are often affected by the community and the country in which students live and attend school. Cultural, social and economic factors can have an influence on the success of an educational system of a country. In our approach, the student ability is not the only interesting aspect useful to evaluate the educational system. Other dimensions should be taken into account to compare students within and between countries. One of the most important is the children reading motivation, as pointed by the educational and psychological literature.

In this work, our attention has been focused on the analysis of children reading motivation in the Italian subsample of the PIRLS data, provided in Italy by the INVALSI [1]. Why motivation is so important to analyze the educational systems and to compare them between countries. The most important reason is related to the fact that motivation is connected to the students educational level and a strong relation have been found with the student abilities. Underlie the motivation development process could be interesting to deeply understand the children ability in reading. Another reason

[1] National Institute for evaluation of the educative system of instruction and formation (http://www.invalsi.it/invalsi/index.php)

is based on the evidences found in the educational research. Many studies affirmed that from the elementary school to the middle school, it is possible to see a decline of motivation and performance. This could be related to some psychological aspects. Other studies demonstrated that this change in motivation depends from the characteristics of the learning environment. The family and the teacher effects have been incorporated into the analysis, because many studies highlighted the important role and the effects that external factors such as home and school context have on motivation. This work has several purposes. Different models have been used to describe the existent theory on motivation and to discover the real situation of the Italian education context.

In our application two classes of models have been implemented: the Mixture Factor Models (MFM) Lubke and Muthén (2005) and the Multilevel Mixture Factor Models (MMFM) Varriale and Vermunt (2009). These two kind of models belong to the generalized latent variable modeling framework. This work is an attempt to develop this kind of latent variable models and increase the use of them in the educational context that present very interesting and challenging aspects. Not so many works have been published on this topic, and in most of them the theoretical approach was focused on models with only one type of latent variables (continuous or categorical). In this thesis, considering both the applied and the theoretical point of view, a very general framework has been proposed. Continuous and categorical latent variables have been used to describe the reading motivation and to classify group level units. In our application, the models implemented combine the features of both the factor analysis and latent class models. The principal difference between the models presented above these models is the position of the latent class (mixture component), because in the Mixture Factor Model, the latent class is at individual level, instead in the Multilevel Mixture Factor Model is at group level.

The Mixture Factor Model has been thought for one level data structure. In our application, the data present a hierarchical structure, but in the implementation of this model we did not include the group level structure. The students have been classified ignoring the fact that they share the same environment and teacher. In this case, we are ignoring the fact that students belonging to the same class. They have in common the class, the way of teaching, the climate and the relations that characterize a class. The aim of this model application is to obtain the classification of students in groups according to their reading motivation.

In the Multilevel Mixture Factor model, the fact that students belong to the same classrooms, sharing the common environments, experiences and interactions is taken into account. To discover the relevance of considering the hierarchical data specification, the multilevel techniques have been used. These techniques recognize the existence of data hierarchies by allowing for residual components at each level in the data structure (Snijders and Bosker (1999)). They allow correct inferences treating the units of the analysis as dependent observations. In this thesis, different specifications of the multilevel mixture factor model have been used. One of the goals is to explain the correlation among items in terms of unobserved variables, called factors. The underlying concept we want to measure is the students reading motivation, and it has been considered as a three dimensions concept. With these models, the target is the classification of teachers according to the students reading motivation. This is an important consideration because not all the classes or the teachers are equal. The analysis of the class differences is a way to understand the teachers contribution on the students reading motivation, after controlling for the home context, described with the family and students covariates. The latent class approach has been introduced by Hagenaars and McCutcheon (2002). The models used in this analysis has been proposed by Vermunt (2003).

The application has been carried on in the following way. Firstly, an exploratory and confirmatory factor analysis have been implemented. The exploratory factor analysis has been used to discover the relation between the items and the latent construct. With the confirmatory factor analysis, we want to show the evidences to confirm the hypothesized latent structure at individual level measuring reading motivation. The second step of the analysis was the introduction of a mixture component at individual level. It has been implemented the mixture factor model to classify students according to the latent factor structure. We want to highlight and isolate the "pure" students reading motivation. Some individual covariates have been introduced to the mixture factor model to explain the influences of the home context (family habits and behaviors towards reading) and personal student characteristics. In this way, it could be possible to discover the home influences on motivation, and trace a profile of motivated students and of an involving context.

Successively, a multilevel mixture factor model has been implemented. In this model, the latent class has been shifted from the individual level to the group level. The individual level structure is the same of the previous models. The use of a continuous latent variables at lower level, reduces the dimensionality of the phenomenon in analysis, indeed few dimensions are more interpretable than many. In our application, we summarized 11 items with three latent factors. The mixture at group level is useful to classify teachers according to the individual latent structure. The aim of the multilevel mixture factor model is to classify the group level observations according to the unobserved concept measured at the individual level. The determination of the effects of the teachers in the process of motivating students is one of the aims of this work.

In this model, both lower and higher level covariates have been added. The purpose of the introduction of the lower level covariates is to describe the influences of the home context and of the personal students characteristics on the individual factor structure. The higher level covariates are useful to describe the teacher abilities, practices and behaviors in classrooms. The explanation of how students motivation increase with teacher practices is another goal of the introduction of the higher level covariates. The role of covariates at both the levels concerns with the possibility of evaluate the "net" effect of teachers on reading motivation controlling for the composition effect.

For the implementation of the mixture factor model and of the multilevel mixture factor model, the syntax module of the Latent GOLD software version 4.5 (Vermunt and Magidson (2007)) has been used. With Latent Gold, it is possible to deal with models containing any combination of categorical and continuous latent variables at each level of the hierarchy.

In our application, the Italian subsample of the PIRLS data, provided in Italy by the *INVALSI* has been selected. In this application, the comparison between countries were not considered, because our interest was the description of the Italian situation. This extension could be investigated in the future. The aim of the PIRLS survey is the analysis of the reading ability of the 4^{th} grade students through a cognitive questionnaire. The external factors (home and school) on reading ability are investigated in the other data as the students, family or teacher datasets.

A short presentation of the data is provided to make clear the data structure. *INVALSI* data are subdivided into five sub-datasets. Each of them deal with a different aspects related to the educational context: Cognitive, Student, Family, Teacher and School. These datasets describe the components of an educative system, analyzing the different aspects related to the involved subjects (students, families, teachers and schools).

Cognitive dataset collects the answers to a questionnaire that investigate the students reading ability. These answers have been summarized through some variables indicating the plausible values obtained according to the different reading areas (overall reading, literacy purpose, informational purpose, interpreting process and straightforward process). Plausible values are obtained with different weighting strategy. These variables have been marginally taken into account during our analysis, because our purpose was the description of the students reading motivation.

Students dataset contains different sections, such as personal information, activities outside school, school and home reading habits, computer and library use, thoughts about reading and school description. This dataset contains *3581* students, (*1742* of them are girls and *1839* boys), *3370* were born in Italy and *188* not.

Family dataset is composed by *3581* observations (one student - one family). It contains information about family and reading habits characteristics. This dataset deals with different topic, such as children activities before and after they were *ISCED Level 1*, types of involvement with child, thoughts about school, importance of literacy at home, level of education, employment situation, kind of job and financial situation.

Teacher dataset collects information on the class environment and the way of teaching, such as language and reading instruction, use of computer and library resources, homework, reading difficulties, assessment and relationship between home and school. Personal information on teachers have been collected like age, sex, level of formal education, license or certificate owned, hours spent in professional development, reading habits and job satisfaction.

In school dataset the information collected describe some characteristics of the school, as students enrolled, instruction, reading, resources, home-school relation, school climate, teacher collaboration and role of the Principal.

The variables used as covariates in the models, have been selected to explain the relations between the reading motivation and the home and school context. These analysis try to highlight the importance of students reading motivation. Reading motivation is a field on which many psychological and educational studies have been conducted. It is an useful instrument to understand children attitudes and to describe the scholastic system. Motivation or attitudes towards reading could also be used to predict the future children choices or to understand the possible intervention/modification to improve the way of teaching.

Chapter 1 is dedicated to the the introduction to the latent variable framework and to a brief historical introduction of this model. The principal works on this topic have been published in the last ten years (Skrondal and Rabe-Hesketh (2004), Vermunt (2007a), Muthén and Muthén (2007)).

Chapter 2 is dedicated to the description of the latent variable model, with particular attention to the mixture factor model Lubke and Muthén (2005) and the multilevel mixture factor model Varriale and Vermunt (2009). Section 2.1 introduces the different kind of outcomes and the way of using the linear predictor in factor model. In section 2.2, the Mixture Factor Model has been presented. The structure could be described in the following way. Three continuous latent factors describe the motivation structure at individual level. A mixture component is introduced to group students according to the reading motivation. After that, the lower level covariates to explain factor mean or class membership have been introduced into the model to describe the home context. Section 2.3 is dedicate to the Multilevel Mixture Factor Model. The specification is similar to the model presented in section 2.2. The difference is that the mixture component is at group level instead of being at individual level. In this way, the classification concerns with the group level unit (teachers/classes) and not with the individual units (students). In this section, the model is described using the linear predictor approach. Covariates at lower level (on factor score) and at higher level (on teacher membership) have been added. Section 2.4 is dedicated to the technical aspects related to the model estimation procedures. In the subsections 2.4.1, 2.4.2, 2.4.3, the most famous estimation procedures, as the Newton-Raphson algorithms, the Fisher scoring and the Expectation-Maximization algorithm has been described. In subsection 2.4.4, the EM variant proposed by Vermunt (2003) for the multilevel latent variable model has been described. Section 2.5 is dedicated to model evaluation. The fit indexes used in the Latent Variable framework have been presented. The subsection 2.5.1 is dedicated to the *posterior analysis* and to the different purposes of using the continuous or the categorical latent variables specification. With the continuous specification, the aim is assigning factor scores to each individual. With the categorical variables, we want to classify units in classes that differ among latent groups.

Chapter 3 is dedicated to the results of the factor analysis and the mixture factor model. Exploratory and confirmatory factor analysis have been implemented to understand the latent structure measuring motivation. Section 3.1 reports an introduction to the reading context. In subsection 3.1.1, the results of the exploratory and confirmatory factor analysis are reported. In this section we show the reading motivation structure and the items selection process. We introduce and explain the latent structure measuring reading motivation. Section 3.2 is a summary of twenty years of literature about students motivation. In subsections 3.2.1 and 3.2.2 are reported respectively the description of the home (family) and class (teacher) environment. In section 3.3, the results of the mixture factor model are reported. The first subsection 3.3.1 is dedicated to the simple mixture factor model without covariates. The aim is the classification of students according to the latent structure. In section 3.3.2, individual covariates have been added to explain how motivation is influenced by the home environment. Section 3.3.3 is dedicated to the comparison of the classification obtained with the two models previously presented. In section 3.3.4, we try to show the connection between motivation and ability.

Chapter 4 is dedicated to the results of the Multilevel Mixture Factor Model (MMFM). The structure of the MMFM could be described in the following way: at individual level, there are three latent continuous factors, while at group level there is a discrete mixture component to model population heterogeneity and classify group level observations. This section is subdivided into three parts. The first one (section 4.2) reports the results of a simple MMFM, where covariates are not inserted. The aim is to obtain the classification of teachers/classes according to the students latent structure. The second subsection (4.3) reports the results obtained with the introduction of the lower level covariates. We want to describe how the latent structure at individual level is affected by the home context and students personal characteristics. The last subsection (4.4) is dedicated to the results obtained with the introduction of the introducing higher level covariates into the model. We want to individuate the teachers characteristics that have an influence on the class motivation to highlight the *pure* effect of teachers, after controlling for students and families characteristics.

Chapter 5 is dedicated to a summary of the obtained results and to the extensions of the models considering both the theoretical and the application side. The comparison of the results obtained through the two models have been presented.

Chapter 1

Conceptual Introduction

In the past thirty years, in the field of applied and social science, as psychology, education, marketing, biology, and medicine, it has been made possible to see a spread of many statistical tools for data analysis. These disciplines deal with unobserved concepts such as intelligence, skills, attitudes, medical conditions, personality traits, preferences, or perceptions. In many cases, researchers observe only direct indicators. They hypothesize the existence of some relations between the observed variables (items) and the latent concepts. It is possible to define the latent variables as concepts or hypothetical constructs that within a statistical process, influence the observed realizations of a phenomenon. Latent variables can be thought as the "*true*" variables or constructs, while the observed variables as the indirect or fallible measures of that concept (Skrondal and Rabe-Hesketh (2004)).

As highlighted by Bartholomew and Knott (1999): *"The interesting questions concern why latent variables should be introduced into a model in the first place and how their presence contributes to scientific investigation"*.

The first reason is practical. The introduction of the latent variables reduces the dimensionality of the model. As Bartholomew and Knott (1999) wrote: "*[...] our ability to visualize relationship is limited to two or three dimensions places us to under strong pressure to reduce the dimensionality of the data in manner which preserve as much of the structure as possible*". Heinen (1996) supported the explanation reported above in the following way: "*Many concepts that play a crucial role in social and behavioral theories cannot be observed directly [...]. The only way to obtained empirical knowledge on these concepts is to look for variables observed that contain information on the theoretical concepts.[...] the theoretical concepts themselves are not measured directly.[...] the theoretical variables that are not observed directly are denoted by the term* ***latent variable****, whereas the variables that are directly observed and contain information on the latent variables are called manifest*

		Manifest Variables	
		Metrical	Categorical
Latent Variables	Metrical	Factor Analysis	Latent Trait Analysis
	Categorical	Latent Profile Analysis	Latent Class Analysis

Table 1.1: Classification of latent variable methods

variables or indicators".

The second reason is more pragmatic. Latent concepts appear in many fields in which statistical methods are often used. Heinen (1996) wrote that "*In many examples of social research, so many different variables are measured that it becomes necessary to compress these data into a smaller set of variables that is assumed to reflect the common substance of a number of original variables*". Following, the general framework for latent variables and manifest variables have been introduced. The Latent Variable (LV) models could be classified following the idea proposed by Bartholomew and Knott (1999) reported in table (1.1). Latent and manifest variables are subdivided in metrical and categorical. Two of the models are: the Factor Analysis (FA) and the Latent Trait Analysis (LTA). These two classes of models handle both with continuous normally distributed latent variables. In LTA, manifest indicators are discrete, while in FA both manifest and latent variables are continuous. The other two classes deal with categorical latent variables: in the Latent Class Analysis (LCA), the manifest and latent variables are both discrete, instead in the Latent Profile Analysis (LPA), the latent variables are discrete while the manifest are continuous.

The first works on Factor Analysis has been published by Galton, Pearson and Spearman (1904) between 19^{th} and 20^{th} century. In the 1950 and 1960s, thanks to the increment of statistical computing capacity, it has been possible to see a huge development of statistical tools for social science data. Nowadays, Factor Analysis is one of the most popular tool used in quantitative social science research, as psychology, education and other applied sciences.

During the past twenty years, many tools for social and applied research have been proposed. The Latent Class analysis is one of those. This method is often applied when the manifest variables are categorical, and the nature of the latent variables is discrete. The underlying assumption of the latent class analysis is on the population heterogeneity. To model population heterogeneity, several approaches have been proposed. In the latent class approach, the class membership for each observation is unknown and so it should be inferred from the data. The purpose of these models is to identify the nature and the number of latent classes. These models were introduced for the first time by Lazarsfeld in 1950s. The purpose was the explanation of the different ways of responding in a survey with dichotomous items (Vermunt and Magidson (2005b)). Twenty years later, the latent class approach was structured and carried on including also the possibility of dealing with nominal responses. In the same years, Day (1969) and Wolfe (1970) proposed a very similar approach to the latent classes: the Finite Mixture (FM) models. These models were based on the assumption that the observed data have been generated by a finite mixture model, in which each mixture component identifies a different subpopulation (McLahan and Peel (2000)). In recent years, the terms latent class and finite mixture have been interchangeable used.

In the latent variables framework, it is important to choose the nature of the latent variables (discrete or continuous) that we want to introduce in the model. In some context as argued by Aitkin (1999), it is possible to approximate a continuous latent variable distribution with a discrete one. In the non-parametric specification, the continuous distribution is replaced with a finite number of probability masses. One of the advantages of this way of proceeding is the introduction of unverifiable assumptions on the random effects. An important result obtained with both the latent variables specification, has been performed by Muthén (2001). In his work, the relation between the factors analysis and latent class has been shown. This evidence supports the importance of using both the specification to deeply understand the phenomenon under study.

The goal of this work is the analysis of the multilevel framework using the latent variables. Latent variables have been used to analyze several statistical concepts. Different names have been used by Muthén (2002) to refer to the latent variable models, such as common factors, latent classes, random effects, underlying variables, frailties, components of variation, missing data, finite mixtures. In recent years, several frameworks for the latent variables have been proposed. One is the Generalized Linear Latent and Mixed Models (GLLAMM) by Skrondal and Rabe-Hesketh Skrondal and Rabe-Hesketh (2004). This framework includes models for continuous and discrete specification. It is possible to handle with different latent variable data structures as multilevel, longitudinal and structural equation models. The generalized linear mixed models, random coefficient models, item response models, factor models are examples of the models included in this specification. This framework is implemented in different statistical softwares, such as the *GLLAMM* software package Rabe-Hesketh *et al.* (2004c). It is a STATA program for generalized linear latent and mixed models, where models with continuous or discrete latent variables are included. Another program created by Muthén, deals with the same arguments. Mplus (Muthén and Muthén (2007)) allows the definition of models with categorical and continuous latent variables at each level of the hierarchy. Latent GOLD (Vermunt and Magidson (2007)) deals with the same topic. With this software it is possible to estimate models with any combination of categorical and continuous latent variables at each structure level.

1.1 Historical introduction

Nowadays, latent class and mixture models are the most popular tools in the field of applied research. As showed in Vermunt (2008), these tools have different uses, such as classification, scaling, clustering and modeling non parametric random effects. One of the possible use of these models is clustering units through latent classes or mixture components. This clustering could involve categorical response variable, continuous items, mixture factor analysis, repeated measures, mixture Markov models or two level datasets where level one units are nested with higher level units (Vermunt (2011)). In many applied works, researches often encounter multilevel data structure, such as individuals with multiple response or repeated measures nested within groups (Vermunt (2003),Vermunt (2008), Bijmolt *et al.* (2004)), or multivariate repeated responses nested within individuals (Vermunt *et al.* (2008)), or three levels data sets (Vermunt (2007a)). Examples of multilevel data structure are patients nested with doctors or with hospitals, students nested with teachers or schools, individuals nested with regions, or repeated measures (as the annual wage) nested with subjects (individual).

Factor Analysis is an important instrument for dealing with the latent variables. It is one of the most used methods to describe the relationships (associations) between manifest variables (indicators or items) and hidden continuous latent variables (factors). In recent years, factor analysis has been associated also with dichotomous or ordinal indicators. One of the assumptions of the common Factor Analysis is the observations independence. In presence of multilevel data structure, this assumption could not be considered valid, because the individual observed indicators are nested with higher level units. In the education context, students are nested with teachers or schools, and so each individual unit shares the same environment with the other subjects. In the school context, the students (lower level units) share with other classmates the teacher (this is a characteristic of the Italian school system), the way of teaching, the interactions and the school environment. The students responses are influenced by these factors and for such reasons, the independence assumption could be not considered valid. To avoid misspecification, in such situation the multilevel techniques should be considered. These methods take into account the relationships between the observations, and illustrate also the influences on the individual subject of the higher level factors (Goldstein (2003), Hox (2002) and Snijders and Bosker (1999)).

Extensions of the standard factor analysis have been treated by several authors (Longford and Muthén (1992), Muthén (1991) and Goldstein and McDonald (1988)). The multilevel factor models is one of these extension and in them it is allowed to some parameters to vary randomly across groups. The aim of the multilevel factor model is multiple. Describing the role of factors at different level of the hierarchy and understanding the relationships between them. As suggested by Muthén (1994), these models can be used both as exploratory and confirmatory.

In this work, we present two extensions of the common factor model. The aim is to integrate both the aspects of the latent class analysis and factor analysis, dealing also with the multilevel data structure. The models are the Multilevel Mixture Factor Model (Varriale and Vermunt (2009)) and the Mixture Factor Models (Lubke and Muthén (2005)).

The multilevel mixture factor model is a statistical tool used to explore unobserved population heterogeneity in presence of hierarchical data structure. This model combines both the aspects of the common factor models (Thurstone (1947)) and the standard latent class model (Lazarsfeld and Henry (1968)). In the latent variables model, the theoretical concepts are indirectly observable, while the observed variables are the indicators for these unobservable variables. The common factor models are appropriated for homogeneous data. These models are useful to investigate the common content from the observed variables and to cluster items. Latent class models deal with unobserved heterogeneity. They are often used in situations where the source of heterogeneity is unknown. The impossibility to assign an individual to a group is a direct consequence of the absence of information on these variables. For these reasons, group membership should be inferred from the data. The aim of the latent class models is to cluster individuals.

The factor mixture model combines both the aspect of the latent class and the common factor models. These models deal with situation in which the observed variables covary within class, and continuous factors are used to model this covariation. The mixture factor model is useful to model latent factor structure in presence of population heterogeneity, where the source of this heterogeneity is unknown. To model the differences between subjects, a latent class is inserted at individual level.

		Group-level (level 3)	
		Discrete	Continuous
Subject-level (level 2)	Discrete	Multilevel Mixture LC	Multilevel random effects LC
	Continuous	Multilevel Mixture IRT/FA	Multilevel random effects IRT/FA

Table 1.2: Four-fold classification of multilevel latent variable models

The Multilevel Mixture Factor Models (MMFM) is an alternative approach to factor analysis in presence of hierarchical data structure. The MMFA is useful to model the between group differences through K latent class or mixture components. In the classical approach, the group level continuous factors or random effects have been used to model group heterogeneity. We decided to use these models because it is possible to relax the assumption on the homogeneous population of the classical factor models. Latent classes or finite mixture (McLahan and Peel (2000)) are useful to verify this hypothesis and to discover the existence of unknown subgroups in the population. The principal difference between the multilevel mixture factor models and the mixture factor models concerns with the level specification of the mixture component (at group level or at individual level).

In table 1.2, it is possible to find a summary of the different possibilities of the models used in the latent variables framework, considering the different level of of the hierarchy.

The four-fold classification table of latent variable is reported in table 1.2 for the three level data sets. It depends on the classification of the variables scale at level 2 and 3 of the hierarchy. The latent variables could be continuous and discrete. If the latent variables at level 2 were discrete, the types that could be used will be the Latent Class or the Finite Mixture Models, with discrete or continuous random effects. The choice between these two models, depends on the nature of latent variables at the higher level. If the latent variables at the second level of the hierarchy were continuous, the models that could be used will be the Multilevel Mixture models (IRT or FA), with discrete or continuous random effects. A lot of authors confronted themselves with type IV models (Fox and Glas (2001), Goldstein and Browne (2002), Rabe-Hesketh *et al.* (2004a)). We concentrate our attention on the models belonging to the type III. We refer to the model specification adopted by Vermunt (2007a) and Varriale and Vermunt (2009), where a continuous latent variable structure at individual level and a discrete mixture component at group level to model population heterogeneity has been specified.

Another combination of models has been described by Palardy and Vermunt (2007). It is a multilevel extension of the mixture growth model proposed by Muthén (2004). The group level observations are classified into homogeneous classes, according to the mean growth trajectories. These models have been obtained when the lower level latent variables is continuous and the higher level latent variables are both discrete and continuous. Several and flexible models could be obtained using different specification for the latent variables at each level of the hierarchy.

Chapter 2

Model specification

In factor analysis, the observed responses for each subject ($i = 1, \ldots, N$) and indicators ($h = 1, \ldots, H$) is indicated with Y_{hi}. With η_{mi} we define the unknown factor scores for each subject on common factor ($m = 1, \ldots, M$). With N, H and M are indicated the total number of individuals, items and factors. The vector of observed items is $\boldsymbol{y_i}$=$(y_{1i}, ..., y_{Hi})$, and the vector of individual factor scores is $\boldsymbol{\eta_i}$=$(\eta_{1i}, ..., \eta_{Mi})$. Factor analysis is a method to investigate whether a number of observed variables of interest $\boldsymbol{y_i}$ are linearly related to a smaller number of unobservable factors $\boldsymbol{\eta_i}$. The regression methods are disqualified because the factors are not observable. Under certain conditions the factor model has some implications, that should be tested on the observed data. The following sections are dedicate to the factor analysis. In the first paragraph, the common factor model has been described. With the factor analysis, it is possible to handle with latent concept using different specifications. The second section is dedicated to the factor mixture model, an extension of the common factor model in presence of population heterogeneity. The introduction of a latent class at individual level is useful to model the population heterogeneity in the data. It represents a different approach to the multi-group factor analysis, where the source of the population heterogeneity is known. In mixture factor model the source of population heterogeneity is unknown and it should be inferred from the data. The third section is dedicated to the technical details of the multilevel mixture factor model. This model compared with the previous one, include both the multilevel structure and the possibility to model the population heterogeneity. In this case, the mixture component is at group level instead of being at individual level. It is possible to group the second level units. The fourth section is dedicated to the estimation procedures in the latent class context. Particular attention has been addressed to the description of the variant of the EM algorithm implemented in *Latent Gold*. In the last section, the model evaluation methods and the posterior analysis have been presented.

2.1 The factor analysis

The aim of the factor analysis is to find a relation between latent variables η_i and items $\boldsymbol{y_i}$ using some regression methods. With these models, it is possible to deal with different item scale types (categorical, nominal, ordinal, continuous or count variables). For this reason, a particular distributional form must be specified for the response ($\boldsymbol{y_i}$). In presence of categorical variables, a multinomial distribution is usually assumed for the items. For continuous variables, the most used specification is the normal distributions (multivariate / censored / truncated). With counts data the Poisson and Binomial (truncated or over dispersed) distributions are used. Referring to the discrete response variables distribution (ordinal, nominal, Poisson count or binomial count), we will use the symbol $P(.)$ instead of $f(.)$ to indicate that we are dealing with a probability instead of a density function (Vermunt and Magidson (2005b)).

Referring to Skrondal and Rabe-Hesketh (2004), we decide to use the response models from the family of the Generalized Linear Modeling (GLM), in which a linear predictor, a link function and an error distribution from the exponential family must be specified. We indicate the linear predictor for the item h and the subject i with υ_{hi}. For the common factor model, the form assumed by the linear predictor is

$$\boldsymbol{\upsilon}_i = \boldsymbol{\mu} + \sum_{m=1}^{M} \boldsymbol{\lambda}_m \eta_{im} \tag{2.1}$$

where $\boldsymbol{\mu}$ is the item intercept, $\boldsymbol{\lambda}_m$ are the factor loadings and η_{im} are the factor scores. The link function between the linear predictor and the items is

$$g\left(E(\boldsymbol{y}_i|\boldsymbol{\eta}_i)\right) = \boldsymbol{\upsilon}_i \tag{2.2}$$

Equation (2.2) shows that with a suitable transformation $g(.)$, the expected value of $\boldsymbol{y}_i$ given the latent factors, is equal to the linear predictor. Obviously, the choice on $g(.)$ will depend on the scale of indicators. Below, the link function for the binomial and continuous response variables are reported. More details could be found on the Technical Guide for Latent GOLD 4.0 (Vermunt and Magidson (2005b)). For the continuous responses, the identity link function reported below is usually used.

$$E\left[\boldsymbol{y}_i|\boldsymbol{\eta}_i\right] = \boldsymbol{\upsilon}_i$$

For the binary response, the logit link function is used

$$ln \frac{E\left[\boldsymbol{y}_i|\boldsymbol{\eta}_i\right]}{1 - E\left[\boldsymbol{y}_i|\boldsymbol{\eta}_i\right]} = \boldsymbol{v}_i$$

This specification for the binary response implies that the expected value of the response given the latent variables is equal to following quantity.

$$E\left[\boldsymbol{y}_i|\boldsymbol{\eta}_i\right] = \frac{\exp \boldsymbol{v}_i}{1 + \exp \boldsymbol{v}_i}$$

The error distribution should be specified in order to complete the *GLM* specification. The error is given by the difference between the observed and the expected value of the response given the latent variables

$$\boldsymbol{e}_i = \boldsymbol{y}_i - E\left[\boldsymbol{y}_i|\boldsymbol{\eta}_i\right]$$

Otherwise, the conditional density function of the items given the latent variables $f(\boldsymbol{y}_i|\boldsymbol{\eta}_i)$ should be defined. Dealing with continuous items in FA models, normal distribution is usually chosen

$$f(\boldsymbol{e}_i) = f(\boldsymbol{y}_i|\boldsymbol{\eta}_i) \sim MN(0, \Sigma)$$

With the dichotomous response, the Bernoulli distribution is typical used. Considering the conditional distribution function of the latent variables ($\boldsymbol{\eta_i}$), it is implicitly assumed that the observed responses of $\boldsymbol{y}_i$ are independent each other, after choosing a fixed number of latent variables. So

$$f(\boldsymbol{y}_i|\boldsymbol{\eta}_i) = \prod_{h=1}^{H} f(y_{hi}|\boldsymbol{\eta}_i) \tag{2.3}$$

This is called the local independence assumption (Bartholomew and Knott (1999)). It is not possible to make some empirical tests on it, because there is no way to fix the latent variables and test for the independence. The aim of the analysis consists of finding the smallest number of latent variables in order that the local independence assumption is adequate.

2.1.1 The variables scale type

In presence of ordinal indicators, many specifications for the response model are available. In Agresti (2002), it is possible to find a broad description of the different choices for the response model. In Latent Gold, for ordinal dependent variables, it is possible to use the adjacent-category logit model, cumulative responses probabilities (cumulative logit, probit, and log-log models) as well as models for continuation-ratio or sequential logits.

Considering Y a categorical response variable with J categories, multinomial logit models for nominal response variables simultaneously describe log odds for all the possible pairs of categories. Following are described some of these methods. The first one is the Baseline-Category Logits. It is possible to define the probability to observe that Y assume the value j given x a fixed setting for explanatory variables in the following way

$$\pi_j(x) = P(Y = j|x)$$

and

$$\sum_j \pi_j(x) = 1$$

. The counts of the J categories of Y are treated as multinomial probabilities $\{\pi_1(x), \ldots, \pi_j(x)\}$. Logit models pair each response category with a baseline category, often the last one or the most common one. The model is:

$$ln\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j' x \tag{2.4}$$

for *j=1,...,J*. It simultaneously describes the effects of x on these *J-1* logits. The effects vary according to the response paired with the baseline.

Another popular logit model for the ordinal responses is the Cumulative Logits. One way to use the category ordering forms logits of cumulative probabilities is

$$P(Y < j|x) = \pi_1(x) + \cdots + \pi_j(x)$$

for *j=1,...,J*. The cumulative logits are defined as

$$logit\left[P\left(Y \leq j|x\right)\right] = ln\frac{P\left(Y \leq j|x\right)}{1 - P\left(Y \leq j|x\right)} = ln\frac{\pi_1(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \cdots + \pi_J(x)} \tag{2.5}$$

for *j=1, ...,J-1*.
Each cumulative logit uses all J response categories. A model for

$$logit\left[P\left(Y \leq j\right)\right]$$

is an ordinary logit model for a binary response in which the categories from *1* to j form one outcome and the categories from *j+1* to J form the second outcome. Models could use all the *J-1* cumulative logits in a single parsimonious model.

The model that simultaneously uses all the cumulative logits is the Proportional Odds Model:

$$logit\left[P\left(Y \leq j|x\right)\right] = \alpha_j + \beta_j' x \tag{2.6}$$

for $j = 1, \ldots, J-1$. Each cumulative logit has its own intercept. The α_j are increasing in j, since $P\left(Y \leq j|x\right)$ increases in j for fixed x, and the logit is an increasing function of this probability.

In our application, we can denote with s a particular category and with S the total number of category assumed by the items. Let $P\left(\boldsymbol{y}_i \leq s|\boldsymbol{\eta}_i\right)$ the probability of responding to a category less or equal to s assuming the conditioning to the latent variables. The proportional odds could be specified in the following way:

$$ln\left(\frac{P\left(\boldsymbol{y}_i \leq s|\boldsymbol{\eta}_i\right)}{1 - P\left(\boldsymbol{y}_i \leq s|\boldsymbol{\eta}_i\right)} = \boldsymbol{\alpha}_s - \boldsymbol{\upsilon}_i\right) \tag{2.7}$$

for *s=1,...,S-1*. The values of α_s have to be estimated from the data. The important characteristic of this model is that υ_{hi} are invariant for the construction of the ordinal categories of $\boldsymbol{y}_i$.

We need also to specify the distribution of the common factors, and usually in FA it is a multivariate normal distribution,

$$\boldsymbol{\eta}_i \sim MN(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \tag{2.8}$$

where $\boldsymbol{\Sigma}$ is the variance/covariance matrix.

2.2 The Mixture Factor Model

The factor mixture models have been designed to model the unknown heterogeneity in the data is modeled through a latent class. This family of models include many different sub-models that could be obtained applying some modification to the number of the latent classes or to structure of the variance/covariance matrix. To obtain the common factor model, it is necessary to set to one the number of the latent class. In this way, all the observations belong to a single latent class and the factor structure is not influenced. To obtain the latent class and the latent profile models, the variance of the within class factors has to be fixed to zero. In this way, any factor structure is imposed on items and the categorical latent variables are useful to model the unknown population heterogeneity. To obtain the growth mixture model, it is necessary to impose some restrictions on the within class covariance matrix (Muthén (2004),Muthén and Muthén (2000)). This kind of models has some similarities with the multi-group confirmatory factor models proposed by Sorbom (1974). The main difference concerns with the subpopulation heterogeneity source. In the factor mixture specification, the source is unknown and the class membership must be inferred from the data. In the approach of Sorbom (1974), the source of heterogeneity is known and so it should not be inferred from the data. Arminger *et al.* (1999) focused their attention on the estimation of the factor mixture model conditional to covariates, because through this conditioning it is possible to assume the multivariate normality. Muthén and Shedden (1999) included also categorical outcome variables that could be predicted by class membership. The implementation of these models is a step by step procedure. The implementation of the common factor model for a single homogeneous population is the first step of this procedure. The second is the addiction of a latent class / mixture component at individual level to model the source of population heterogeneity. The following step is the estimation of the mixture factor model with different numbers of components to find the best solution. Further steps are the addiction of covariates to explain class membership and factors scores. The common factor model is a linear regression where the observed variables are regressed on factors. Both items and factors could be regressed on covariates.

Referring to the notation used in paragraph 2.1, the Mixture Factor Model used in our application could be specified in the following way:

$$\boldsymbol{\upsilon} = \boldsymbol{\mu} + \sum_{m=1}^{M} \boldsymbol{\lambda}_m \boldsymbol{\eta}_m^{(2)} \tag{2.9}$$

$$\boldsymbol{\eta}_m^{(2)} = \sum_{k=1}^{K} \boldsymbol{\beta}_{km} \boldsymbol{c}_k^{(2)} + \boldsymbol{\epsilon}_m \tag{2.10}$$

$$\boldsymbol{\epsilon}_m \sim MN(0, \boldsymbol{\Sigma}) \tag{2.11}$$

$$\boldsymbol{c}_{kj}^{(2)} = \begin{cases} 1 & \text{if group j belongs to k-}th\text{ latent class at individual level} \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

As it is possible to see the only quantity that is class specific is the factor score mean in equation 2.10. In equation 2.9, $\boldsymbol{\mu}$ represents item mean intercept and $\boldsymbol{\lambda}_m$ the factor score loadings. The distribution form for the mixture component at group level is specified in equation 2.13.

$$\pi_k = P\left(c_{kj}^{(2)} = 1\right) = \frac{exp(\gamma_k)}{\sum_{t=1}^{K} exp(\gamma_t)} \tag{2.13}$$

In our application we added lower level covariates on the factor score, to describe the Home context and its effect on children motivation. As it is possible to see from equation 2.9 to 2.13, many of the quantities involved are not class specific. Some extensions to this model could specified below.

$$\boldsymbol{\upsilon} = \boldsymbol{\mu_k} + \sum_{m=1}^{M} \boldsymbol{\lambda}_m \boldsymbol{\eta}_m^{(2)} \tag{2.14}$$

$$\boldsymbol{\eta}_m^{(2)} = \sum_{k=1}^{K} \boldsymbol{\beta}_{km} \boldsymbol{c}_k^{(2)} + \boldsymbol{\theta}_m^{(2)} \boldsymbol{z}_m^{(2)} + \boldsymbol{\epsilon}_m \tag{2.15}$$

$$\boldsymbol{\epsilon}_m \sim MN(0, \boldsymbol{\Sigma}_k) \tag{2.16}$$

$$\boldsymbol{c}_{kj}^{(2)} = \begin{cases} 1 & \text{if group j belongs to k-}th\text{ latent class at individual level} \\ 0 & \text{otherwise} \end{cases} \tag{2.17}$$

$$\pi_k = P\left(c_{kj}^{(2)} = 1\right) = \frac{exp(\gamma_k + \delta_{mk}^{(2)} x_m^{(2)})}{\sum_{t=1}^{K} exp(\gamma_t + \delta_{mt}^{(2)} x_m^{(2)})} \tag{2.18}$$

In equation 2.14, the item intercepts ($\boldsymbol{\mu}_k$) are class specific and this means that the mixture component has an influence on the way of responding of children. In equation 2.15 and 2.18, $\boldsymbol{z}_m^{(2)}$ and $x_m^{(2)}$ represents respectively the covariates affecting the factor score means and the class membership. In equation 2.27, the index k in $\boldsymbol{\Sigma_k}$ model the effect of the mixture component on the variance/covariance structure of latent factor score.

A graphic representation of the different models is reported in figure 2.1, from Lubke and Muthén (2005). The common factor model is a sub-model of the general factor mixture model, in which the number of classes is fixed to one. The classic latent class model and the latent profile model could be derived from the factor mixture model fixing the factor variance to zero.

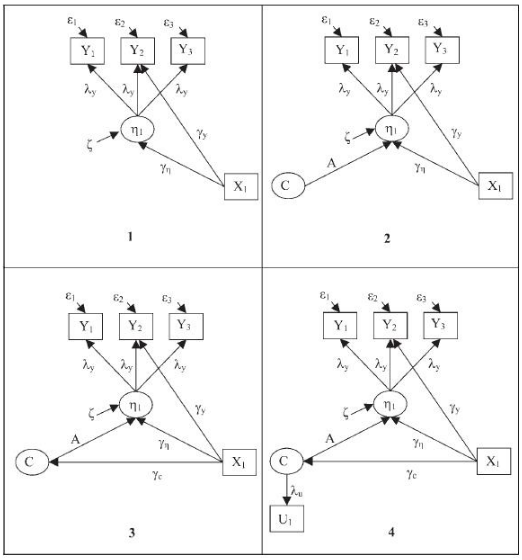

Figure 2.1: From Common Factor Models to Factor Mixture Models

2.3 The Multilevel Mixture Factor Model

In this paragraph, we present a different approach to the multilevel data, where the between group heterogeneity is not modeled in the classic way using continuous factors or random effects. In this model, group differences are modeled hypothesizing the existence of K latent classes or mixture components. This is called the Multilevel Mixture Factor Model, where the unknown source of population heterogeneity is hypothesized to be, not at individual level (as in the mixture factor model), but at group level. Some authors proposed different approaches to deal with hierarchical data as the mixture models for factor analysis (FA) or item response theory (IRT)(McLahan and Peel (2000), Lubke and Muthén (2005), Yung (1997)). These models try to relax the key assumption of the common factor analysis, where the observations came from a single homogeneous population. Both the finite mixture models and the latent class analysis try to verify this assumption. Discovering the existence of unknown subpopulations that generated the sample is one of the goals of these models.

Multilevel mixture factor models and mixture of FA or IRT described in section 2.2 have some similarities and differences. The key difference concerns with the position of the mixture distribution (group level instead of individual level). A fundamental assumption of the factor analysis is the fact that the sample has been generated by a single population. These models deal with situations where the population is composed of different subpopulations. In our application, the mixture factor model has been implemented. It has been assumed that the individual heterogeneity source was at student level (lower level), with the Assumption that students differences do not depend on teachers. In multilevel mixture factor model the source of individual heterogeneity depends on teacher level (higher level). The explicit idea of this model is that the students differences depends on the environment and the teachers. Multilevel mixture factor model assumes that the group level observations could be classified into homogeneous classes. This specification is useful if the aim of the research is finding a classification of the subjects in groups. Using a discrete specification for the mixture component at the higher level is an advantage to approximate the continuous higher level variation without introducing strong assumptions on the distributions of the higher level latent variables, making in this way the computation easier.

We indicate with K the total number of latent classes or the number of the finite mixture components, and with k one of the latent classes ($k = 1, \ldots, K$). The observed variables (items) are denoted by y_{hij}, where h identifies items *($h = 1, \ldots, H$)*, i subjects *($i = 1, \ldots, n_j$)* and j groups *($j = 1, \ldots, J$)*. The number of observations in each group is n_j and it could change between groups. The total number of subjects is $N = \sum_{j=1}^{J} n_j$. This model could be used both for the two level regression model for multivariate responses (individuals i represent lower level units, while groups j represent higher level units) and for the three level model for univariate responses (item h is the level one, individual i is the level two and group j is level three of the hierarchy).

We adopted the same notation reported in the book of Skrondal and Rabe-Hesketh (2004). The subscripts *2, 3* identify the latent variables at individual and group level. The linear predictor for the item y_{hij} is identified by υ_{hij}. $\eta^{(2)}_{hij}$ for *($h = 1, \ldots, M$)* represents the latent variable at individual level, following indicated with $\boldsymbol{\eta^{(2)}_{ij}}$. $\boldsymbol{\eta^{(3)}_{j}} = \left(\eta^{(3)}_{1j}, \ldots, \eta^{(3)}_{Kj}\right)$ indicates the latent variable at group level.

Reusing the previous notation, the group membership has been defined in the following way

$$\eta^{(3)}_{kj} = \begin{cases} 1 & \text{if group j belongs to latent class k} \\ 0 & \text{otherwise} \end{cases} \tag{2.19}$$

In this specification, the linear predictor for y_{hij} is identified by υ_{hij}. Latent variables at individual and group level are denoted by $\boldsymbol{\eta}^{(2)}_{ij}$ and $\boldsymbol{\eta}^{(3)}_{i}$. The following equations describe the two level mixture factor models used in our application:

$$\boldsymbol{\upsilon}_{ij} = \boldsymbol{\mu}_j + \sum_{m=1}^{M} \boldsymbol{\lambda}_m \eta_{ijm}^{(2)} \tag{2.20}$$

$$\eta_{ijm}^{(2)} = \sum_{k=1}^{K} \boldsymbol{\beta}_{km} \eta_{kj}^{(3)} + \epsilon_{ijm} \tag{2.21}$$

$$g\left(E(\boldsymbol{y}_{ij}|\boldsymbol{\eta}_{ij}^{(2)}, \boldsymbol{\eta}_{j}^{(3)})\right) = \boldsymbol{\upsilon}_{ij} \tag{2.22}$$

Equation 2.22 shows the connection of the conditional expectation of the response y_{hij} (given the latent variables at different levels of the hierarchy) with the linear predictor υ_{hij}.

In this specification, the factor score means are class specific (equation 2.21), and $\boldsymbol{\mu}_j$ represents the item mean intercepts. Equation 2.23 describes the multinomial distribution for the mixture component.

$$\pi_k = P\left(\eta_{kj}^{(3)} = 1\right) = \frac{\exp(\gamma_k)}{\sum_{t=1}^{K} \exp(\gamma_t)} \tag{2.23}$$

The sum over the K components of the π_k in equation 2.23, is equal to one, while γ_k represents the class intercept. Some restrictions on the factor score means (the sum over the K mixture components is equal to zero) and on factor loadings (the first factor loading for each dimension is fixed to one) have been imposed. A multivariate normal distribution is typically assumed for factor scores (see equation 2.24).

$$\boldsymbol{\epsilon}_m \sim MN(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.24}$$

These equations show the inclusion of the multilevel data structure inside the model definition. The sum in equation 2.21 and 2.24 are over the K latent indicators. The mixture component at group level classifies the second level observations according to the latent structure measured at individual level. The added value of the *MMFA* is the contemporaneous inclusion of the factor analysis to classify items and the mixture component (at higher level) to classify the group level observations.

How is it possible to see, equation 2.20 is similar to 2.1. The differences concern with the quantities involved in the estimation (as the linear predictor, items intercept and factors scores), because they depend on the index j, representing the groups. The other modification is related to the notation, because it is necessary to distinguish the quantities related to the level two from the ones related to the level three.

The quantities in equation 2.21 are: the mean intercept for item h ($\boldsymbol{\mu}_j$), the factor loadings ($\boldsymbol{\lambda}_m$) and the error indicator at level three ($\boldsymbol{\epsilon}_m$). In equation 2.21, it is possible to see the effect of the third level on the factor scores mean at the second level. $\boldsymbol{\beta}_{km}$ represents the latent variable mean at individual level and ϵ_{ijm} is the residual term error.

Some extensions of the model (presented in equations 2.20 to 2.24) have been proposed. The addition of predictors to explain the factor mean or class membership, and the effect of the mixture component on the item responses or on the latent variable covariance matrix, are some possible examples. These kind of extensions have been reported below.

$$\boldsymbol{\upsilon}_{ij} = \boldsymbol{\mu}_k + \sum_{m=1}^{M} \boldsymbol{\lambda}_m \eta_m^{(2)} \tag{2.25}$$

$$\eta_{ijm}^{(2)} = \sum_{k=1}^{K} \boldsymbol{\beta}_{km} \eta_{kj}^{(3)} + \theta_m^{(2)} z_{ij}^{(2)} + \epsilon_{ijm} \tag{2.26}$$

$$\boldsymbol{\epsilon}_m \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_k) \tag{2.27}$$

$$\pi_k = P\left(\eta_{kj}^{(3)} = 1\right) = \frac{\exp(\gamma_k + \theta_k^{(3)} z_j^{(3)})}{\sum_{t=1}^{K} \exp(\gamma_t + \theta_t^{(3)} z_j^{(3)})} \tag{2.28}$$

The index k in $\boldsymbol{\mu}_k$ describes the influence of the mixture component on the item mean. This effect describes the fact that the group level units (in our application the teachers) have an influence on the way of responding of the children. In equation 2.26 and 2.28, $z_{ij}^{(2)}$ and $z_j^{(3)}$ are the individual and group level covariates, affecting the factor score means and class membership. In equation 2.27, $\boldsymbol{\Sigma}_{\boldsymbol{k}}$ represents the class specific latent factor covariance matrix.

In figure 2.2 a two level factor mixture model has been shown. It is possible to see the two level structure, where index i identifies individual level and index j the group one. The latent variables are present at both the level of the hierarchy and in this specification the latent variable at individual level is connected directly to items, while the latent variable at higher level is connected both to the latent variables at lower level and to the items. In our application the latent variable at higher level will affect only the lower level latent variables. In figure 2.3 this situation has been

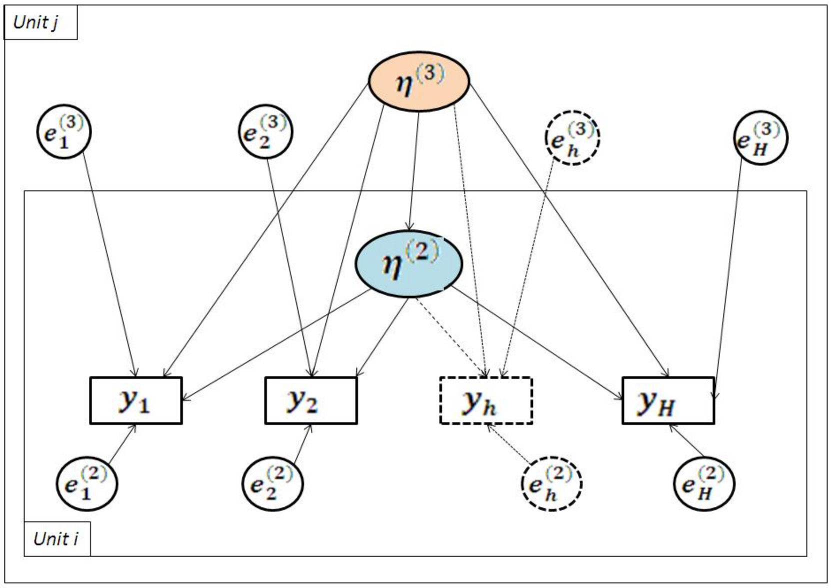

Figure 2.2: Two level mixture factor model

reported. It is possible to see the effect of the group level covariates on the class membership. In our application these models have been estimated to show the different approaches between the mixture factor model and the multilevel factor model. We want to show the importance of introducing the covariates to explain the class membership or the relations between the latent factor scores and the lower level covariates.

2.4 Parameter Estimation

As pointed by Skrondal and Rabe-Hesketh (2004) different estimation methods have been proposed in the last years for the latent variables framework. These have been classified into three different classes on the base of the quantities that are fixed or random. The possibilities are: random latent variables and fixed parameters, fixed latent variables and parameters, random latent variables and parameters. Further details on these estimation methods are

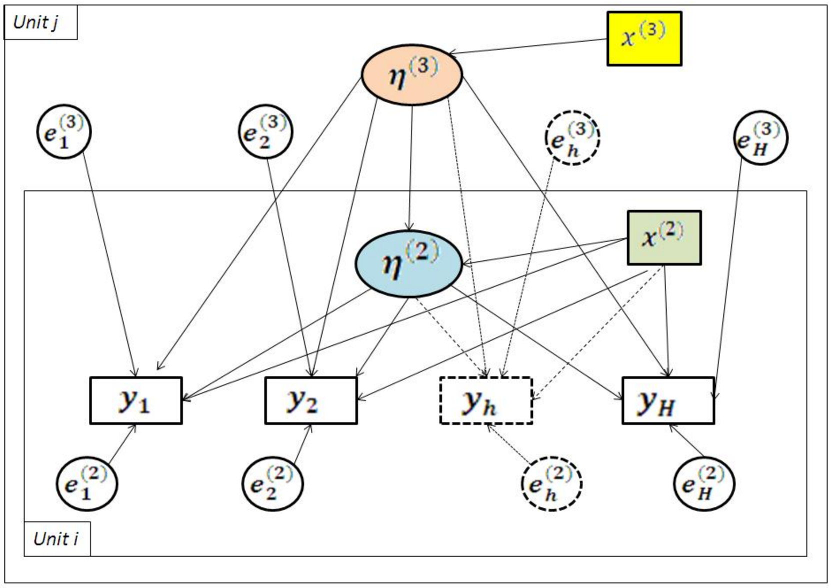

Figure 2.3: Two level mixture factor model with covariates

reported in the book of Skrondal and Rabe-Hesketh (2004). The parameters of the mixture factor models and the multilevel mixture factor models could be estimated via maximum likelihood (ML). Considering the assumptions for the error distribution (for items and latent variables), it is possible to derive the density functions for the items vector of an independent subject. The model is a two level mixture models. At lower level (students), the existence of a set of continuous latent variables has been hypothesized. At higher level (teachers), a discrete distribution for the latent variables has been assumed. The maximum likelihood estimation involves the maximization of the marginal likelihood function:

$$L = \prod_{j=1}^{J} f(y_j) \tag{2.29}$$

In our application, we deal with $\boldsymbol{\eta}_{ij}^{(2)}$ and $\boldsymbol{\eta}_{j}^{(3)}$ that are respectively the continuous latent variables at lower level, and the discrete latent variables at higher level. In Varriale and Vermunt (2009) and Vermunt (2008), the estimation details are reported. Following, in the described situation, the latent variables are always considered as continuous. In model estimation one of the fundamental assumptions is groups independence, given the group membership (for the discrete higher level latent variables) or the group level random coefficient (for the continuous higher level latent variables). For this reason, it is possible to write

$$f\left(\boldsymbol{y_{ij}}|\boldsymbol{\eta_j^{(3)}}\right)=\int_{\eta_{ij}^{(2)}}\left[\prod_{h=1}^{H} f\left(y_{hij}|\boldsymbol{\eta_{ij}^{(2)}},\boldsymbol{\eta_j^{(3)}}\right)\right] f\left(\boldsymbol{\eta_{ij}^{(2)}}|\boldsymbol{\eta_j^{(3)}}\right) d\boldsymbol{\eta_{ij}^{(2)}} \quad (2.30)$$

The possibility of writing the internal terms of square brackets depends on the local independence (see equation 2.3). The term $f\left(\boldsymbol{\eta_{ij}^{(2)}}|\boldsymbol{\eta_j^{(3)}}\right)$ represents the conditional distribution of the item h given the latent variables $\boldsymbol{\eta_{ij}^{(2)}}$ and $\boldsymbol{\eta_j^{(3)}}$.

In our application, the integration over $\eta_j^{(3)}$ has been replaced with the summation over the K latent classes. The quantities involved in the estimation procedures are:

$$f(\boldsymbol{y}_j)=\sum_{k=1}^{K}\left[\prod_{i=1}^{n_j} f\left(\boldsymbol{y}_{ij}|\boldsymbol{\eta_j^{(3)}}=1\right)\right]\pi_k \quad (2.31)$$

$$f\left(\boldsymbol{y}_{ij}|\boldsymbol{\eta_j^{(3)}}=1\right)=\int_{\eta_{ij}^{(2)}}\left[\prod_{h=1}^{H} f\left(y_{hij}|\boldsymbol{\eta_{ij}^{(2)}},\boldsymbol{\eta_j^{(3)}}=1\right)\right] f\left(\boldsymbol{\eta_{ij}^{(2)}}|\boldsymbol{\eta_j^{(3)}}=1\right)d\boldsymbol{\eta_{ij}^{(2)}} \quad (2.32)$$

The individual level observations are independent each other, given the latent variables at both levels $\boldsymbol{\eta_{ij}^{(2)}}$ and $\boldsymbol{\eta_j^{(3)}}$. To find the parameter estimates it is necessary to solve the integrals involved in the likelihood function. It is necessary to maximize the likelihood function. Many different approaches have been proposed to evaluate the integrals involved in the equations 2.31 and 2.32, such as the Laplace integration, the numerical integration with adaptive and non adaptive quadrature or the Monte Carlo integration. In Skrondal and Rabe-Hesketh (2004), it is reported the description of these different estimation methods. The Gauss Hermite quadrature approximates the multivariate normal mixing distribution using a limited number of discrete points. Summarizing the information provided in equations 2.31,2.32 and using the Gauss Hermite quadrature numerical integration methods, it is possible to write:

$$f_k(\boldsymbol{y}_k)=\sum_{s=1}^{T^{(3)}} P_k\left(y_k|\boldsymbol{\eta_s^{(3)}}\right)\pi\left(\boldsymbol{\eta_s^{(3)}}\right)=\sum_{s=1}^{T^{(3)}}\prod_{j=1}^{n_k} P_{jk}\left(y_{jk}|\boldsymbol{\eta_s^{(3)}}\right)\pi\left(\boldsymbol{\eta_s^{(3)}}\right)= \quad (2.33)$$

$$\sum_{s=1}^{T^{(3)}} \left[\prod_{j=1}^{n_k} \sum_{r=1}^{T^{(2)}} \left\{ \prod_{i=1}^{n_{jk}} P_{ijk} \left(y_{ijk} | \boldsymbol{\eta_r^{(2)}}, \boldsymbol{\eta_s^{(3)}} \right) \right\} \pi \left(\boldsymbol{\eta_r^{(2)}} \right) \right] \pi \left(\boldsymbol{\eta_s^{(3)}} \right)$$

The $\eta_r^{(2)}$ and $\eta_s^{(3)}$ are the quadrature nodes and $\pi\left(\boldsymbol{\eta_r^{(2)}}\right)$ and $\pi\left(\boldsymbol{\eta_s^{(3)}}\right)$ the quadrature weights. In the normal distribution, they represent the multivariate normal densities. In presence of missing values, the most used algorithms for maximizing likelihood function are the Expectation-Maximization (EM) (Dempster *et al.* (1977)), the Newton-Raphson (NR), the Fisher Scoring and the Quasi-Newton Methods. Softwares for the latent class model estimation are: GLLAMM (Rabe-Hesketh *et al.* (2004b)), MPlus (Muthén and Muthén (2007)) and Latent GOLD (Vermunt and Magidson (2005b, 2007)). Some differences between these programs concern with the estimation methods. GLLAMM solves integrals using both the adaptive or non-adaptive Gauss-Hermite quadrature. The Newton-Raphson method is used to maximize the marginal likelihood function. Mplus uses the rectangular, or the Gauss-Hermite, or the Monte Carlo integration to solve the integrals. For the maximization of the likelihood function, Mplus combines the EM and quasi-Newton methods. Gauss-Hermite integration is used in Latent GOLD to solve integrals, while to maximize the likelihood function a combination of the EM and NR algorithms is used.

In the parameters estimation it is necessary to implement a method to find factor scores or latent class membership. To compute these quantities using the Bayes rule, we must refer to the latent variable posterior distributions given the observed data. The situation deals in the same way latent variables as discrete or continuous, and the distribution given to observed data have been considered. In equation 2.34, it is reported the higher level latent variable distribution given the observed data for the group j:

$$f(\boldsymbol{\eta_s^{(3)}}|\boldsymbol{y}_j) = \frac{f(y_j, \boldsymbol{\eta_s^{(3)}})}{f(\boldsymbol{y}_j)} = \frac{f(\boldsymbol{y}_j|\boldsymbol{\eta_s^{(3)}})f(\boldsymbol{\eta_s^{(3)}})}{\int_{\eta_s^{(3)}} f(\boldsymbol{y}_s|\boldsymbol{\eta_j^{(3)}})f(\boldsymbol{\eta_s^{(3)}})} \tag{2.34}$$

In equation 2.34, the $f(\boldsymbol{y}_j|\boldsymbol{\eta_j^{(3)}})$ and $f(\boldsymbol{\eta_j^{(3)}})$ have been estimated before.

Considering the lower level latent variable distributions given the observed data for the subject i in group j, we have:

$$f(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{y}_{ij}, \boldsymbol{\eta}_s^{(3)}) = \frac{f(\boldsymbol{y}_{ij}, \boldsymbol{\eta}_r^{(2)}|\boldsymbol{\eta}_s^{(3)})}{f(\boldsymbol{y}_{ij})} = \frac{f(\boldsymbol{y}_{ij}|\boldsymbol{\eta}_r^{(2)}, \boldsymbol{\eta}_s^{(3)})f(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{\eta}_s^{(3)})}{\int_{\eta_r^{(2)}} f(\boldsymbol{y}_{ij}|\boldsymbol{\eta}_r^{(2)}, \boldsymbol{\eta}_s^{(3)})f(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{\eta}_s^{(3)})} \tag{2.35}$$

The *Empirical Bayes* (EB) prediction is one of the methods used for factor scores. In equation 2.36, it has been showed a method to obtain the mean value of the posterior distribution.

$$\boldsymbol{\eta}_j^{(3)EB} = E\left(\boldsymbol{\eta}_j^{(3)}|\boldsymbol{y}_j\right) = \int_{\boldsymbol{\eta}_j^{(3)}} f(\boldsymbol{\eta}_j^{(3)}|\boldsymbol{y}_j) \tag{2.36}$$

The same reasoning could be done for the latent variables at level two. More details could be found in Skrondal and Rabe-Hesketh (2004). In our application a discrete higher level latent variable has been hypothesized. The posterior probability is defined in the following way:

$$P(\boldsymbol{\eta}_j^{(3)} = 1|y_k) = \frac{f(\boldsymbol{y}_j, \boldsymbol{\eta}_s^{(3)} = 1)}{f(\boldsymbol{\eta}_s^{(3)})} = \frac{f(\boldsymbol{y}_j|\boldsymbol{\eta}_s^{(3)}\pi_k)}{\sum_{k=1}^{K} f(\boldsymbol{y}_j|\boldsymbol{\eta}_s^{(3)}\pi_k)} \tag{2.37}$$

A subject is assigned to the group j if the $P(\boldsymbol{\eta}_j^{(3)} = 1|\boldsymbol{y}_k)$ reaches its maximum.

In the next sub-paragraphs some methods for the likelihood maximization have been presented. The methods are: the Expectation-Maximization (EM) algorithm, the Fisher scoring and the Newton - Raphson algorithms. A sub-section has been dedicated to the Upward-Downward EM Algorithm, a variant of the EM algorithm used in Latent Gold.

2.4.1 The Newton-Raphson Algorithm

The *Newton-Raphson Algorithm* is a method used to find the maximum likelihood estimates. Suppose that it is possible to differentiate twice the likelihood function respect to the unknown parameters θ. The matrix of these second order partial derivatives is individuated by Θ. The updating estimation procedure of the *Newton-Raphson algorithm* is based on a previous solution and an iterative process. To individuate the next parameter estimates, it is necessary to compute the following step:

$$\theta^{m+1} = \theta^m - \left(\frac{\partial L^2}{\partial\theta\partial\theta^T}\right)^{-1} \frac{\partial L}{\partial\theta} \tag{2.38}$$

θ^m represents the previous estimate.

$$\frac{\partial L^2}{\partial\theta\partial\theta^T}$$

The quantity above reported is the inverse of the matrix of the second order partial derivatives calculated in $\theta = \theta^m$. Under particular regularity conditions, this algorithm converges to the unique maximum of the likelihood function. If some of these conditions are not valid, as the definition of the space parameters that must be opened or the negative definite of the second order partial derivatives matrix, it is not possible to be sure that the found solution corresponds to the global maximum.

2.4.2 The Fisher Scoring

The Fisher scoring differs from the Newton-Rapshon algorithm because in the updating estimation procedure, the inverse of the second order partial derivatives matrix is replaced with the Expected Information Matrix:

$$H(\theta) = -E\left(\frac{\partial L^2}{\partial\theta\partial\theta^T}\right) \tag{2.39}$$

The updating procedure could be written in the following way:

$$\theta^{m+1} = \theta^m - (H(\theta^m))^{-1}\left.\frac{\partial L}{\partial\theta}\right|_{\theta=\theta^m} \tag{2.40}$$

These two methods have similar convergence properties. Both the Newton-Raphson and the Fisher scoring algorithms use the second order derivatives of the log-likelihood function in updating procedures. The computation of these derivatives could be analytically difficult and increase the computation time. To solve this problem, some quasi-Newton algorithms have been suggested. One of these has been proposed by Bemdt *et al.* (1974), called the *BHHH* or BH^3 algorithm. Under a correct model specification, the information matrix is equal to the covariance matrix of the gradients:

$$H(\theta)_{BH^3} = E\left(\frac{\partial L}{\partial\theta}\frac{\partial L}{\partial\theta^T}\right) \tag{2.41}$$

The estimation procedure could be written in this way:

$$\theta^{m+1} = \theta^m - (H(\theta)_{BH^3})^{-1}\left.\frac{\partial L}{\partial\theta}\right|_{\theta=\theta^m} \tag{2.42}$$

This algorithm computes only the first derivatives and it represents the added value of this algorithm. Indeed, neither the Hessians nor the Fisher information matrices must be computed. In the book of Skrondal and Rabe-Hesketh (2004) more details on the other estimation procedures could be found.

2.4.3 EM Algorithm

The Expectation-Maximization (EM) algorithm is one of the most used methods for the maximization of the likelihood function. Dempster *et al.* (1977) introduced the EM approach as an iterative procedure based on the maximum likelihood estimation for models in presence of missing data. For a more general view on the EM Algorithm it is possible to consult the book of McLachlan and Krishnan (1997). The general idea is to simplify the estimation procedures. The observed data is augmented with the missing data in order to permit to the estimation procedure to go on in a simpler estimation steps.

In latent variable models, the missing data are the values of the latent variables. Let C be equal to $(\boldsymbol{y};\boldsymbol{\eta})$, that are the complete data. $\boldsymbol{y}$ is the incomplete observed data. $\boldsymbol{\eta}$ are the unobservable or latent data. The complete data log-likelihood, imaging that the latent data were observed, is denoted by $logL^c = logL(\theta|C)$. EM algorithm is characterized by two steps. More details could be found in Skrondal and Rabe-Hesketh (2004):

- **E-step**: the posterior expectation is calculated as follow:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^m) = E_\eta[logL^c|\boldsymbol{y};\boldsymbol{\theta}^m]$$

  It represents the conditional expectation of the complete data log-likelihood with respect to $\boldsymbol{\eta}$, given the incomplete data and the estimates of $\boldsymbol{\theta^m}$ at the previous iteration.

- **M-step**: the likelihood function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^m)$ is maximized with respect to $\boldsymbol{\theta}$, in order to update the estimate of $\boldsymbol{\theta}^{m+1}$

The proceeding way of the EM algorithm is:

1. To impute the missing values using the predicted values,
2. To estimate the parameters treating the imputed values as data,
3. To impute the missing values again using the new estimates as the true parameters,
4. To re-estimate the parameters and so on until convergence.

The advantage of the *EM* algorithm is the easier implementation compared with the other methods. Theoretical advantages include the increment of the likelihood value at each iteration and the certainty of the convergence to a local maximum or saddle point, if the sequence θ^m converges. A disadvantage of the *EM* algorithm is the slowness of the convergence procedure in presence of a large fraction of missing data. Another disadvantage is that the estimated information matrix is not a direct byproduct of maximization, compared with the gradient methods such as Newton-Raphson. A way to solve these problems, is augmenting the EM algorithm with a final Newton-Raphson step after the convergence. This way has been adopted in Latent Gold. This kind of problems arises using *EM* algorithm in the multilevel latent variable framework, and a variant of this algorithm has been proposed.

2.4.4 Upward-Downward Algorithm

The log-likelihood function in the Multilevel Latent Variable Model (MLVM), assumes the following form:

$$\sum_{s=1}^{T^{(3)}}\sum_{r=1}^{T^{(2)}}\sum_{k=1}^{K}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}} P_{jk}(\boldsymbol{\eta}_r^{(2)}, \boldsymbol{\eta}_s^{(3)}|\boldsymbol{y}_k) log f_{ijk}(\boldsymbol{y}_{ijk}|\boldsymbol{\eta}_r^{(2)}, \boldsymbol{\eta}_s^{(3)}) \tag{2.43}$$

In the E step of the EM algorithm, it is necessary to compute the expectation of the complete data log-likelihood. In the E step, we compute the posterior probabilities $P_{jk}(\boldsymbol{\eta}_r^{(2)}, \boldsymbol{\eta}_s^{(3)}|y_k)$. In the M step, the log-likelihood function is maximized after the model parameters updating procedures. In the multilevel latent variable framework, the implementation problems are related to the E step, where the posterior probabilities should been obtained. Further details are reported in Vermunt (2004, 2008). We can solve the computational problems referring to the conditional independence assumption, because it is possible to compute only the n_k marginal posterior probabilities $P_{jk}(\boldsymbol{\eta}_r^{(2)}, \boldsymbol{\eta}_s^{(3)}|\boldsymbol{y}_k)$ without referring to the full posterior distribution. The procedures adopted in this situation are similar to the forward-backward algorithm used for the estimation of hidden Markov models with large numbers of time points (Baum *et al.* (1970), Juang and R. (1991), Fruhwirth-Schatter (2006)). Both Forward/Backward and Upward/Downward algorithm are propagation algorithm.

In the upward-downward algorithm, latent variables are integrated out going from the lower to the higher levels. The marginal posterior probabilities are computed going from the higher to the lower levels. With these procedures, the increment of computer storage and time follows a linear trend instead of exponentially, as in a standard EM algorithm. Following are reported some details to explain how the algorithm, implemented in Latent Gold, works. The marginal posterior probability $P_{jk}(\boldsymbol{\eta}_r^{(2)}, \boldsymbol{\eta}_s^{(3)}|y_k)$ can be decomposed in this way:

$$P_{jk}(\boldsymbol{\eta}_r^{(2)}, \boldsymbol{\eta}_s^{(3)}|\boldsymbol{y}_k) = P_k(\boldsymbol{\eta}_s^{(3)}|\boldsymbol{y}_k) P_{jk}(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{y}_k, \boldsymbol{\eta}_s^{(3)})$$

and we also known that

$$P_{jk}(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{y}_k,\boldsymbol{\eta}_s^{(3)}) = P_{jk}(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{y}_{jk},\boldsymbol{\eta}_s^{(3)})$$

Furthermore $\boldsymbol{\eta}_r^{(2)}$ is independent from the observed responses of the other level *2* units within the same level *3* unit given $\boldsymbol{\eta}_s^{(2)}$ and level *2* observations are mutually independent conditionally to level *3* latent variable (class membership or random effect). So we can decompose :

$$P_{jk}(\boldsymbol{\eta}_r^{(2)},\boldsymbol{\eta}_s^{(3)}|\boldsymbol{y}_k) = P_k(\boldsymbol{\eta}_s^{(3)}|\boldsymbol{y}_k)P_{jk}(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{y}_{jk},\boldsymbol{\eta}_s^{(3)}) \tag{2.44}$$

So we need to compute $P_k(\boldsymbol{\eta}_s^{(3)}|\boldsymbol{y}_k)$ and $P_{jk}(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{y}_{jk},\boldsymbol{\eta}_s^{(3)})$ instead of computing marginal posterior probabilities. This term $P_{jk}(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{y}_{jk},\boldsymbol{\eta}_s^{(3)})$ could be obtained in the following way:

$$P_{jk}(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{y}_{jk},\boldsymbol{\eta}_s^{(3)}) = \frac{P_{jk}(\boldsymbol{y}_{jk},\boldsymbol{\eta}_r^{(2)}|\boldsymbol{\eta}_s^{(3)})}{P_{jk}(\boldsymbol{y}_{jk}|\boldsymbol{\eta}_s^{(3)})}$$

where

$$P_{jk}(\boldsymbol{y}_{jk},\boldsymbol{\eta}_r^{(2)}|\boldsymbol{\eta}_s^{(3)}) = \pi(\boldsymbol{\eta}_r^{(2)})\prod_{i=1}^{n_{jk}} P_{ijk}(\boldsymbol{y}_{ijk}|\boldsymbol{\eta}_r^{(2)},\boldsymbol{\eta}_s^{(3)})$$

$$P_{jk}(\boldsymbol{y}_{jk}|\boldsymbol{\eta}_s^{(3)}) = \sum_{r=1}^{T^{(2)}} P_{jk}(\boldsymbol{y}_{jk},\boldsymbol{\eta}_r^{(2)}|\boldsymbol{\eta}_s^{(3)})$$

while the term $P_{jk}(\boldsymbol{y}_{jk}|\boldsymbol{\eta}_s^{(3)})$ is obtained by

$$P_{jk}(\boldsymbol{y}_{jk}|\boldsymbol{\eta}_s^{(3)}) = \frac{P_k(\boldsymbol{y}_k,\boldsymbol{\eta}_s^{(3)})}{P_k(\boldsymbol{y}_k)} \tag{2.45}$$

in which

$$P_k(\boldsymbol{y}_k,\boldsymbol{\eta}_s^{(3)}) = \pi(\boldsymbol{\eta}_s^{(3)})\prod_{j=1}^{n_k} P_{jk}(\boldsymbol{y}_{jk}|\boldsymbol{\eta}_s^{(3)})$$

$$P_k(y_k) = \sum_{s=1}^{T^{(3)}} P(\boldsymbol{y}_k,|\boldsymbol{\eta}_s^{(3)})$$

The level-2 posterior probabilities $P_{jk}(\boldsymbol{\eta}_r^{(2)}|\boldsymbol{y}_{jk},\boldsymbol{\eta}_s^{(3)})$ are obtained from the level-1 $P_{ijk}(\boldsymbol{y}_{ijk}|\boldsymbol{\eta}_r^{(2)},\boldsymbol{\eta}_s^{(3)})$ and after level-3 posterior probabilities are obtained from the level-2 information $P_{jk}(\boldsymbol{y}_{jk}|\boldsymbol{\eta}_s^{(3)})$. This way of proceeding is the *Upward* step of the algorithm because goes from the lower level to the higher level. In the *Downward* step $P_{jk}(\boldsymbol{\eta}_r^{(2)},\boldsymbol{\eta}_s^{(3)}|\boldsymbol{y}_k)$ are computed.

The maximum likelihood estimation has been described for models with continuous latent variables and numerical integration. In the same way, this algorithm could be extended without problems for models with discrete latent variables. The difference concerns with the quadrature weights because they are not fixed, but contains free parameters to be estimated (Vermunt (2003, 2004)). Using the estimation procedure based on the Newton-like methods, as the EM algorithm, one of the problems is that it does not provide the standard errors of the parameters. It is necessary to estimate asymptotic standard errors through the matrix of the second-order derivatives of the log-likelihood function. The variance-covariance matrix is the inverse of the second-order derivatives. The second derivatives were obtained numerically using analytic first derivatives provided by the EM algorithm. The utility of the information matrix is the possibility to check for identifiability. A sufficient condition for local identification is that the Jacobian matrix is of full column rank (Rothenberg (1971)).

2.5 Model Evaluation

As said in section 2.4 the estimation procedures deal with the maximum likelihood and in latent variable context it is necessary to decide the number of the latent components that we want to insert into the model. Many studies tried to evaluate nested models using the likelihood-ratio tests, that under certain regularity conditions, assume a chi-square distribution. A problem arising with the LRT is the impossibility to use it to compare models with different number of classes, because of the specification of the null hypothesis, where for a smaller number of classes one or more parameter from the alternative hypothesis have been fixed.

Two types of goodness of fit tests have been principally used to solve this problem: Chi-squared type tests and tests based on the empirical distribution function. The Chi-square tests are usually used when the data are grouped into discrete classes and are based on the comparison between the observed and the expected frequencies. While tests based on the empirical distribution function are frequently used with continuous data.

Usually the Chi-squared type goodness-of-fit tests are used to compare models. They are based on the comparison between the observed frequencies f_s and the expected frequencies under the model specified in the null hypothesis $\hat{F}_s$. There are two types of goodness of fit tests that are usually employed: the likelihood ratio chi-squared statistic L^2 and the Pearson chi-squared goodness of fit test statistic X^2, defined in the following way

$$L^2 = 2\sum_s f_s ln\left(\frac{f_s}{\hat{F}_s}\right)$$

and

$$X^2 = \sum_s \frac{\left(f_s - \hat{F}_s\right)^2}{\hat{F}_s}$$

These two statistic tests give a measure of the variables association that the model is unable to explain. Under the null hypothesis the distribution that these two tests follow is a χ^2. The disadvantages of using L^2 and X^2 are related to the applications, because their usefulness decreases. The χ^2 statistic is related to the sample size and in presence of large samples the performances get worse, because it tends to be too much conservative and the value of χ^2 could lead to the rejection of the model even if the differences between the model and the data are not so evident. Another problem arises in presence of sparse data, where many cells are empty. In these situations it is not possible to use the chi-squared distribution to compute the p-values.

Some authors (McLachlan (1987), Nylund *et al.* (2007)) tried to solve this problem using a bootstrap-based testing procedures, but this was not the best solution because these procedures were computationally intensive. The Bootstrap approach empirically estimates the p-value associated with the L^2 statistic by means of a parametric bootstrap, instead of assuming a known distribution.

Another approach is the *information criteria.* As Agresti (2002) pointed, a simple model should be preferred to a more complex one. The aim is to provide a better estimate of the true model considering the parsimony criteria. Information criteria are based on some indexes that incorporate model fit (log likelihood value) and parsimony (number of parameters). This kind of approach is the most used in model selection in the latent variables framework. Typically, the log-likelihood grows, increasing the number of latent classes (complexity of the model), but with these indexes the complexity of the model is taken into account, because they include also the number of the estimated parameters. Usually the Information Criteria (IC) are expressed in the following way:

$$IC = -2lnL + Cr, \tag{2.46}$$

where $-2lnL$ is the part of the index that incorporates the model fit, while C r represents the complexity of the model (r is the number of parameters and C is the penalty for additional parameters). Looking for the best solution is equal to search for the model with the lowest information criterion. The information criteria differs from each other because of the value of C. Many texts on Latent Variables models make reference to the *Bayesian information Criterion* (Schwarz (1978)) to assess the number of classes (Hagenaars and McCutcheon (2002), Magidson and Vermunt (2004)). *BIC* is defined in the following way:

$$BIC = -2lnL + ln(n)r \tag{2.47}$$

where n represent the sample size. In some studies had been showed that *BIC* has good performance, but the number of classes will be underestimate if the structure of the classes is not so clear (Dias (2004) and Nylund *et al.* (2007)).

Another possible index used to determine the number of latent variables is the *Akaike Information Criterion* (AIC, Akaike (1974)). It could be expressed as

$$AIC = -2lnL + 2r. \tag{2.48}$$

It has been showed through simulation studies that the *AIC* tends to overestimate the number of classes (McLahan and Peel (2000), Dias (2004)).

Two versions of adjusted *AIC* has been proposed by Bozdogan: *AIC3* (Bozdogan (1993)) and *CAIC* (Bozdogan (1987)) a consistent version of the *AIC*. *AIC3* and *CAIC* can be expressed, respectively, as:

$$AIC3 = -2lnL + 3r \tag{2.49}$$

and

$$CAIC = -2lnL + (1 + ln(n))r. \tag{2.50}$$

Some authors, as Andrews and Currim (1974) and Dias (2004), have shown through simulation studies that it is better to use the *AIC3* in Latent Class models with categorical response variables.

Lukočienė and Vermunt (2009) analyzed the use of the fit index in presence of hierarchical data structure. It was not clear if for sample size was indicated the number of groups K or the total number of individuals N. As shown in a recent simulation study by Lukočienė *et al.* (ress), in the context of the multilevel mixture models, adopting the number of groups as sample size seems to be a better choice. This evidence is due to the fact that the underestimation of the number of mixture components is much more likely using the *BIC* with the number of groups than with the number of observations. The reason of this choice concerns with the fact that using the fit indexes with the number of observations, it is possible to underestimated the number of mixture components, especially if the separation between components is weak or moderate.

Sometimes, model fit is affected by the incomplete explanation of association. This unexplained association between couples of variables can be detected using the bivariate residuals: "*measures can be interpreted as lower bound estimates for the improvement in fit if corresponding local independence constraints were relaxed*"(Vermunt and Magidson (2005b)). This is a diagnostic statistic corresponding to the Pearson chi-square statistic divided by the degrees of freedom.

The indexes calculated in this table are based on the computation of the observed frequencies in a two-way table of a pair of variables using the expected frequencies estimated under the corresponding model. Values larger than *3.84* suggest that the association is not well explained by the model. The presence of residual associations even after inserting latent variables in the model, may be solved with some different methods. Increasing the number of latent classes, deleting one or more than one redundant indicators, or adding a direct effect between these items are three examples of these methods. Many empirical studies have been conducted to investigate and to find the best index to assess the model fit. The same conclusions were reached and some indexes could have a better performance under certain condition and worse in others. It depends on different factors as the sample size, the level of separation of the classes, the scale type of the response variables, or the model assumptions.

2.5.1 Posterior Analysis

In the latent variables modeling, there are two important goals: scoring and classification. Some typical situations are: assessing ability scores in IRT models or medical diagnosis in latent class models. In our application the purpose is scoring students in function of their reading motivation and classify teachers in function of the class motivation. The posterior analysis is based on the frequentist methods, where the estimated structural parameters are considered as known. Using the Bayesian approach, both the latent variables and the parameters are considered as latent variables. Any difference is based on the latent scoring or classification and the parameter estimation. Indifferently of the hierarchical level, latent variables conditional density given the manifest variables, could be expressed in the following way:

$$h(\eta|y) = \frac{h(\eta)g(y|\eta)}{f(y)} \tag{2.51}$$

With this specification it is possible to represent a units profile finding a location for subjects belonging to the latent space (assessing factor scores) or classifying units in different classes. If a unit has the same response pattern should belong to the same latent class or report the same factor scores. In presence of continuous latent variables, the goal consists of assessing factor scores to each unit. In the categorical context the aim is to classify subjects in one of the latent classes.

Using the empirical Bayesian approach, the conditional posterior distribution of latent variables given the observed variables assumes the following representation:

$$f(\eta|y,\theta) = \frac{f(y,\eta|\hat{\theta})}{f(y,\hat{\theta})} \tag{2.52}$$

where the involved quantities are: the estimated vector of parameters $\hat{\theta}$, the observed variables distribution $f(y,\hat{\theta})$ and the joint distribution of the observed and latent variables $f(y,\eta|\hat{\theta})$. Both the latent and the observed variables are considered random variables. In a complete Bayesian specification we assume a "'*a priori*"' distribution for the parameters θ.

The previous specification (2.52) could be extended inserting also the conditional distribution for the latent variables η:

$$f(\eta|y,\theta) = \frac{f(y,|\eta,\hat{\theta})f(\eta|\hat{\theta})}{\int f(y,|\eta,\hat{\theta})f(\eta|\hat{\theta})}$$

Often the calculus of the posterior distributions is not so simple because a closed form does not always exists and for this reason the numerical integration methods are required.

In the empirical Bayesian approach, conditional posterior distributions, given the parameters, are used to make inference on the latent variables. The most famous approaches used to estimate the factor scores or the latent class membership for each subject, are based on the prediction using the Empirical Bayes o *a posteriori* (EB) and Empirical Bayes Modal *modal a posteriori* (EBM). The first one is used for the subject scoring, while the second for the classification.

With regards to the *EB*, the predictors are the mean of the posterior empirical Bayesian latent variables distribution reported in equation (2.52):

$$\eta^{EB} = E\left[\eta|y,\hat{\theta}\right]$$

and under the normal distributional assumption of the latent variables η, the η^{EB} is the Best Linear Unbiased Predictor (BLUP) (Skrondal and Rabe-Hesketh (2004)). For the prediction based on the EBM, the posterior modal is used:

$$\eta^{EBM} = \begin{cases} 1 & \text{if } \max_\eta \eta|y, \hat{\theta} \\ 0 & \text{otherwise} \end{cases}$$

To solve this equation, numerical integration is not requested. As pointed from Skrondal and Rabe-Hesketh (2004) the expected misclassification rate is minimized through these methods. In a common factor models context, the results are the same both using the Empirical Bayes and the Empirical Bayes Modal.

In latent categorical variable context, it is often used the classification statistics (Vermunt and Magidson (2005a)). This statistics describes the level of separation of the latent classes. This index could be expressed in the following way:

$$E = \frac{\sum_{i=1}^{I} \left[1 - max f(\eta|y, \hat{\theta})\right]}{N}$$

where $f(\eta|y, \hat{\theta})$ is the empirical Bayesian posterior distribution and i represents the different response pattern.

Another tool used to measure the classification, is the *Classification Table* (Vermunt and Magidson (2005b)). The entries, called *(x,x')*, are the sum of class x posterior membership probabilities for the cases allocated to modal class *x'*. In the principal diagonal there are the elements classified in the right way, while the elements out of the principal diagonal are the number of misclassification units. Larger is the misclassification, less the class *(x, x')* are separated.

Chapter 3

The mixture factor model

The application described below has a multiple perspective. One of the most important is to provide a widen description of the students reading motivation. The data used in this work has a two level structure. We have items collected for the 4th grade students. The two level of the hierarchy are the student/family (called first or individual level) and the teacher/classes (second or group level). In the Italian system for the primary school, to one class corresponds one teacher. The different aspects examined in this application are:

- To discover the factor structure underlying students reading motivation,
- To describe the reading context,
- To score students according to the reading motivation,
- To classify teachers according to the student reading motivation,
- To evaluate the impact of covariates on motivation at individual (student/family) level in order to describe the effect of the home context,
- To evaluate the effect of the higher level (teacher) covariates on class membership in order to describe the way of teaching,

This chapter has been dedicated to the description of the results obtained with the implementation of both the factor analysis and the mixture factor models described in sections 2.1 and 2.2. With these models, attention has been focused on the individual level ignoring the hierarchical data structure. Section 3.1 reports an introduction to the reading context and the results obtained with the exploratory factor analysis. This model are useful to describe the motivation structure and to introduce the way of dealing with the latent structure. Section 3.2 outline the contributes of the past twenty years of literature on motivation. The environmental descriptions of the influences on student motivation as the home and class context are deeply described. In section 3.3, the results of the mixture factor model are reported. The subparagraph 3.3.1 is dedicated to the results of the mixture factor model without the introduction of the individual covariates. The purpose of this particular specification is the classification of students according to the latent structure. In paragraph 3.3.2, the individual covariates have been added to describe the effects of home environment on motivation. With the mixture factor model we want to model the heterogeneity at individual level using a categorical latent variable, and find unobserved groups of students that show a similar behavior.

3.1 Students Reading Motivation

In the past twenty years, many studies have been conducted to describe both the class situation and the environmental influences (family or school) on the student reading motivation. Several articles have been published in recent years on this topic. In most of them reading motivation was often considered as a multidimensional construct. Guthrie and Wigfield (2000) concentrated their efforts to study the reading development and to describe which factors influence motivation. This is one of most important dimensions in the analysis of the reading development. Reading motivation is influenced by different aspects as the home environment, parental attitude, teacher involvement, teacher strategy of rewards and prices and evaluation.

As said before, the educational and psychological approaches suggest to consider the reading motivation as a multidimensional construct. For this reason, different ways to measure reading motivation have been hypothesized. Following a brief summary of the several approaches proposed has been reported.

Gambrell *et al.* (1995) developed a method to measure the process of reading motivation of children and to describe the personal/background factors that affect motivation. They constructed a questionnaire called the "'*Motivation to Read Profile*"' (MRP) that deals with three different underlying dimensions describing reading motivation: self-concept as a reader, the value of reading and the reasons for reading.

Guthrie and Wigfield (2000) proposed another approach: the "'*Motivations for Reading Questionnaire* "'(MRQ). It includes *82* items measuring *11* different underlying dimensions. The final version of the *MRQ* contains *54* items measuring the following dimensions: efficacy (belief that one can be successful at reading), challenge (willingness to take on difficult reading material), work (avoidance desire to avoid reading activity), curiosity (desire to read topics of interest), involvement (enjoyment received from reading), importance (value placed on reading), recognition (pleasure of receiving a tangible form of recognition for success in reading), grades (desire for positive school evaluations by teacher), competition (desire to outperform others in reading), social (sharing meaning gained from reading with others) and compliance (reading to meet others expectations). This is a very all-encompassing proposal, because it takes into account several dimensions that could be found in the motivation developing process.

Saracho and Dayton (1989) proposed the "'*Motivation for Reading Scale*"' (MRS). In this approach, the authors hypothesized that four dimensions were connected to the reading motivation. This specification is similar to the one proposed by Gambrell *et al.* (1995), because the four dimensions are: the enjoyment of reading (e.g., likes to be read to), the value of reading (e.g., people can learn from books), the self concepts as a reader (e.g., thinks he or she is a good reader) and the interest in library-related activities (e.g., likes to take book from library).

As it is possible to see some dimensions are recurrent in the approaches above presented. In our proposal, this theoretical common point has been taken into account. Our attention has been focused on a short modification of the *Motivation for Reading Scale*. The *P.I.R.L.S.* survey has not been thought to study and describe the reading motivation of the students. The principal aims were the evaluation of the reading performance and of the home and class influences on the students achievement. We decide to utilize the items collected in this survey to study the children reading motivation and the influences of the home and class environment. The results obtained could be influenced by the nature of the data. We modified the structure proposed by Saracho and Dayton (1989). The results obtained in the exploratory factor analysis suggest to considerer only three latent factors instead of four as hypothesized in the *Motivation for Reading Scale*. We restricted the reading motivation structure to the first three dimensions: the enjoyment of reading, value of reading, and self concepts as a reader.

3.1.1 The Factor Analysis results

In this section the results obtained with the implementation of both the exploratory and confirmatory factor analysis are presented. We selected three dimensions to measure the student reading motivation. These factors have been treated as continuous latent factors. Six items were initially selected to measure enjoyment of reading (ER): read only if have to, like talk about

Items	Factor 1	Factor 2	Factor 3	Factor 4
I Enjoy reading	**,79**	,24	-,18	,10
Reading is boring	**-,78**	-,11	,19	,010
Book as a present	**,64**	,26	-,01	,12
Well for future	**,53**	,09	,08	-,02
Read only if have to	**-,42**	,11	**,33**	-,01
Talk with friends	,01	**,67**	-,03	,08
Talk with family	,07	**,67**	-,06	-,13
Like talk about books	**,34**	**,57**	-,02	,16
Reading for info	,10	**,54**	-,05	,03
Read Aloud	,14	**,52**	,04	**-,38**
Reading Books	,15	**,49**	-,07	,23
Read slower than others	-,06	-,01	**,80**	,07
Not as well as other	-,04	-,09	**,80**	-,02
Reading is easy	,14	,14	**-,52**	,31
Understand	,10	,09	-,04	**,84**

Table 3.1: Exploratory Analysis - Rotated Factor Loadings

books, book as a present, reading is boring, well for future, I enjoy reading. For each of these items, students have to express a judgment from a four level scale (from agree a lot to disagree a lot). The enjoyment of reading dimension measures the students feelings towards reading, the pleasure of reading books, if they see reading as a way to interact with people and the utility for their own life.

For the value of reading (VR) dimension, four items have been selected: to read aloud, to talk with friends, to talk with family and to read for info. This factor measures the value of reading for the students, using items that investigate the outside school activities related to reading. These items describe the importance of reading in the student social life and the fact that it is a way to interact with the other children.

Four items have been selected to measure the self concept as a reader (SCR). These items were: reading is easy, reading not as well as the other students, understand everything and to read slower than others. This factor measures the perceived reading ability compared to the other classmates.

We firstly run an exploratory factor analysis to describe the relations between items and factors. The results are reported in table 3.1. Some items are not highly related with the factors and this makes the interpretation really difficult. We decide to restrict the factorial structure. The items that have been inserted in the model are:

- What do you think about reading? Tell how much you agree with each of these statements? *(Agree a lot / Agree a little /Disagree a little / Disagree a lot)*

  - I would be happy if someone gave me a book as a present,

  - I think reading is boring,

  - I need to read well for my future,

  - I enjoy reading.

- How often do you do these things outside of school? *(Every day or almost every day / Once or twice a week / Once or twice a month /Never or or almost never)*

  - I read aloud to someone at home,

  - I talk with my friends about what I am reading,

  - I talk with my family about what I am reading,

  - I read to find out about things I want to learn.

- How well do you read? Tell how much you agree with each of these statements. *(Agree a lot / Agree a little /Disagree a little / Disagree a lot)*

  - Reading is very easy for me,

  - I do not read as well as other students in my class,

  - I read slower than other students in my class.

After the item selection process, we run an exploratory and confirmatory factor analysis. The rotated factor loadings and the fit indexes are reported in tables 3.2 and 3.3. From table 3.2, it is possible to see the three factor structure. The items are highly correlated with the three underlying dimensions and the factor loadings over *0.4* have been highlighted. With this table it is possible to highlight the relations between the observed items and the latent variables and the entity of this relation. Some factor loading appears to be negative and it depends on the item orientation. Table 3.3 reports the fit indexes for the confirmatory factor analysis. The values of the *GFI* and the *NNFI* are both over 0.90, while the value of the *RMSEA* is under 0.05. These indexes are the empirical proof of the goodness of model fit, and for these reasons the three factors structure has been confirmed by the preliminary analysis.

Tables 3.4,3.5 and 3.6 are the frequencies tables of the selected items.

Items	Enjoyment of Reading	Value of Reading	Self Concept as Reader
I Enjoy reading	**.81**	.18	-.21
Reading is boring	**-.76**	-.07	.20
Book as a present	**.70**	.18	-.05
Well for future	**.54**	.08	.05
Talk with friends	.06	**.67**	-.06
Talk with family	.07	**.74**	-.06
Reading for info	.16	**.56**	-.06
Read Aloud	.13	**.59**	.06
Reading is easy	.22	.07	**-.57**
Not as well as other	-.03	-.06	**.81**
Read slower than others	-.02	.02	**.80**

Table 3.2: Exploratory Analysis - Rotated Factor Loadings

Index	Value
Goodness of Fit Index (GFI)	0.9877
GFI Adjusted for Degrees of Freedom (AGFI)	0.9803
Chi-Square	226.6319
Chi-Square DF	41
Pr $\geq$ Chi-Square	$\leq$ .0001
RMSEA Estimate	0.0349
Bentler's Comparative Fit Index	0.9617
Bentler & Bonett's (1980) Non-normed Index	0.9487
Bentler & Bonett's (1980) NFI	0.9538

Table 3.3: Fit indexes for the Confirmatory Factor Analysis

	Book as a present	Reading is boring	Well for future	I enjoy reading
Agree a lot	56.91	8.39	83.90	61.39
Agree a little	25.68	6.83	11.79	25.37
Disagree a little	9.32	15.62	2.48	7.16
Disagree a lot	8.09	69.16	1.83	6.08

Table 3.4: Frequency table for reading statements

	Read Aloud	Talk with friends	Talk with family	Reading for info
Every day or almost every day	31.36	23.27	37.68	54.98
Once or twice a week	30.34	25.01	26.24	27.88
Once or twice a month	9.48	17.63	13.75	10.69
Never or almost never	28.82	34.09	22.33	6.46

Table 3.5: Frequency table for things out school

	Reading is easy	Not as well as other	Rdg slower than other
Agree a lot	58.64	16.15	11.28
Agree a little	31.44	20.76	15.73
Disagree a little	6.70	23.27	25.16
Disagree a lot	3.22	39.82	47.83

Table 3.6: Frequency table for reading ability

3.2 Reading Context

One of the questions related to the school context to which many researchers tried to answer is about the way in which is possible to recognize high motivated students. The principal characteristics of a highly motivated student are: enthusiasm, interest, involvement and curiosity. These students are fascinated by hard challenge and persist on it. These children like being at school, learning new concepts and to continue education after school. It is clear that the problem is not to recognize highly motivated students, because the characteristics could been easily recognized. The real question concerns with the possibility to find these motivated students. The aim of this work is a little bit different from the previous approaches, where all the efforts where concentrate on the recognition of different level of students reading motivation. In this paragraph a brief historical literature introduction on the reading motivation and the different point of view have been reported.

Many research studies on motivation have focused their attention on the decrement of student motivation in the passage from the preschool to the high school and the way in which teachers have to deal with it. The fall of motivation depends on the fact that students feel increasingly alienated from learning in the passage from one type of school to another (Harter (1981)). The educational and the psychological point of view are the principal approaches used in this context.

In the educational literature some evidences on the reading motivation are related to the fact that students with a positive attitudes towards reading should like to read, see reading as a desirable activities and read voluntarily. As pointed by Cothern and Collins (1992), many factors have an influence on the attitude of children towards reading, such as the school and home environment, family and teacher behavior, socioeconomic status, gender and intelligence. Baker *et al.* (1997) noted that different reasons could explain the development of reading motivation, as the perception of reading as an enjoyable activity, the personal value or the opportunity of social interactions. Involvement, challenge or curiosity are examples of these causes.

In the psychological research, attention has been focused on the intra psychic influences on motivation, such as attributions (Weiner (1986)), self-efficacy (Schunk (1991)), perceived ability (Mclver *et al.* (1991)), perceived control and competence (Chapman *et al.* (1990); Weisz and Cameron (1985)), self-concept (Wigfield and Karpathian (1991)), intrinsic motivation (Corno and Rohrkemper (1985) ; Deci and Ryan (1985)), interest (Schiefele (1991)), learning strategies (Pintrich and De Groot (1990)) and goal orientations (Ames and Ames (1984), Dweck and Elliot (1983), Nicholls (1984)).

In the educational research, teachers and their behavior are one of the most important actors. Brophy (1986) identified some teacher guidelines to reach good results, such as guidance, modeling, enthusiasm, provision of choice, sincere praise, reinforcement, curiosity, dissonance and interest-induction. Keller (1983) included four basic strategies to stimulate students motivation: attention focusing, relevance, confidence building and satisfaction.

In our work both the psychological and the educational approaches have been considered. In this way, it is possible to provide a more wide and interesting perspective, because teacher behavior and student motivation has been considered at the same time. Educational literature individuates the best classroom practices that have a positive influence on the student attitudes and beliefs towards reading. The psychological literature give a considerable contribution to explain the reasons because student engagement is influenced by these beliefs.

3.2.1 Home context

The home environment is one of the most important factors useful to describe and analyze pupil reading motivation. Csikszentrnihalyi (1991) pointed that a child who see adults reading for pleasure, take for granted that reading is worthwhile. Baker *et al.* (1995) observed that a child living in an environment where literacy is promoted as a source of entertainment, are likely to become more motivated to read. Several methods have been proposed in literature to evaluate the home environment effect. The self-reports of interest (Scher and Baker (1996)) or the use of some behavioral indicators (Morrow (1983)) are two of the most famous examples.

In many articles on reading motivation, empirical results pointed out that children reading motivation is positively related with some student and family characteristics, as sex, income level and ethnicity. We decide to use these variables as covariates at individual level. We want to verify if the results reported in literature are also valid for the Italian educational context. Hansen (1969) analyzed the reading attitudes in relation with the home experiences and personal reading habits. Several variables could be used to define the home literacy environment as the available materials, the amount of reading done with child, the parents guidance, behavior and encouragement. The empirical results previously found, highlighted a significant relationship between the home environment and the ability of reading independently. Greaney and Hegarty (1987) pointed out that children with more positive motivation towards reading, engage their selves in a greater amount of leisure reading.

In other studies, the attention have been focused on the effect on motivation of parental beliefs, values, attitudes and expectation towards literacy. Guthrie and Greaney (1991) hypothesized the existence of two kinds of parental belief effects: direct (e.g., explicit communications to children, who internalizes parents ideas about reading) and indirect belief (e.g.,being enrollment in a library). These two effects have been considered to describe the parental role in the process of pupil reading motivation development.

3.2.2 Teacher and Class context

The teacher/class context is useful to analyze motivation. Students are influenced by class and school environment. Way of teaching, teacher enthusiasm and ability are examples of the external factors. We summarize the literature on the school environment effects on reading motivation.

Many authors observed that reading motivation decreases when children begin to go to school, because they start to evaluate and compare them selves with the other students. In this way, some of them could realize their own inability as readers and this awareness could cause a negative effect on motivation. Another explanation proposed by Guthrie and Wigfield (2000). They argued that: "*Explanation focuses on how instructional practices may contribute to a decline in some children motivation*". This means that bad habits such as social comparison between students which can cause extreme competition, may lead to the crush of a student confidence, competence belief, motivation, or mastery goal.

These two explanations introduce the problematic related to the school context. Motivation is influenced by students relationships and approach to reading. The teaching practices, involvement attitudes, classroom activities and self motivation are examples of these influences.

Oldfather and Dahl (1994) found a significant relationship between the class conditions and the change in motivation of the pupils. Roeser *et al.* (1996) pointed out that the way of teaching in a classroom fostered self-efficacy in students. In situation where teachers show their students that understanding is more important than answering correctly, pupils believe in their own capacity and seem to be able to do the most difficult work. Wentzel (1997) proved that students report higher level of motivation for learning when they perceive that teachers care more about their progress and well being. Sweet *et al.* (1998) established some guidelines for teachers such as choices (autonomy support), social interaction (relatedness support) and activity connections (competence support) in class. Empirical evidences showed that using textbooks and linking with outside resources as libraries and Internet, facilitate motivational development (Morrow and Young (1997)) and increase reading achievement (Guthrie *et al.* (2000)).

In the development process of motivation, an important role is "'played "' by the collaboration. It has been proved that social collaboration in class increases learning interest and sets the conditions to be more independent in future reading activities (Morrow (1983)). Great importance has been assessed to teachers personal involvement. Skinner and Belmont (1993) found that involved teachers (interested in students progress) and autonomy support, help to promote the motivation development. In this way, students feel more engaged and involved in class activities. Skinner and Belmont (1993) showed that influences between students and teachers are reciprocal because students engagement affects teachers involvement and vise versa.

3.3 The Mixture Factor Model Results

This section describes the results of the Mixture Factor Model(MFM), presented in paragraph 2.2. In the first part of this section, the result of the mixture factor model have been described.The second sub-paragraph reports the result obtained after the introduction of the individual covariates. The home context variables have been used to explain the differences between groups. With these models, the hierarchical data structure has been ignored because the aim is to group the students and not the classes. The effect on motivation caused by sharing the same environment have been ignored. We make the assumption that the teachers do not influence student motivation. This hypothesis is adverse to all the educational theory context, because literature proved that different environments affect student motivation and behavior. We group students considering three motivation dimensions: enjoyment of reading, value of reading and self concept as reader. The classification of the students has been obtained through the introduction of a mixture component at individual level. Ignoring the multilevel structure of the data could present some disadvantages, as the fact that some aspects related to motivation could be hide because of the class effects. The added value of ignoring teacher behaviors and class environment is useful to individuate the characteristics of a motivated students and the influences of the home context on the motivation development.

This model has been implemented to show how some bias could be committed ignoring the data structure. The next chapter is dedicated to the multilevel mixture factor analysis, because we want to show the advantages of considering the multilevel structure and how these evidences are supported by the data. In this work, the mixture factor model has been used as an exploratory tool to analyze and understand the latent construct at individual level, ignoring the multilevel structure.

3.3.1 The Mixture Factor Model without covariates

In this section the results obtained with the mixture factor models have been presented. Figure 3.1 reports a scheme of the mixture factor model. The factor structure is the same described in paragraph 2.1. The only difference concerns with the introduction of a mixture component at individual level. The aim is obtaining the student classification according to the latent factor means.

In the literature, there are not evidences to select the number of latent class. For this reason, several models have been implemented. A well known way of proceeding in presence of mixture component is varying the number of latent classes until the best solution has been found. The fit indexes have been reported in table 3.7. Usually, in this context several fit indexes have been proposed. The index reported in table 3.7 are the *BIC*, *AIC*, *AIC3* and *CAIC* (for more details see paragraph 2.5). All the fit indexes reported in table 3.7 indicate the same solution: the four latent classes model. The *BIC* assumes a value equal to 81071 and the *AIC* is equal to 80707.

To make the model estimable, some restrictions should be imposed on the factor loading. In table 3.8 the parameter estimates of the mixture factor

Number of Classes	Log Likelihood	BIC N=3158	AIC N=3158	AIC3 N=3158	CAIC N=3158	Number of Parameters	Classification Error
2	-40388	81194	80879	80930	81245	51	0.22
3	-40313	81077	80736	80791	81132	55	0.28
4	-40294	**81071**	**80707**	80766	81130	59	0.35
5	-40292	81099	80709	80772	81162	63	0.44

Table 3.7: Fit indexes for the Mixture Factor Model without covariates

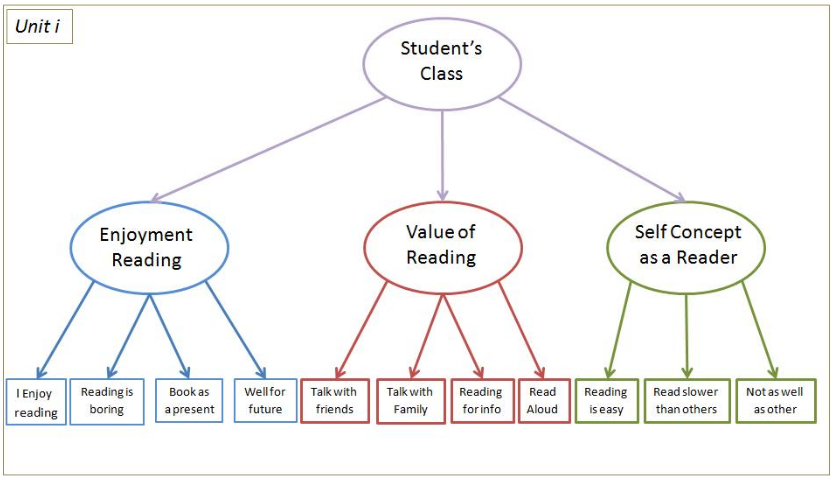

Figure 3.1: Mixture Factor Model

model have been reported. The factor loadings of the following items “Book as a present”, “Read aloud” and “Reading is easy” have been fixed to one. We impose also a restriction on the items of the factor loadings. The sum over the h modality of the variables must be equal to zero. The first restriction is useful to evaluate the importance (weight) of each item on the factor score mean. Values higher than one indicates the importance of items on factor score. Values smaller than one indicates that items give a less contribution to the factor means. From table 3.8 it is possible to see that for the enjoyment of reading dimension, the most important item is "‘I enjoy reading”’, "‘Talk with family”’ for value of reading, and "‘Not as well as other”’ for the self concept as reader. The sign minus before the coefficients, means that items are ordered in the opposite direction with regards to the measured dimension.

In table 3.9 and in figure (3.2) the factor mean values for each latent class are reported. The four groups represent respectively the 13%, 34%, 27% and 26% of the population. The students that belongs to the *High Motivated* class, are characterized by positive values for the enjoyment of reading (*0.69*) and self concept as reader (*2.17*).The *Motivated Bad Reader* class includes students that find reading an enjoyable activity because the value related to the enjoyment of reading is positive (*0.65*), but they are not good readers because the value for the self concept as reader is negative (-*0.61*). The students classified as *Unmotivated* reports negative values for the dimensions enjoyment of reading (-*0.44*) and self concept as reader (-*0.63*), and positive value for the dimension measuring value of reading (*0.18*). The students classified as *High unmotivated* reports the lowest values for the three

Factor	Item	Coef	s.e.	z-value	p-value
Enjoyment of reading	Book as a present	1	.	.	.
	Reading is boring	-1.65	0.13	-12.51	6.7E-36
	Well for future	0.70	0.06	11.71	1.2E-31
	I enjoy reading	**4.13**	0.52	7.96	1.7E-15
Value of reading	Read Aloud	1	.	.	.
	Talk with friends	1.36	0.13	10.11	5.0E-24
	Talk with family	**1.77**	0.18	10.13	4.1E-24
	Reading for info	1.31	0.13	10.35	4.2E-25
Self concept as reader	Reading is easy	1	.	.	.
	Not as well as other	**-2.32**	0.35	-6.63	3.3E-11
	Read slower than others	-1.65	0.18	-8.97	2.9E-19

Table 3.8: Factor loadings for the Mixture Factor Model without covariates

	High Motivated	Motivated Bad Reader	Unmotivated	High Unmotivated
Enjoyment of reading	0.69	0.65	-0.44	-0.90
Value of reading	0.06	0.02	0.18	-0.26
Self concept as reader	2.17	-0.61	-0.63	-0.93
Size	0.13	0.34	0.27	0.26

Table 3.9: Class mean factor scores for Mixture Factor Model without covariates

dimensions measuring motivation.

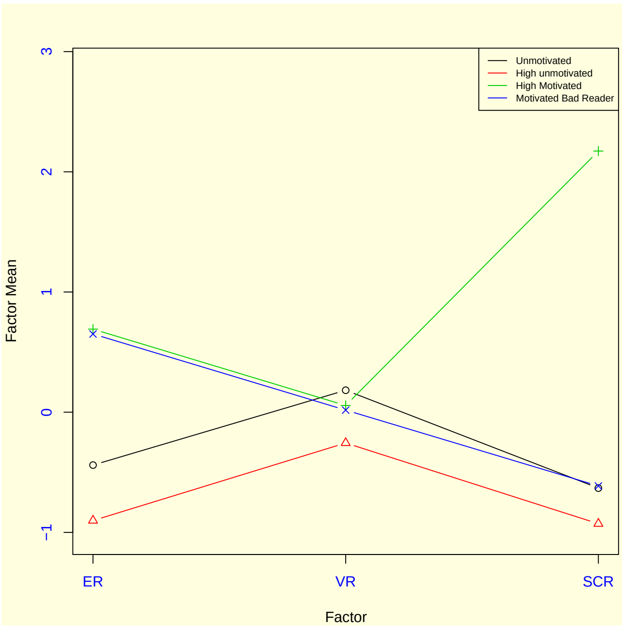

Figure 3.2: Plot class mean factor scores mixture factor model without covariates

These results support the hypothesis made on the population heterogeneity. The best solution is the model with four latent classes, and each of them identify a subpopulation of students that differ for the factor mean values. These differences concerns with two aspects of motivation: enjoyment in reading and self concept as reader. To explain the source of these differences, it is necessary to introduce student and family covariates to analyze the differential effect of the home context. These covariates could be useful to individuate the factors that have an influence on student motivation.

		Modal				
		High Motivated	Motivated Bad Reader	Unmotivated	High Unmotivated	Total
Probabilistic	High Motivated	424	28	22	9	482
	Motivated Bad Reader	156	868	100	76	1199
	Unmotivated	65	267	424	211	967
	High Unmotivated	22	134	171	607	933
	Total	716	902	667	1296	3581

Table 3.10: Classification Table for the Mixture Factor Model without covariates

In tables 3.7 and 3.10, it is possible to see the high percentage of misclassified observations. Table 3.10 is the classification table, and it is automatically implemented in Latent Gold. In this table it is possible to see the fact that many observations are incorrectly classified, because the number of elements out of the diagonal principal is huge. The misclassification rate is equal to 35.17%. The introduction of covariates is a way to reduce the percentage of misclassified observations.

3.3.2 The mixture factor model with covariates

This section describes the effects of the home context on the students motivation through the student and family covariates. The aim is to explain the differences in the factor score means. We want to individuate the factors (e.g., home environment) that have a positive or negative influence on the student reading motivation. We classify students ignoring their class membership (hierarchical structure). In the multilevel mixture factor model, we evaluate the class environment influences on motivation. The results have been reported below. We introduce different covariates to explain the factor score means of the three dimensions. Taking into account the educational and psychological literature, the selected covariates are: sex, nationality, parental education, importance of reading at home, predisposition towards school, duty of reading, library use, books at home and pre-entrance ability. Parental characteristics or behavior towards reading are useful to describe the effect of the home context, while the other covariates describe the attitude or tools used to increase reading ability.

In figure 3.3, a representation of the Mixture Factor Model with covariates is reported. Table 3.11 reports the fit indexes for this model. The best solution is the model with five latent classes. The *BIC* and the *AIC* indexes assume the lowest value (BIC=**80626.93**, AIC=**79965.31**). Comparing the results obtained in tables 3.11 and 3.7, it is possible to note that the introduction of covariates decreases the values of the fit indexes. The classification error for the mixture factor model decreases from 35.17% to 27.08%. These prove that covariates improve the classification power of the model.

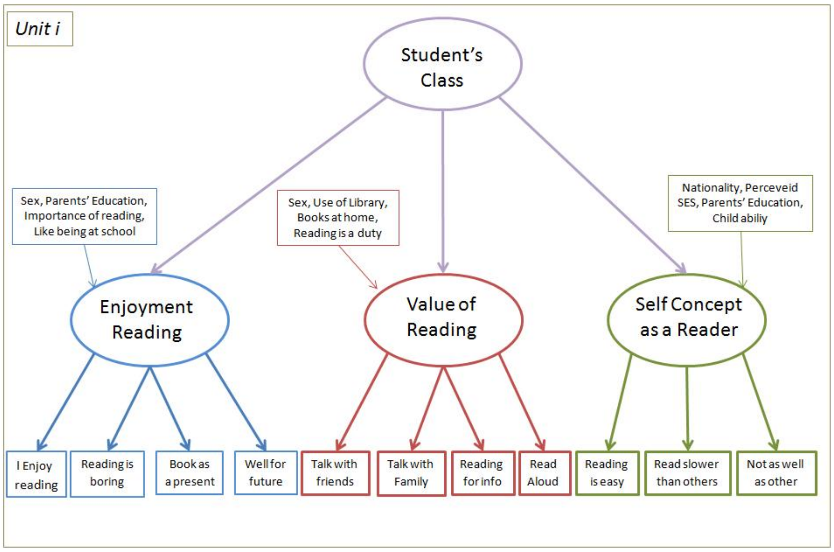

Figure 3.3: Mixture Factor Model with family and student covariates

Table 3.12 reports the value of the factor score means for the model with five latent classes. From figure 3.4 it is possible to see that the value of reading dimension do not give an important contribution in the classification of the students.

In table 3.12, 9% of the population represents the *High Motivated* students. The values for the enjoyment of reading (*1.65*) and self concept as reader are positive (*2.05*). The students classified as *Motivated Bad Reader* (32%) are characterized by positive value for the enjoyment of reading (*0.98*), while the value associated to the the self concept as reader is negative (*-0.85*). This means that the students belonging to this class like reading but they are not good readers. The difference between the classes *Medium Motivated Good Reader* (8%) and *Medium Motivated Bad Reader* (50%) is related to the factor mean value of the self concept as reader. The factor mean values for the students belonging to the *High Unmotivated* class (1.5%) are all negative, and so these students identify the critical subjects of our analysis.

The covariates inserted to explain the enjoyment of reading dimension are: sex, parental education, importance of reading and pleasure of being at school. These variables have been selected according to the theoretical context (paragraph 3.2.1). Positive values in table 3.13 indicate a positive effect on the enjoyment of reading dimension. Sex has a discriminant role, indeed being a girl (*0.15(0.02)*) compared to being a boy have a positive effect. This means that girls find reading a more enjoyable activity compared

Number of Classes	Log Likelihood	BIC N=3158	AIC N=3158	AIC3 N=3158	CAIC N=3158	Number of Parameters	Classification Error
2	-39991	80776	80176	80273	80873	97	0.18
3	-39941	80693	80081	80180	80792	99	0.25
4	-39907	80656	80019	80122	80759	103	0.29
5	-39876	**80627**	**79965**	80072	80734	107	0.27
6	-39872	80652	79966	80077	80763	111	0.38

Table 3.11: Fit indexes for the Mixture Factor Model with covariates

Factor	High Motivated	Motivated Bad Reader	Medium Motivated Good Reader	Medium Motivated Bad Reader	High Unmotivated
Enjoyment of reading	1.65	0.98	-0.25	-0.32	-2.06
Value of reading	0.22	0.24	0.01	0.07	-0.54
Self concept as reader	2.05	-0.85	0.93	-1.12	-1.01
Size	0.09	0.32	0.08	0.50	0.02

Table 3.12: Factor Mean for the Mixture Factor Model with covariates

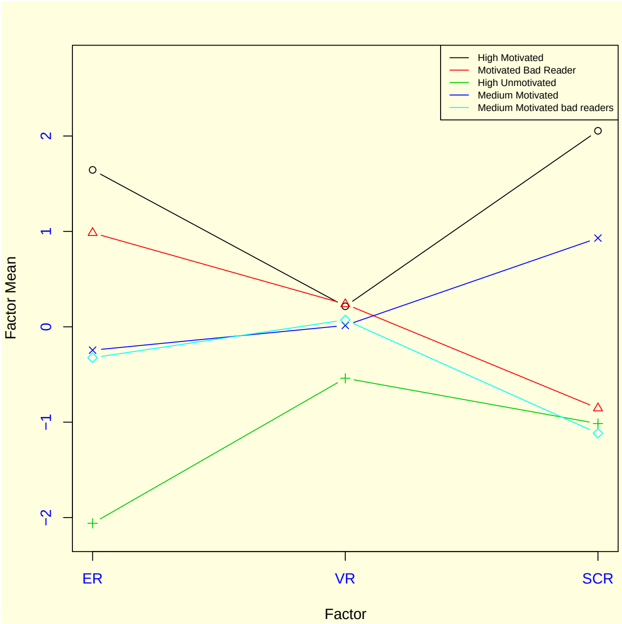

Figure 3.4: Plot class mean factor scores Mixture Factor Model with covariates

to boys. Positive effect have also been found for the parental education (University or Higher (*0.13 (0.04)*)). The importance of reading activities at home has positive effect on motivation (Agree a lot (*0.15(0.03)*)) compared to an environments where reading do not have an important role (Disagree a lot (-*0.15(0.06)*)). Student attitudes towards school have a positive influence on the reading motivation (Agree a lot (*0.41(0.03)*) and Agree a little *0.14(0.03)*) compared to bad attitudes of children towards school (Disagree a little $-0.08(0.03)$ and Disagree a lot -*0.47(0.05)*).

It is possible to trace the profile of a motivated student and to highlight the factors that affect motivation. Living in a context where parental education is high, parents assess high importance to reading activities and children who like being at school have a positive influence on reading motivation.

In table 3.14 the parameter estimates for the covariates inserted in the model to explain the value of reading dimension have been reported. The covariates are: sex, duty of reading, like being at school, use of the library and number of books at home. For sex and like being in school, the same results for the enjoyment of reading dimension have been found. The interpretation of the results related to the variable "'Reading is a duty"' is in contrast with the theoretical context, because a good climate towards reading at home should have a positive effect on motivation, while in this situation the results

Factor	Item	Coef	s.e.	z-value	p-value	Wald(0)	df	p-value	
Enjoyment of reading	sex	Girl	0.15	0.02	9.09	9.50E-20	82.70	1	9.50E-20
		Boy	-0.15	0.02	-9.09	9.50E-20			
	Finished post-secondary Education	University or Higher	0.13	0.04	3.29	0.001	17.51	4	0.0015
		0.02	0.05	0.51	0.61				
		Finished upper-secondary	0.03	0.03	1.09	0.28			
		Finished lower-secondary	-0.05	0.03	-1.66	0.01			
		Some primary or no schooling	-0.13	0.09	-1.51	0.13			
	Importance of reading	Agree a lot	0.15	0.03	4.90	9.50E-07	25.50	3	1.20E-05
		Agree a little	0.01	0.03	0.33	0.74			
		Disagree a little	0.00	0.04	-0.04	0.97			
		Disagree a lot	-0.15	0.06	-2.74	0.006			
	Like being in school	Agree a lot	0.41	0.03	12.80	1.60E-37	170.448	3	1.00E-36
		Agree a little	0.14	0.03	5.49	4.00E-08			
		Disagree a little	-0.08	0.03	-2.40	0.016			
		Disagree a lot	-0.48	0.05	-9.89	4.40E-23			

Table 3.13: Parameters estimates of the latent factor Enjoyment of Reading for the Mixture Factor Model with covariates

go the other way. One of the reasons of this discrepancy depends on the ambiguity of the question, because it is not clear which aspect of reading should be considered as a duty. A frequent use of the library has a positive effect on motivation, because students feel more involved in reading activities. This is an example of an indirect effect of parental beliefs on the motivation of the children. Living in a high cultural developed environment (the numbers of books at home is a proxy of this dimension) has a positive effect on reading motivation (More than 200 *0.05(0.02)*).

In table 3.15, the parameter estimates for the self concept as reader have been reported. The covariates are: nationality, perceived SES, parental education level and pre-entrance ability. Being an Italian student has a positive influence on the reading awareness. The results related to the perceived Socio Economic Status have a strange interpretation and this variable is not useful to describe motivation. The reason is related to the fact that it is not an objective measure of the family SES. The same results have been found for the parental education. Interesting results have been highlighted with the introduction of the covariates describing the pre-entrance abilities. These variables describe the role of the abilities acquired before starting school have as predictor of the actual reading awareness. A good ability in reading sentence (Very well *-0.13(0.05)*) or in writing some words (Very well *-0.08 (0.03)*) has a positive effect on the self concept as reader dimension. A developed environment (More than 200 books *-0.07 (0.03)*) has a positive effect on the children reading awareness.

It is possible to outline the profile of a motivated student and to describe the positive influence on the reading motivation of the home context.

An Italian girl who likes being at school, going to the library and that was able to read sentences and write some words before starting school, perfectly identify the profile of a motivated student. It is possible to identify the best environment that affect positively the student reading motivation development. High level of parental education, a promoting and challenging environment where reading is seen as an important activity in daily life, describe the best context where a child should live.

Many of these results were found by several authors (see section 3.2) in the past educational and psychological literature. One of the goals of these analysis was the validation of the existent theories on reading motivation. The applicative context was not the same of the cited articles in section 3.2, but the results they are very similar.

Factor	Item	Coef	s.e.	z-value	p-value	Wald(0)	df	p-value	
Value of reading	sex	Girl	0.06	0.01	-5.6	2.10E-08	31.3602	1	2.10E-08
		Boy	-0.06	0.01	5.6	2.10E-08			
	Reading is a duty	Agree a lot	0.08	0.02	3.19	0.0014	41.33	3	5.60E-09
		Agree a little	0.08	0.02	3.60	0.00032			
		Disagree a little	-0.06	0.02	-2.94	0.0032			
		Disagree a lot	-0.10	0.02	-5.66	1.50E-08			
	Like being in school	Agree a lot	0.07	0.02	3.42	0.0006	12.66	3	0.0054
		Agree a little	0.02	0.02	1.03	0.3			
		Disagree a little	-0.04	0.03	-1.48	0.14			
		Disagree a lot	-0.05	0.04	-1.22	0.22			
	Frequency of library use	At least once a week	0.03	0.02	2.06	0.04	10.89	3	0.012
		Once or twice a month	-0.02	0.02	-0.93	0.35			
		A few times a year	0.03	0.02	1.25	0.21			
		Never or almost never	-0.04	0.02	-2.70	0.0069			
	Books at home	0-10	-0.12	0.02	-5.00	5.80E-07	26.24	4	2.80E-05
		11-25	0.04	0.02	2.02	0.043			
		26-100	0.02	0.02	1.05	0.29			
		101-200	0.02	0.02	0.91	0.37			
		More than 200	0.05	0.02	2.18	0.03			

Table 3.14: Parameters estimates of the latent factor Value of Reading for the Mixture Factor Model with covariates

Factor	Item	Coef	s.e.	z-value	p-value	Wald(0)	df	p-value	
Self concept as reader	Nationality	Italian	0.11	0.03	3.77	0.0002	14.20	1	0.0002
		Not Italian	-0.11	0.03	-3.77	0.0002			
	Perceived SES	Very well-off	-0.11	0.06	-1.72	0.086	16.36	4	0.003
		Somewhat well-off	0.06	0.04	1.51	0.13			
		Average	0.09	0.3	3.35	0.0008			
		Not very well-off	0.10	0.04	2.58	0.01			
		Not at all well-off	-0.14	0.06	-2.33	0.02			
	Education	University or Higher	0.21	0.04	5.35	8.70E-08	32.58	4	1.50E-06
		Finished post-secondary	-0.06	0.04	-1.35	0.18			
		Finished upper-secondary	0.06	0.03	2.14	0.032			
		Finished lower-secondary	-0.04	0.03	-1.26	0.21			
		Some primary or no schooling	-0.18	0.07	-2.49	0.013			
	Child ability read sentences	Very well	0.13	0.05	2.83	0.005	21.02	3	0.0001
		Moderately well	0.04	0.03	1.30	0.19			
		Not very well	-0.03	0.03	-1.39	0.17			
		Not at all	-0.13	0.03	-4.54	5.70E-06			
	Child ability write some words	Very well	0.08	0.03	2.69	0.007	10.28	3	0.016
		Moderately well	0.03	0.02	1.36	0.17			
		Not very well	-0.04	0.02	-1.60	0.11			
		Not at all	-0.08	0.03	-2.40	0.017			
	Books at home	0-10	-0.10	0.03	-3.02	0.003	13.07	4	0.011
		11-25	-0.03	0.02	-1.14	0.26			
		26-100	0.04	0.02	1.54	0.12			
		101-200	0.02	0.03	0.74	0.46			
		More than 200	0.07	0.03	2.16	0.03			

Table 3.15: Parameters estimates of the latent factor Self Concept as a Reader for the Mixture Factor Model with covariates

		With covariates					
		High Motivated	Motivated Bad Reader	Medium Motivated Good Reader	Medium Motivated Bad Reader	High Unmotivated	Total
Without covariates	High Motivated	**511**	37	112	7	0	667
	Motivated Bad Reader	1	**1072**	0	223	0	1296
	Unmotivated	0	55	72	**581**	8	716
	High Unmotivated	0	18	60	**787**	37	902
	Total	512	1182	244	1598	45	3581

Table 3.16: Comparison between student class membership for the mixture factor model without and with lower level covariates on factor score

		With covariates					
		High Motivated	Motivated Bad Reader	Medium Motivated Good Reader	Medium Motivated Bad Reader	High Unmotivated	Total
Without covariates	High Motivated	**99.80%**	3.13%	**45.90%**	0.44%	0.00%	667
	Motivated Bad Reader	0.20%	**90.69%**	0.00%	13.95%	0.00%	1296
	Unmotivated	0.00%	4.65%	**29.51%**	**36.36%**	17.78%	716
	High Unmotivated	0.00%	1.52%	**24.59%**	**49.25%**	**82.22%**	902
	Total	512	1182	244	1598	45	3581

Table 3.17: Comparison between student class membership for the mixture factor model without and with lower level covariates on factor score: column percentage

3.3.3 Covariates effect in the Mixture Factor Model

In this section, we highlight the classification differences due to the covariates in the mixture factor model. Tables 3.16, 3.17 and 3.18 are dedicated to the comparison between the two classification obtained. Table 3.16 reports the cross classification of the subjects. The observations belonging to the same class (same classification) have been highlighted in bold. The differences between the two classification are related to the Medium Motivated Bad Reader class in the mixture factor model with covariates and to the *Unmotivated* and *High Unmotivated* classes in the mixture factor model without covariates. In tables 3.17 and 3.18, the covariates effects on model classification have been reported. The individual level covariates reduce the differences between students. Many students classified as *Unmotivated* and *High Unmotivated* in the model without covariates, change their membership after the introduction of covariates. This means that student differences are flattened out. The principal effect of covariates is on the awareness of reading, because it is a discriminant dimension to describe the student reading motivation. Table 3.18 reports the raw percentage. This table show the matching between the model classification. Table 3.17 reports the column percentages. This table highlights the matching between the models classification.

In table 3.19 the within and between variances of the latent factors for the mixture factor model with and without covariates have been reported.

		With covariates					
		High Motivated	Motivated Bad Reader	Medium Motivated Good Reader	Medium Motivated Bad Reader	High Unmotivated	Total
Without covariates	High Motivated	**76.61%**	5.55%	16.79%	1.05%	0.00%	667
	Motivated Bad Reader	0.08%	**82.72%**	0.00%	17.21%	0.00%	1296
	Unmotivated	0.00%	7.68%	10.06%	**81.15%**	1.12%	716
	High Unmotivated	0.00%	2.00%	6.65%	**87.25%**	4.10%	902
	Total	512	1182	244	1598	45	3581

Table 3.18: Comparison between student class membership for the mixture factor model without and with lower level covariates on factor score: row percentage

	Within Variance			Between Variance		
	With Covariates	Without Covariates	Percentage of reduction	With Covariates	Without Covariates	Percentage of reduction
Enjoyment of reading	0.16	0.37	-58.37%	2.00	0.64	217.91%
Value of reading	0.16	0.16	0.00%	0.10	0.03	192.52%
Self concept as reader	0.18	0.20	-10.79%	2.02	2.12	-4.53%

Table 3.19: Percentage reduction of Within and Between Variance with and without covariates for the mixture factor model

The covariates effect on the within variance is related to the enjoyment of reading and self concept as a reader dimensions. The covariates cause a 58.37% reduction of the variance for the enjoyment of reading and a 10.79% for the self concept as reader dimension. This means that the differences within group have been reduced by the covariates, making the subpopulations more homogeneus. Analyzing the between variance, it is possible to see an increment of the variance for the enjoyment of reading. This means that the differences between group are more accentuated.

3.3.4 Construct validity for the Mixture Factor Model

This section is dedicated to the construct validity, that is a way to find the connection between a theoretical concept and a specific measuring device or procedure. This is often used in the context of the applied science, as psychometrics or marketing. For example, a researcher inventing a new IQ test might spend a great deal of time attempting to "define" intelligence in order to reach an acceptable level of construct validity. In our application, we measure the reading motivation choosing to use some items to define a structure in order to measure this unobservable concept.

The construct validity is subdivided into two categories: convergent validity and discriminate validity. The first one is the actual general agreement among ratings, gathered independently of one another, where measures should be theoretically related. The second is the lack of a relationship among measures which theoretically should not be related.

A three steps procedure has been outlined by Carmines and Zeller (1979) to approach the construct validity:

1. Specification of the theoretical framework,
2. Examination of the empirical relationships between the measures of the concepts,
3. Interpretation of the empirical evidences in terms of how it clarifies the construct validity of the particular measure being tested.

Construct validity is related to the theoretical ideas behind the trait under consideration. The evaluation of a construct validity requires the examination of the correlations of the measures to understand which variables should be related to the construct. In our application, we show how motivation has a positive effect on student ability, highlighting in this way the importance of studying motivation to evaluate and compare the development level of a country. For these reasons, the Analysis of Variance (ANOVA) has been used. We want to analyze the reading ability differences between the group of students considering the classification obtained with the mixture factor model. In the ANOVA the observed variance is partitioned into components due to the different explanatory variables. The explicative variable is the classification obtained with the mixture factor model, where students have been grouped according to their level of motivation. ANOVA tests the null hypothesis that the samples have been drawn from the same population. We decide to compare the group means using the Dunnett test, designed specifically for situations where all the groups are to be pitted against one, called the *Reference* group. It is often used after the rejection of the null hypothesis, that is the equality of all the group means. The goal is the identification of the groups whose means are significantly different from the reference group mean. The null hypothesis is that no group has a mean significantly different from the reference group mean.

In tables 3.20 and 3.22 the result of the ANOVA for the variables measuring the plausible values have been reported. The different aspects of the reading ability are: overall reading, informational purpose, literacy purpose, interpreting process and straightforward process. These variables measure the effective reading ability of children. They have been calculate weighting the answers from the cognitive questionnaire.

In these two tables, we show how the classification obtained with mixture factor model is useful to improve the recognition of the clever students. In

Variables	Category	Mean Difference	Standard Error	p-value	Lower Bound	Upper Bound
Overall	High Unmotivated	-11.07	3.29	2.16E-03	-18.76	-3.39
	High Motivated	38.86	3.53	4.40E-08	30.60	47.12
	Motivated Bad Reader	16.80	3.06	1.64E-07	9.65	23.94
Information	High Unmotivated	-9.36	3.08	6.67E-03	-16.57	-2.16
	High Motivated	35.81	3.31	4.40E-08	28.07	43.56
	Motivated Bad Reader	16.26	2.87	8.68E-08	9.56	22.96
Literacy	High Unmotivated	-13.55	3.54	3.74E-04	-21.81	-5.29
	High Motivated	39.19	3.80	4.40E-08	30.31	48.08
	Motivated Bad Reader	18.73	3.29	8.12E-08	11.04	26.42
Interpreting	High Unmotivated	-8.98	3.13	1.12E-02	-16.29	-1.68
	High Motivated	35.44	3.36	4.40E-08	27.58	43.29
	Motivated Bad Reader	13.88	2.91	5.86E-06	7.08	20.67
Straightforward	High Unmotivated	-8.65	3.35	2.59E-02	-16.48	-0.83
	High Motivated	38.46	3.60	4.40E-08	30.05	46.87
	Motivated Bad Reader	20.93	3.11	4.40E-08	13.65	28.21

Table 3.20: ANOVA on Plausible Values where the factor is the class membership of the mixture factor model without covariates, reference category Unmotivated class

table 3.20 the reference category is the *Unmotivated* class. Each contrast is significant for all the plausible values. Only the contrast related to the variable *Overall Reading* has been commented, because the same results have been found for the other plausible values. The value of the contrast for the *High Unmotivated* class is negative (*-11.07 (2.16E-03)*). This means that, on the average, students classified as *High Unmotivated* are less able than the *Unmotivated* students . The values of the contrasts for the *High Motivated* and the *Motivated Bad Reader* students are positive (*38.86* and *16.80*). It is possible conclude that the student reading motivation has a high influence on each aspects of reading ability. This means that studying reading motivation could be useful to predict the students reading ability. In table 3.21 the mean values for each plausible values have been reported. Figure 3.5 shows the means plot, considering the latent class membership.

The same reasoning has been done for the mixture factor model with covariates. Table 3.22 reports the results of the ANOVa. The variables are the plausible values and the factor is the classification obtained with the mixture factor model without covariates. The reference category is the *Medium motivated Bad Reader* class. This class is composed of students that are motivated but they do not present high value for the reading awareness. The results for the Overall Reading variable have been reported. The contrast between the *High Motivated* class and the reference category is *41.5 (4.40E-08)*

	Overall	Information	Literacy	Interpreting	Straightforward
Unmotivated	541	537	540	546	531
High Unmotivated	530	528	527	538	522
High Motivated	580	573	580	582	569
Motivated Bad Reader	558	554	559	560	551

Table 3.21: Plausible Values Means considering the classification obtained with the mixture factor model without covariates

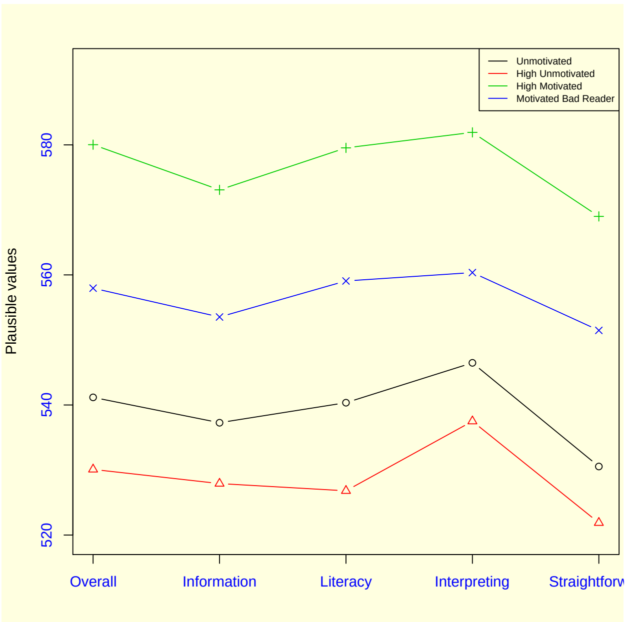

Figure 3.5: Means plot for the plausible value considering the classification obtained with the mixture factor model without covariates

Variables	Category	Mean Difference	Standard Error	p-value	Lower Bound	Upper Bound
Overall	High Motivated	41.50	3.36	4.40E-08	33.14	49.85
	Motivated Bad Reader	22.17	2.54	4.40E-08	15.86	28.48
	High Unmotivated	-5.37	10.00	9.70E-01	-30.22	19.49
	Medium Motivated Good Reader	30.20	4.55	4.41E-08	18.90	41.51
Information	High Motivated	41.76	3.62	4.40E-08	32.76	50.77
	Motivated Bad Reader	22.84	2.74	4.40E-08	16.04	29.64
	High Unmotivated	-10.87	10.78	7.67E-01	-37.67	15.93
	Medium Motivated Good Reader	29.90	4.90	4.86E-08	17.71	42.09
Literacy	High Motivated	37.00	3.15	4.40E-08	29.16	44.85
	Motivated Bad Reader	19.37	2.38	4.40E-08	13.45	25.30
	High Unmotivated	1.54	9.39	1.00E+00	-21.80	24.88
	Medium Motivated Good Reader	27.32	4.27	4.46E-08	16.71	37.93
Interpreting	High Motivated	37.02	3.19	4.40E-08	29.10	44.95
	Motivated Bad Reader	18.70	2.41	4.40E-08	12.71	24.69
	High Unmotivated	-10.64	9.49	6.92E-01	-34.24	12.96
	Medium Motivated Good Reader	29.90	4.32	4.40E-08	19.17	40.63
Straightforward	High Motivated	40.93	3.42	4.40E-08	32.43	49.42
	Motivated Bad Reader	24.15	2.58	4.40E-08	17.73	30.57
	High Unmotivated	-3.61	10.18	9.94E-01	-28.91	21.68
	Medium Motivated Good Reader	26.48	4.63	8.87E-08	14.98	37.98

Table 3.22: ANOVA on Plausible Values where the factor is the class membership of the mixture factor model with covariates, reference category Medium Motivated Bad Reader class

and the contrast for the *Motivated Bad Reader* is *22.17 (4.40E-08)*. These contrasts highlight the important role played by motivation in the analysis of the reading achievement. Comparing the average value for the *High Unmotivated* class with the reference group, it is possible to note that these two groups are not different (p-values=0.97). The last contrast highlight the importance of self concept as reader, because the mean difference between the groups of students classified as *Medium Motivated good reader* and *Medium Motivated bad reader* is equal to *30.20 (4.41E-08)*. In table 3.23, the mean values for each plausible values have been reported. Figure 3.6 reports the mean values for each latent class individuated.

	Overall	Information	Literacy	Interpreting	Straightforward
High Motivated	578	571	577	580	568
Motivated Bad Reader	559	553	558	561	552
High Unmotivated	531	535	525	532	524
Medium Motivated Good Reader	567	561	566	572	554
Medium Motivated Bad Reader	536	534	536	542	528

Table 3.23: Plausible Values Means considering the classification obtained with the mixture factor model with covariates

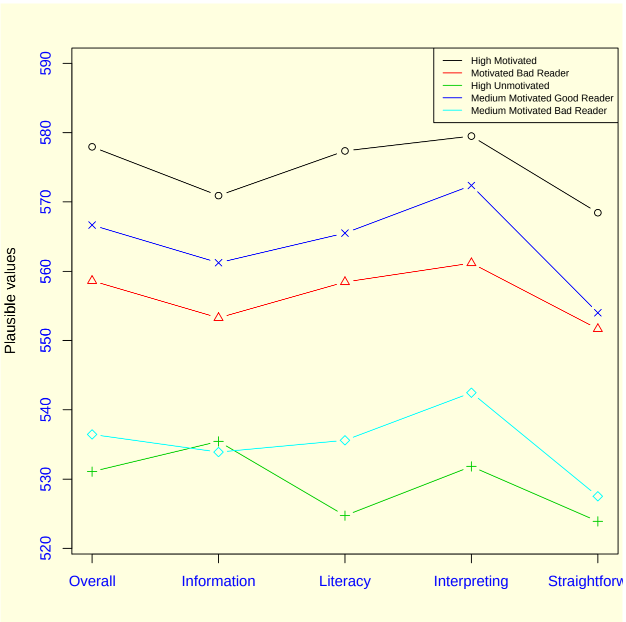

Figure 3.6: Plot mean plausible value for considering the classification obtained with the mixture factor model with covariates

Chapter 4

The multilevel mixture factor model

This chapter describes the results obtained with the multilevel mixture factor model ($MMFM$) and it is subdivided into three sections. different models are reported. Section 4.2 reports the results of the multilevel mixture factor model, where any covariates have been inserted. The goal of this model is the classification of the classes according to the reading student motivation measured at individual level. We want to model the group level heterogeneity. Section 4.3 reports the results obtained after the introduction of the lower level covariates to explain the factor means. The purpose of this model is to describe how the individual latent structure is affected by the home context. We want to highlight the factors that influence the students reading motivation. Section 4.4 reports the results obtained after the introduction of the higher level covariates. The aim is to individuate and describe the teacher effects on the entire class motivation.

Below it has been shown the advantages of considering the multilevel structure to model this kind of data. In table 4.1 the values of the Intraclass Correlation Coefficient (ICC) and the Likelihood Ratio Test(LRT) have been reported. The ICC index

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma_{00}}$$

represents the percentage of the variance accounted to the group level considering the multilevel structure. The quantity τ_{00} represents the variance at the higher level for a specific item, while σ_{00} represents the lower level variance for that item. In our application, the value of σ_{00} is equal to the variance of a multinomial distribution ($\pi^2/3$), because of the specification of the scale type of the items. The other values reported in table 4.1, are

Item	τ_h	ICC_h	LRT	p-value
Book as a present	0.18	5.1%	41.4	1.24E-10
Reading is boring	0.36	9.8%	82.3	1.17E-19
Well for future	0.28	7.8%	29.0	7.24E-08
I enjoy reading	0.20	5.6%	44.0	3.28E-11
Read Aloud	0.49	13.1%	224.4	9.92E-51
Talk with friends	0.32	8.7%	126.0	3.08E-29
Talk with family	0.29	8.1%	106.0	7.37E-25
Reading for info	0.34	9.4%	114.6	9.63E-27
Reading is easy	0.29	8.2%	82.0	1.36E-19
Not as well as other	0.23	6.4%	75.6	3.47E-18
Read slower than others	0.15	4.3%	24.3	8.24E-07

Table 4.1: Intraclass Correlation Coefficient and Likelihood Ratio Test for the Multilevel Mixture Factor Model

the Likelihood Ratio Test. This test is used to compare two models, one of which is nested within the other, and it is based on the comparison of the log-likelihood of the two models.

$$D = -2ln\left(\frac{likelihood\ under\ H_0}{likelihood\ under\ H_1}\right)$$

In our application, the null hypothesis (H_0) is represented by the model without the multilevel structure (the one we want to reject), while the alternative hypothesis is represented by the model that include the multilevel structure. Large values for the statistic test, suggest that the model under the null hypothesis does not give a satisfying description of the data, compared to the model under the alternative hypothesis. In other words, this statistics provide evidences about the model fit lack. The intra correlation coefficient and the likelihood ratio test have been calculated for the *11* items measuring the students reading motivation.

The values of the ICC_h (see table 4.1) are all into a range from 4.27% for item “Read slower than others” to 13.06% for the item “Read Aloud”. The p-values for the likelihood ratio tests (LRT) are all significant. The results reported in table evidences 4.1 confirm the fact that the multilevel structure should be considered.

4.1 The results of the multilevel mixture factor model

The multilevel mixture factor model has been presented in section 2.3. The structure could be easily summarized in the following way: at individual level, the items are connected to three different factors that measure the student reading motivation, while at group level, a mixture component has been hypothesized to model the population heterogeneity. The teachers and the classes could be described, classified and compared on the basis of the latent structure at student level. The individual structure is the same described in section 3.3. The principal difference concerns with the position of the mixture component. In the mixture factor model, the latent class is at individual level, because the aim is the classification of the level-one subjects. In the multilevel mixture factor model, the interest has been addressed to the classification of the classes according to the student motivation. The goal of this model is multiple, because we want to compare the teacher practices in class and analyze the influences on the reading motivation. Section 4.2 is dedicated to the simplest version of the Multilevel Mixture Factor Model. We want to highlight the relationships between items and factor scores. The classification of teachers is one of the goals of this model. Section 4.3 describes the effects of the individual level covariates. We want to discover and understand the nodal points on which attention should be turned to increase student reading motivation.

The results of the multilevel mixture factor model after the introduction of the higher level covariates have been reported in the section 4.4. After controlling for family and student characteristics (lower level covariates) to eliminate the differences related to the home context, it is possible to compare the classes motivation and describe the different effects of teachers. Introducing the group level covariates, we want to discover how teacher characteristics impact on the motivation of students. The hypothesis is that teachers differ in the way of motivating students. One of the focal point consist of highlighting the best practices (material and resources use) in class.

4.2 Multilevel mixture factor model

The multilevel mixture factor model has been described in paragraph 2.3 in equations 2.20, 2.21 and 2.24. In the first level of the hierarchy, a standard factor model has been specified (see equation 2.20). It represents the relationship between the observed items and the latent continuous construct measuring motivation. Figure 4.1 represents the two level mixture factor model, where the multilevel structure is clearly visible.

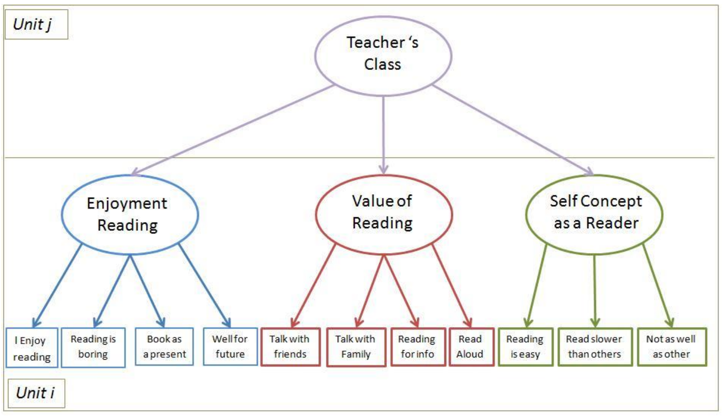

Figure 4.1: Multilevel mixture factor model

The aim of this model is to obtain the classification of teachers/classes (the second level units) into homogeneous groups, according to the reading motivation of the classes. The assumption that teachers come from different sub-populations seems to be reasonable because it is well known that teachers could have a different influences on students. To model this unobserved heterogeneity, we introduce a latent categorical variable at the group level.

In the multilevel context, it has been proved that the best fit index is the *BIC* with sample size equal to number of higher level observation. For more details, it is possible to refer to the article of Lukočienė *et al.* (ress). In that article, a simulation study showed that in the multilevel mixture models, the best fit index is the *BIC* with sample size equal to the group level observations. This evidence is due to the fact that the underestimation of the number of mixture components is much more likely using the *BIC* with the number of groups than with the number of observations. The reason of this choice concerns with the fact that using the fit indexes with the number of observations, it is possible to underestimated the number of mixture components, especially if the separation between components is weak or moderate Lukočienė and Vermunt (2010). In table 4.2, the fit indexes for the MMFM without covariates have been reported. The seven classes solution is the best model (BIC=81024).

Some restrictions must be imposed to make the model estimable. In table 4.3 the factor loadings have been reported and it is possible to see that some of them have been fixed to one. Values greater than one show that the importance of that item to explain motivation. Vice versa, values smaller than one indicate a low contribution of the item for that dimension. Table

Number of Classes	Log Likelihood	BIC N=3158	BIC N=198	Number of Parameters
3	-40415	81281	81121	55
4	-40381	81245	81074	59
5	-40352	81220	81037	63
6	-40336	81220	81026	67
7	-40324	81229	**81024**	71
8	-40320	81254	81037	75

Table 4.2: Fit indexes for the Multilevel Mixture Factor Model

4.4 reports the the latent classes size and the factors score means. Positive values indicate high level of motivation. The situation have been illustrated in figure 4.2.

To avoid misunderstanding, in the next paragraphs the term *class* will indicate the latent class or mixture component, while the term *classroom* the group of students sharing then same environment and teacher. The classrooms belonging to the *Very Highly Motivated* class (5%) are charactherized by the highest value for two thirds of the dimensions measuring pupil reading motivation, enjoyment of reading (*ER=.56*), and self concept as reader (*SCR=.69*). The *High Motivated* class (16%) is composed of classrooms that ascribe high importance to reading (*VR=.32*), and find enjoyable reading activities (*ER=.36*). The classrooms classified as *Motivated Bad Reader* (21%), are composed of students that like reading (*ER=.39*), but these classrooms of children are not good readers (*SCR=-.16*). 24% of classrooms belong to the *Medium Motivated* class because of the values assumed by the factor means that are close to zero. Students belonging to classrooms classified as *Unmotivated* (12%), are classified as bad readers (*SCR=-.21*). They don't see reading as an unenjoyable activity (*ER=-.39*), but as a valuable activity (*VR=.12*). The *Highly Unmotivated* class (13%), exihibits a similar behavior compared to the previous class, excluding the value of reading *VR=-.22*). Lastly, classrooms belonging to the *Very Highly Unmotivated* class (9%) are characterized by the lowest value for factor means. As it has been possible to see from the results reported above, the hypothesis on the teachers population heterogeneity, has been confirmed by the data analysis.

Table 4.5 reports the bivariate residuals (see paragraph 2.5) that is a useful tool to verify if the local independence assumption is reasonable. We highlight the situation in which the values are bigger than the critical value. For those items, it is possible to assume that the local independence assumption does not hold. Considering the items connected to the same factor, we

Item	Factor	Coef	s.e.	z-value	p-value
Enjoymento of reading	Book as a present	1	.	.	.
	Reading is boring	-1.56	0.13	-12.45	1.5E-35
	Well for future	0.71	0.06	11.49	1.5E-30
	I enjoy reading	**2.56**	0.18	14.01	1.3E-44
Value of reading	Read Aloud	1	.	.	.
	Talk with friends	1.39	0.14	10.05	9.1E-24
	Talk with family	**1.66**	0.16	10.46	1.3E-25
	Reading for info	1.27	0.12	10.35	4.3E-25
Self concept as reader	Reading is easy	1	.	.	.
	Not as well as other	**-2.00**	0.23	-8.61	6.8E-18
	Read slower than others	-1.44	0.13	-10.69	1.2E-26

Table 4.3: Factor loadings for the Multilevel Mixture Factor Model without covariates

	Very Highly Motivated	High Motivated	Motivated Bad Reader	Medium Motivated	Unmotivated	Highly Unmotivated	Very Highly Unmotivated
Enjoyment of Reading	.56	.36	.39	.02	-.39	-.58	-.36
Value of Reading	.18	.32	.01	-.03	.12	-.22	-.38
Self Concept as Reader	.69	-.01	-.16	.12	-.21	-.09	-.34
Size	.05	.16	.21	.24	.12	.13	.09

Table 4.4: Class mean factor scores for the Multilevel Mixture Factor Model

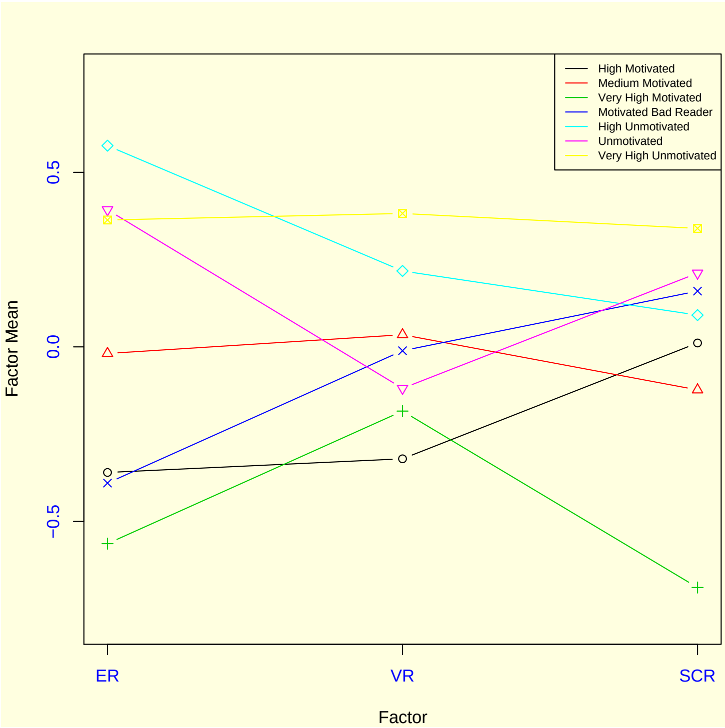

Figure 4.2: Plot class mean factor scores for the Multilevel Mixture Factor Model without covariates

can note that the local independence assumption is valid for the most part of them. For the situations in which the local independence assumption is violated, it is possible to solve the problem replacing or deleting these items. In our work, we decided to remove some of the items initially inserted in the model. Table 4.6 reports the variance and covariance structure of the latent factors. The between variance for the *value of reading* dimension assumes the smallest value(0.15). This means that there is a class effect on the value of reading for the students. The classification table for the Multilevel Mixture Factor (table 4.7) is a tool used to analyze the classification power of the model. In this table, the class memberships based on the empirical Bayesian posterior distribution (empirical Bayes and empirical Bayes modal) have been reported. The classification error is 20.62% and it represents the percentage of misclassified units. This index represents the ratio between the number of observations out of the principal diagonal and the total number of observations.

4.3 Multilevel Mixture Factor Model with individual level covariates

In the educational and psychological context, the multilevel mixture factor models are useful to individuate the latent classes at group level and classify the level one units according to the latent structure. In our application, we introduce the individual level covariates to describe the home context. In this way, we analyze the covariate effects on the factor score means and how the student reading motivation is influenced by parental beliefs. The latent structure at group level has not been modified (table 4.8), indeed the best solution is the model with seven latent classes.

Figure 4.3 reports a scheme of the multilevel mixture factor model with the lower level covariates. The individual covariates have been selected on the base of the educational and psychological literature about reading motivation. The covariates inserted at the beginning of the analysis were: sex, nationality, number of books at home, parental education, perceived socio economic status, library frequency use, lending books from library, children ability (read sentences and write some words), children statement (like being in school), parental behavior towards reading and parental statements (importance of reading and if reading is a duty). The student variables have been selected to understand the aspects directly related to the student behavior and attitudes. The family influences on children motivation have been described with different variables and the parental attitude and behav-

	Var	1	2	3	4	5	6	7	8	9	10	11
1	I Enjoy reading	.										
2	Reading is boring	1.98	.									
3	Book as a present	0.63	0.74	.								
4	Well for future	**11.84**	**11.15**	1.23	.							
5	Talk with friends	2.56	0.45	**5.81**	0.64	.						
6	Talk with family	**4.48**	2.58	0.29	0.61	**6.03**	.					
7	Reading for info	2.46	**6.42**	0.44	0.04	1.65	1.75	.				
8	Read Aloud	**13.56**	0.29	**9.39**	**6.58**	0.13	0.04	0.52	.			
9	Reading is easy	**22.84**	**14.22**	2.39	**59.48**	0.73	3.68	**6.59**	**21.48**	.		
10	Not as well as other	**18.42**	0.15	**4.54**	1.04	1.69	0.63	0.19	1.27	1.18	.	
11	Read slower than others	**14.42**	0.77	3.29	0.93	**8.11**	2.82	0.25	0.22	2.86	1.34	.

Table 4.5: Bivariate residual for the Multilevel Mixture Factor Model without covariates

Term	Estimates	S.E.	z-value	p-value
Enjoyment of Reading	0.55	0.05	10.23	1.4E-24
Value of Reading	0.15	0.02	7.31	2.6E-13
Self Concept as Reader	0.40	0.05	7.52	6.1E-14
Enjoyment of Reading - Value of Reading	0.13	0.01	9.91	3.7E-23
Enjoyment of Reading - Self Concept as Reader	0.20	0.02	9.28	1.7E-20
Value of Reading - Self Concept as Reader	0.03	0.01	3.10	0.002

Table 4.6: Between Variance and Covariance for the Latent Factor in the Multilevel Mixture Factor Model without covariates

ior are two examples of these variables. Not all the variables were significant and for this reason the assumptions on the reading motivation context could be verified at all. The selected covariates are: sex, number of books at home, parental education, perceived socio economic status, library frequency use, children ability (read sentences) and parental statements (importance of reading).

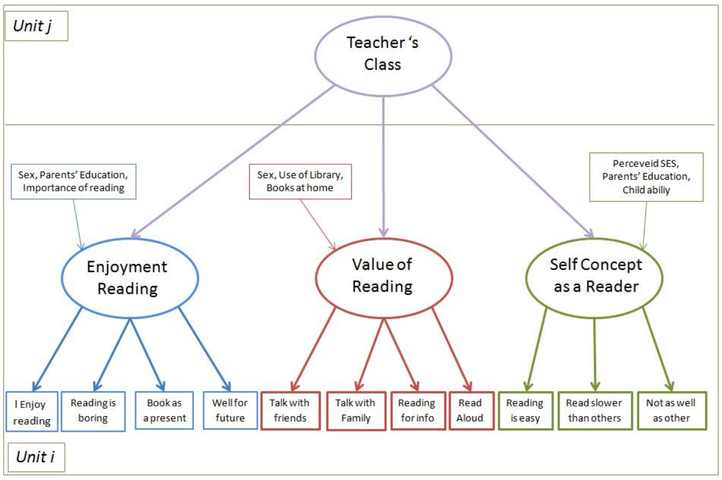

Figure 4.3: Multilevel Mixture Factor Model with lower level covariates

In table 4.9 the factor score means and the size of the latent classes have been reported. This situation is also represented in figure 4.4. By comparing the tables 4.4 and 4.9, it is possible to note that the group level structures are quite similar. The explicative variables affect the factor scores mean and not the class membership.

Looking table 4.9, the 4% of the classrooms have been classified as *Very Higly Motivated*, because of the highest values assumed both the factors: enjoyment of reading and self concept as reader dimension. The classrooms belonging to the *High Motivated* class (15%) are characterized by students that

		Modal							
		Very Highly Motivated	High Motivated	Motivated Bad Reader	Medium Motivated	Unmotivated	Highly Unmotivated	Very Highly Unmotivated	Total Total
Probabilistic	Very Highly Motivated	28.32	**3.84**	0.47	**2.63**	0.02	1.02	0.00	36.31
	High Motivated	**2.15**	40.02	0.83	**3.86**	**2.52**	**2.12**	0.12	51.61
	Motivated Bad Reader	0.56	0.70	9.45	0.01	0.00	0.00	0.00	10.73
	Medium Motivated	**4.17**	**3.46**	0.22	29.29	0.28	0.40	0.03	37.85
	Unmotivated	0.03	1.23	0.00	0.37	23.11	1.71	1.59	28.05
	Highly Unmotivated	0.77	1.67	0.03	0.15	1.85	17.73	0.01	22.22
	Very Highly Unmotivated	0.00	0.07	0.00	0.69	1.21	0.02	9.24	11.24
	Total	36	51	11	37	29	23	11	198

Table 4.7: Classification table for the Multilevel Mixture Factor Model

Number of Classes	Log Likelihood	BIC N=3158	BIC N=198	Number of Parameters
3	-40195	81061	80823	82
4	-40158	81020	80771	86
5	-40130	80997	80736	90
6	-40098	80965	80693	94
7	-40080	**80962**	**80678**	98
8	-40077	80989	80693	102

Table 4.8: Fit indexes for the Multilevel Mixture Factor Model with individual level covariates

see reading as an enjoyable ($ER = .50$) and valuable activity ($VR = .39$). The *Motivated Bad Reader* class (27%) represents the classrooms that have a positive disposition towards reading ($ER = .13$). They have some problems related to their own ability as reader ($SCR = -.24$). 24% of the classrooms belongs to the medium motivated class because the factor mean values of this class are all close to zero. The *Unmotivated* class (10%) is composed of classrooms that have negative values for the two-thirds of the dimension measuring motivation. These classrooms do not see reading as an enjoyable activity and they are also bad reader. The 15% of the classrooms has been classified as *High Unmotivated*. The students of these classrooms do not see reading as an enjoyable ($ER = -.50$) and valuable activity ($VR = -.21$). The *Very Highly Unmotivated* (5%) class is characterized by the negative values of all the factor score means. This class identify the classrooms that do not have any interest for reading.

In table 4.10 the parameter estimates for the explicative variables for the enjoyment of reading dimension are reported. The results obtained seem to support the theoretical context on the student reading motivation (in brackets the standard errors are reported). The lower level covariates inserted to explain the enjoyment of reading dimension are: sex, parental education and attitude towards reading. Sex (girl) has a positive effect (*0.22 (0.02)*) on motivation. This fact confirms the theoretical background concerning the positive impact of sex on reading motivation. In many studies, it has been proved that girls appear to be more motivated than boys. These results confirm the importance of parental education in the process of motivation development. High level of parental education gives a positive contribution on student reading motivation. The coefficients are all positive: University or Higher (*0.14 (0.04)*), Post-secondary (*0.05 (0.05)*) and Upper-secondary (*0.03 (0.03)*), and this means that a positive effect on the factor score mean

	Very Highly Motivated	High Motivated	Motivated Bad Reader	Medium Motivated	Unmotivated	High Unmotivated	Very Highly Unmotivated
Enjoyment of Reading	.50	.47	.13	-.01	-.28	-.50	-.31
Value of Reading	-.01	.39	-.03	.06	.28	-.21	-.48
Self Concept as Reader	.74	.03	-.24	.11	-.22	-.07	-.35
Size	.04	.15	.27	.24	.10	.15	.05

Table 4.9: Class mean factor scores for the Multilevel Mixture Factor Model with lower level covariates

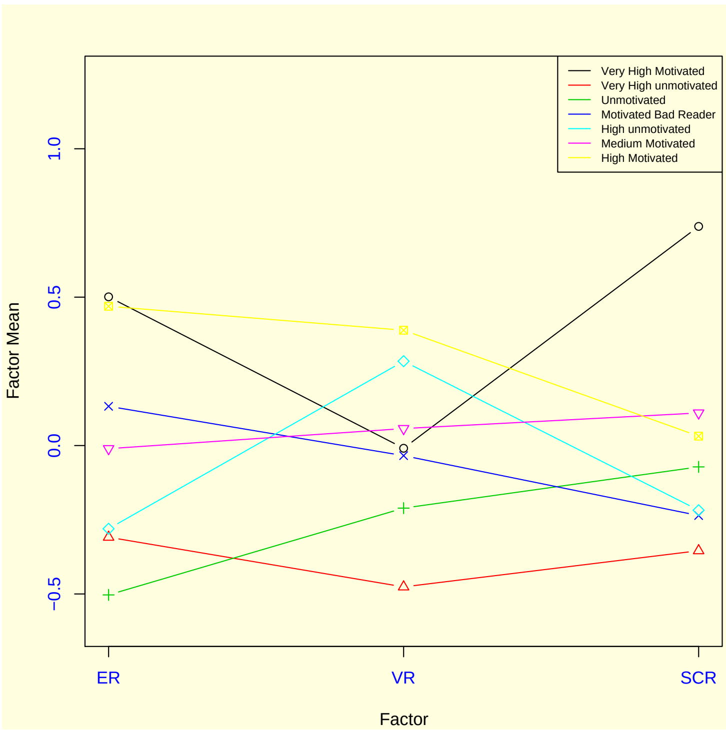

Figure 4.4: Plot class mean factor scores for the Multilevel Mixture Factor Model with lower level covariates

could be observed compared with low level of parental education (Lower-secondary (*-0.10 (0.03)*), Primary or lower (*-0.12 (0.08)*)).

The covariates describe the importance addressed by parents to reading activities at home. Previous studies outline that a well-disposed familiar context has an incremental effect on student motivation. Parents with positive disposition towards reading increase children motivation (Agree a lot *0.18(0.03)*). The effect on motivation is negative (Disagree a lot *-0.18 (0.05)*) in families where parents do not see reading as an important activities.

For the value of reading, the covariates selected are: sex, frequency of using library and number of books at home. The parameter estimates are reported in table 4.11. These variables take into account the results obtained in previous studies. The importance of living in a challenging environment and the use of library are examples of the positive influences on value of reading. Sex (girl) has a positive effect (*0.06 (0.01)*) on the value of reading, because it has been proved that girls give more consideration than boys to reading. Analyzing the library use, it is possible to note the effect of a frequent use of the library (At least once a week (*0.09(0.02)*)) on the reading value, compared with situation where students go few times to the library (Never or almost never (*-0.09(0.02)*)).

A challenging environment has an effect on the value of reading. The coefficients related to $101 - 200$ and more than 200 books at home assume positive values (*0.03* and *0.03*), while the coefficient related to $0 - 10$ books is negative (*-0.10 (0.02)*). This shows a negative effect on the reading value. It could be considered as an empirical proof of the importance of living in a challenging environment for a child.

The covariates used to explain the awareness as reader of the students are: perceived Socio Economic Status (SES), parental education and pre-entrance ability (read some sentences). In some studies, it has been proved that a good economic situation and cultural environment have a positive effect on the children awareness as reader. Other studies have focused their attention to the importance of the pre-entrance abilities. This covariate has a double function. It could be seen as a proxy of the interest towards reading, but also as a predictor of the awareness as reader. Perceived SES is significant (p-value=*0.00035*), but the interpretation is not so easy. The coefficient associated to *Very well-off* and *Not at all well-off* are in contrast with the theoretical background. This depends on the fact that it is not a measure of the effective Socio Economic Status because it has been asked a personal opinion to parents about their own financial situation. Parental education has a positive effect on children self concept as reader (University or Higher 0.27(0.04)), compared with low level of parental education (Primary or lower $-0.29(0.08)$). In the *P.I.R.L.S.* survey many variables have been inserted

Factor	Variables	Terms	Coef.	S.E.	Z-value	p-value	Wald(0)	df	p-value
Enjoyment of Reading	Sex	Girl	.22	.02	-12.06	1.70E-33	145.45	1	1.70E-33
		Boy	-.22	.02	12.06	1.70E-33			
	Parental Education	University or Higher	.14	.04	3.26	.001	2.70E+01	4	2.00E-05
		Post-secondary	.05	.05	1.00	.32			
		Upper-secondary	.03	.03	1.09	.28			
		Lower-secondary	-.10	.03	-3.12	.002			
		Primary or lower	-.12	.08	-1.5	.13			
	Importance of Reading	Agree a lot	.18	.03	6.15	7.50E-10	4.31E+01	3	2.30E-09
		Agree a little	-.01	.03	-.20	.84			
		Disagree a little	.005	.04	.12	.91			
		Disagree a lot	-.17	.05	-3.75	.0002			

Table 4.10: Parameters estimates of the latent factor Enjoyment of Reading for the MMFM with lower level covariates

Factor	Variables	Terms	Coef.	S.E.	Z-value	p-value	Wald(0)	df	p-value
Value of Reading	Sex	Girl	.06	.01	5.95	2.80E-09	3.54E+01	1	2.80E-09
		Boy	-.06	.01	-5.95	2.80E-09			
	Frequency of library use	At least once a week	.09	.02	4.94	8.00E-07	37.51	3	3.60E-08
		Once or twice a month	.02	.02	1.4	.17			
		A few times a year	-.02	.02	-.81	.42			
		Never or almost never	-.09	.02	-5.10	3.30E-07			
	Books at home	0-10	-.10	.02	-4.56	5.20E-06	21.14	4	.0003
		11-25	.03	.02	1.50	.14			
		26-100	.02	.02	1.40	.16			
		101-200	.03	.02	1.20	.23			
		more than 200	.03	.02	1.52	.13			

Table 4.11: Parameters estimates of the latent factor Value of Reading for the MMFM with lower level covariates

to analyze the pre-entrance ability. We consider the reading activities. In table 4.12, it is possible to see the positive effect on the student awareness as reader (Very well 0.19(0.05)), compared with situations where children were not able to read before starting school (Not at all $-0.18(0.03)$).

The individual level covariates (student and family) are useful to describe the effects of the home context on the reading motivation. In this way, it is possible understand which are the factors that influence the students reading motivation.

One of the most important results concerns with sex difference. Girls appear to be more motivated than boys. Sex differences emerges in two different aspects: enjoyment of reading and value of reading. This difference could be connected to the different approach that girls have towards reading, because they like reading more than boys at this age because reading is an interesting activity. For the value of reading, the explanations could depend on the high importance that girls give to reading in their day life. Reading is a way to interact with other children and for this reason in our application they appear to be more motivated than boys. Differently, boys do not give a high consideration to reading, because they are attracted by other types of pastimes.

Another result is the importance of growing in a high educated context. A high level of parental education has a positive effect on two aspects of reading motivation: enjoyment of reading and self concept as reader. The results show how owning a superior parental qualification from *Upper Secondary* to *University* increase the factor score mean. Low qualification produces a negative effect on both the aspects of motivation. Parental education has a positive effect on the attitude of children. An involving context increases both the children reading enjoyment (parents with high education have a good disposition towards reading activities) and the awareness as reader. Table 4.11 highlights the positive influence of being enrolled in reading activities (e.g., library use) and living in a involvement context (e.g., books at home). A child absorbs in this way, the significance of reading for the future, increasing their actual motivation. The last result underlines the positive role of the pre-entrance abilities into the school system (e.g., reading some sentences). When children are able to read before starting school, it has been proved that they are more autonomous in reading and this has a positive effect on the perceived pupil ability.

Table 4.13 represent the classification table (paragraph 2.5.1). The classification error for the Multilevel Mixture Factor Model with lower level covariates on factor score is 18.62%. Compared with the classification error of the multilevel mixture factor model (paragraph 4.2), it is possible to appreciated the effect of the introduction of the individual level covariates. This has a

Factor	Variables	Terms	Coef.	S.E.	Z-value	p-value	Wald(0)	df	p-value
Self Concept as Reader	Perceived SES	Very well-off	-.13	.07	-2.00	.05	2.08E+01	4	.00035
		Somewhat well-off	.010	.04	2.30	.021			
		Average	.10	.03	3.74	.00018			
		Not very well-off	.10	.04	2.39	.017			
		Not at all well-off	-.17	.06	-2.67	.0077			
	Parents Education	University or Higher	.27	.04	6.175	6.60E-10	4.71E+01	4	1.40E-09
		Post-secondary	.00	.05	-.12	.91			
		Upper-secondary	.09	.03	3.00	.0028			
		Lower-secondary	-.06	.03	-1.88	.06			
		Primary or lower	-.29	.08	-3.87	.00011			
	Child ability read some sentences	Very well	.19	.05	4.08	4.60E-05	48.4966	3	1.70E-10
		Moderately well	.07	.03	2.34	.02			
		Not very well	-.08	.03	-2.90	.0037			
		Not at all	-.18	.03	-6.64	3.20E-11			

Table 4.12: Parameters estimates of the latent factor Self Concept as Reader for the MMFM with lower level covariates

	Modal								
		Very Highly Motivated	High Motivated	Motivated Bad Reader	Medium Motivated	Unmotivated	High Unmotivated	Very Highly Unmotivated	
Probabilistic	Very Highly Motivated	7.76	0.00	0.00	0.08	0.00	0.23	0.06	8.13
	High Motivated	0.00	7.73	0.87	0.78	0.01	0.01	0.00	9.39
	Motivated	0.00	1.01	22.67	2.03	0.64	2.36	0.01	28.72
	Medium Motivated	0.01	0.24	0.58	45.24	0.22	**5.67**	1.87	53.82
	Unmotivated	0.00	0.00	0.40	0.87	15.86	1.54	1.29	19.97
	High Unmotivated	0.13	0.03	1.48	**5.95**	1.03	36.52	2.42	47.56
	Very Highly Unmotivated	0.09	0.00	0.00	2.05	1.24	1.66	25.35	30.40
Total	8	9	26	57	19	48	31	198	

Table 4.13: Classification table for the Multilevel Mixture Factor Model with lower level covariates

	Within Variance			Between Variance		
	With Covariates	Without Covariates	Percentage of reduction	With Covariates	Without Covariates	Percentage of reduction
Enjoyment of reading	0.48	0.55	-13.60%	0.15	0.20	-24.78%
Value of reading	0.14	0.15	-9.83%	0.08	0.06	46.97%
Self concept as reader	0.40	0.40	-1.64%	0.13	0.11	15.80%

Table 4.14: Percentage reduction of Within and Between Variance with and without covariates for the multilevel mixture factor model

positive effect on the classification performance of the model. The reduction of misclassification is almost the 2%. It is possible to see from table 4.13, that the classification errors are related to the *motivated bad reader* and *medium motivated* classes.

Table 4.14 reports the comparison of the variances between the two models described above. The variance has two components: between (group mean variance) and within (mean of the group variances). We want to show how the variance reduction is related to the individual level covariates.

The most interesting result is related to the variance within. The within variance of the enjoyment of reading dimension, in the model without covariates is equal to *0.55* and the variance within is *0.48* after the introduction of the covariates. The variance reduction is negative (*-13.60*%). For the other two latent variables (value of reading and self concpet as reader), it is possible to observe a reduction of the variance within (*-9.83*%) and (*-1.64*%). This reduction means that the introduction of covariates affect the variability within the groups. Those are more homogeneous after the introduction of covariates. In the second part of the table, the between variance reduction is reported. It is possible to see that the only interesting result is related to the enjoyment of reading dimension. The Between variance in the model without covariates is equal to *0.20*, while the one with the covariates is equal to *0.15*. The percentage reduction is equal to *24.78*%. This means that the difference between groups dicrease with the introduction of covariates.

Table 4.15 compare the classification obtained with the models described above. The latent classes obtained for the two classification are not equal.

The elements on the principal diagonal should be the elements classified in the same way. A difference could be noted between the *High Unmotivated* class in the model without covariates and the *Unmotivated* class in the model with covariates. The introduction of the covariates decrease the value of the factor means. The same situation happens for the classes *Very highly unmotivated* in the model without covariates, and the *High Unmotivated* in the model with covariates.

4.4 Multilevel Mixture Factor Model with individual and group level covariates

This section describes the effects of the high level covariates on the model classification. The aim of this section is to explain the latent class membership in order to find the different effects of the teachers on the class motivation using the group level covariates. These covariates investigate different aspects as: the use of resources and texts type in class, the way of teaching new words to enlarge student vocabulary, the way of involving and evaluate student and teacher specialization. In the teacher questionnaire, a lot of different aspects have been investigated. Some of these variables have been selected according to their significance and usefulness to describe the latent class membership. The variables describing the teachers characteristics have not been inserted, because we want to trace a profile of teacher in function of their ability and way of teaching. In figure 4.5 the situation above described has been represented.

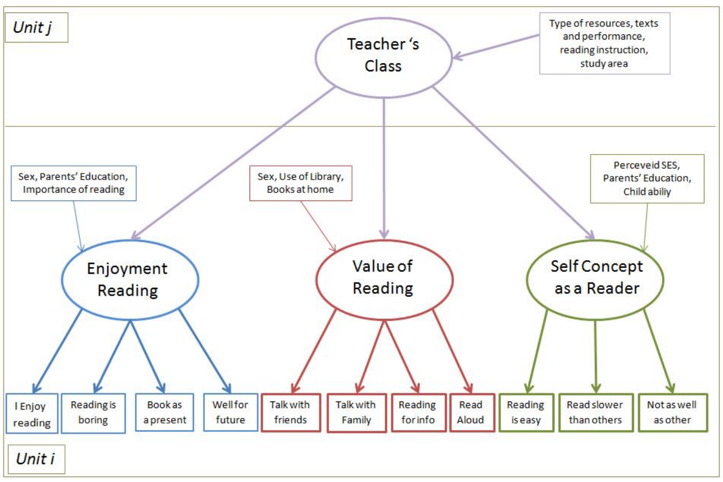

Figure 4.5: Multilevel Mixture Factor Model with high level covariates

		With covariates							
		High Motivated	Motivated Bad Reader	Medium Motivated	Medium Motivated Bad Reader	Unmotivated	High Unmotivated	Very High Unmotivated	Total
Without covariates	Very Highly Motivated	**5**	2	1	0	1	0	2	11
	High Motivated	1	**20**	7	5	0	0	3	36
	Medium Motivated	1	5	**30**	12	3	0	3	54
	Motivated Bad Reader	1	1	0	**32**	1	0	0	35
	Unmotivated	0	3	4	2	**1**	0	9	19
	High Unmotivated	0	0	5	5	17	**1**	2	30
	Very Highly Unmotivated	0	0	0	1	3	8	**0**	12
	Total	8	31	47	57	26	9	19	197

Table 4.15: Comparison between student class membership for the multilevel mixture factor model without and with lower level covariates on factor score

Number of Classes	Log Likelihood	BIC N=3158	BIC N=198	Number of Parameters
2	-40208	81095	80855	83
3	-40157	81066	80800	92
4	-40105	81037	80744	101
5	-40068	81037	**80718**	110
6	-40050	81073	80729	119

Table 4.16: Fit indexes for the Multilevel Mixture Factor Model with individual and group level covariates

In this section, we show how the composition of the latent classes and so the classification of the changes after the introduction of the high level covariates. This is a problem that affect this kind of models. After the introduction of covariates as in this case affecting the class membership, the latent structure and the results obtained could be very different compared with the results previously obtained. This is still an open question in the field of the latent class models that many researchers are trying to solve to assure the stability of the results. The model is the same that has been introduced in paragraph 2.3. The reference equation are 2.25, 2.26, 2.27 and 2.18. In table 4.16, the fit indexes have been reported and the best solution is the model with five latent classes. The *BIC* with sample size equal to the number of groups equal to **80718**.

In table 4.16, the latent class means for the three factor have been reported. This situation could be visualized in figure 4.6. The factor score means have been modified after the introduction of the higher level covariates. The *High Motivated* (*5*%) class represents the classrooms that have the highest values for two dimensions: enjoyment of reading (*.40*) and self concept as reader(*.63*). *17.32*% of the classrooms have been classified as *Motivated*, because the enjoyment of reading (*.38*) and the value of reading (*.31*) assumes positive values. For the *Medium Motivated* class (*43.3*%, the factor score means are close to zero, because this is the benchmark class. *15.65*% of the population represents the *Unmotivated* class because the mean values for the enjoyment in reading and the self concept as reader are negative (respectively -*.31* and -*.19*). The *High Unmotivated* class (*19*%) is characterized by the lowest value for all the dimension measuring motivation.

In this section, we describe the effects of the higher level covariates on the teacher class membership. One of the purposes of this model is the description of the way of teaching and behaving in classrooms. The teachers (class) have been classified according to the factor score means of the three different

	High Motivated	Motivated	Medium Motivated	Unmotivated	High Unmotivated
Enjoyment of Reading	.40	.38	.01	-.31	-.49
Value of Reading	-.05	.31	-.07	.17	-.36
Self Concept as Reader	.63	-.04	-.17	-.19	-.23
Size	.05	.17	.43	.16	.19

Table 4.17: Class mean factor scores for the Multilevel Mixture Factor Model with lower and higher level covariates

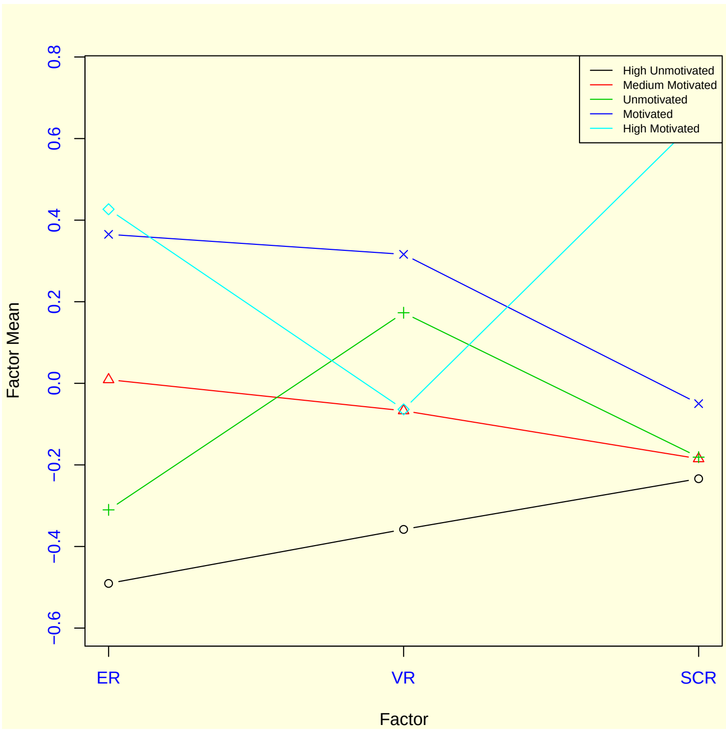

Figure 4.6: Plot class mean factor scores for the Multilevel Mixture Factor Model with lower and higher level covariates

aspects of the reading motivation. We highlight the teacher influences on the reading motivation after regulating for student and family characteristics. In this way, it is possible to describe the pure teacher effect on the student motivation.

Table 4.12 reports the parameter estimates for the group level covariates. These variables have been considered as continuous, although they were categorical to make the interpretation easier. We are hypothetical assuming that there is an unobserved continuous dimension that measure these aspects related to teaching. The variables as the use of textbooks, the text types and the way of teaching the meaning of new words assume four modalities (every day or almost every day, once or twice a week, once or twice a month and never or almost never). The categories for the variable relative to the number of evaluations are: once a week, once or twice a month, once or twice a year and never. The variable related to the area of specialization of teachers assumes only three categories: not at all, overview or introduction to topic and area of emphasis. Attention has been directed only on the variables that have been found significant. In round brackets, the p-values associated to the exponential value of the coefficient have been reported below.

For the classrooms belonging to the *High Unmotivated* class, only two variables are significant. This means that they give a contribution to trace the profile of a teacher that positively affect class motivation. The coefficient related to the use of textbooks is positive (*4.7 (0.04)*). This result indicates a negative effect on the probability of belonging to the *High Unmotivated* class. A decrement of the number of times of using textbooks in classroom, increases the probability to belong to *High Unmotivated* class. We can conclude that using often textbooks in classrooms increase the student reading motivation. The coefficient related to the form of the Evaluation is positive (*5.2 (0.0006)*). This means that a decrement of the number of oral reports in classrooms, increases the probability of belonging to the *High Unmotivated* class. A frequent oral evaluation increases the student motivation. Being evaluated about reading is seen by students as a possibility to be rewarded for their work.

For the *medium motivated* class, we found that almost all the variable were significant. The coefficient associated to the use of textbooks is *7.25 (0.007)*. This mean that decreasing the number of times of using textbooks in classrooms, has a positive effect on the probability to belong to the *medium motivated* class. Using textbooks during lessons, increases the classroom reading motivation. The use of textbooks help students to see reading as a desirable activity. The coefficient related to the *Text Types* is *0.46 (0.002)*. This means that decreasing the number of times of using books with long chapter in classrooms decreases the probability to belong to the *medium* *motivated* class. Using books with long chapter increase motivation. Student should be more concentrated during reading activity also because they are in contact with a more challenging environment. The coefficient related to the way of teaching new words is *0.32 (0.01)*. This result highlights that decreasing the number of times of teaching new words in classrooms decreases the probability to belong to the medium motivated class. The reason is related to the fact that enlarging student vocabulary could make them more autonomous in reading in the future. The coefficient related to the form of evaluation is *3.15 (0.02)*. This means that a decrease in the number of oral report increases the probability of belonging to the *medium motivated* class. Frequent evaluations increase the student motivation. It is an opportunity for students to be rewarded for their work.

Considering the *unmotivated* class, the only variable that was found significant was the teacher competences. The coefficient associated is *0.19(0.0006)*. This means that teacher with more abilities in child language development decreases the probability to belong to the *unmotivated* class. The explanation is related to the fact that more a teacher is able to develop child language more the students will appear motivated, because teachers with this kind of specialization could help students in a better way.

For the *high motivated* class, three variables were significant. The interpretation is not so easy. A reason could be assessed to the small number of classrooms classified as high motivated. The first variable is related to the use of text types in classrooms. The coefficient is *3.81 (0.02)*. The meaning is that a decrease of the number of times of using textbooks increases the probability to belong to *high motivated* class. This is in contrast with the results obtained previously. For the variable related to the way of teaching new words, the coefficient is *4.70 (0.005)*. The interpretation seems to be in contrast with previous results. While for the variable related to the teacher competences, the coefficient is equal to *3.78 (0.05)*. The interpretation is that an increment of the teacher abilities in child language development increases the probability to belong to *high motivated* class.

We obtained a lot of important results after the introduction of the higher level covariates. The teacher effects on classrooms motivation is one of the principal purposes of this analysis. The teacher variables investigate different aspects of teaching. The use of textbooks has a positive effect on classroom reading motivation, because it is a way to develop student abilities that are strictly connected to reading motivation. Using often books with long chapter increase student motivation. The theoretical context suggests that a challenging environment for students, has a positive effect on motivation. For such reasons, promoting in class this climate is an important dimension that should be taken into account in the analysis of teacher influences. Positive

Class	Variable	Coef.	Exp Coef.	S.E.	z-value	p-value
	Textbooks	1.55	4.70	.75	2.06	**.039**
	Text Types	-.33	.72	.27	-1.20	.23
High Unmotivated	Teaching new words	-.04	.97	.38	-.09	.93
	Oral Evaluation	1.65	5.19	.48	3.44	**.00059**
	Specialization	-.38	.68	.38	-1.03	.31
	Textbooks	1.98	7.25	.74	2.70	**.007**
	Text Types	-.78	.46	.25	-3.15	**.0016**
Medium Motivated	Teaching new words	-1.15	.32	.43	-2.65	**.008**
	Oral Evaluation	1.15	3.15	.47	2.42	**.016**
	Specialization	.40	1.49	.34	1.18	.24
	Textbooks	-2.78	.06	2.26	-1.23	.22
	Text Types	.17	1.18	.41	.41	.68
Unmotivated	Teaching new words	-.56	.57	.75	-.75	.45
	Oral Evaluation	-.87	.42	.80	-1.09	.28
	Specialization	-1.65	.19	.48	-3.43	**.00061**
	Textbooks	-.08	.92	1.02	-.08	.93
	Text Types	-.39	.67	.34	-1.17	.24
Motivated	Teaching new words	.20	1.22	.49	.41	.68
	Oral Evaluation	-1.65	.19	1.11	-1.48	.14
	Specialization	.31	1.36	.40	.78	.44
	Textbooks	-.66	.52	1.23	-.54	.59
	Text Types	1.34	3.81	.56	2.39	**.017**
High Motivated	Teaching new words	1.55	4.7	.56	2.79	**.0054**
	Oral Evaluation	-.27	.76	.77	-.36	.72
	Specialization	1.33	3.78	.68	1.96	**.05**

Table 4.18: Parameters estimates of the covariates on class membership for the Multilevel Mixture Factor Model with lower and higher level covariates

		Modal					
		High Unmotivated	Medium Motivated	Unmotivated	Motivated	High Motivated	Total
Probabilistic	High Unmotivated	33.33	**4.09**	0.32	0.00	0.00	37.73
	Medium Motivated	**2.44**	79.28	1.48	**2.73**	0.02	85.95
	Unmotivated	0.21	**2.74**	26.70	1.22	0.02	30.89
	Motivated	0.00	**3.68**	1.50	28.91	0.14	34.23
	High Motivated	0.02	0.22	0.00	0.14	8.82	9.19
	Total	36	90	30	33	9	198

Table 4.19: Classification table for the Multilevel Mixture Factor Model with lower and higher level covariates

effects have been found also for the way of teaching the meaning of new words. In this way, teachers involve students in reading. The aim is to increase in the future, the students autonomy as readers. This increases their ability as readers. Teaching the meaning of new words encourages students to take more seriously reading, in order to face their lack of knowledge. Another important evidence is related to the evaluation methods. More often a student is evaluated (with oral reports) by teachers, more motivated a student will appear. The reason has been assessed to the fact that students need to be evaluated. In this way, they could assign a value to their efforts and way of studying. Students feel more motivated to read, because they could demonstrate their ability in the comparison with the other students. This has a positive effect on the motivation of the entire classroom. Teacher background and competences has been analyzed in this work. More a teacher is able in child language development activities, more a student will appear as motivated. The reason is related to the fact that teaching ability increase the student motivation and involvement in classroom.

Table 4.19 is the classification table. The percentage of misclassified units is equal to 10.59%. Compared to the previous model, the reduction (8%) depends on the teacher covariates. This reduction could be explained by the introduction of the higher level covariates, becasuse they affect the class membership. The aim is to classify teachers according to the latent structure at lower level and the results have shown that they increase the model classification power. In bold the errors commited by the model have been highlighted. Most of the classification errors are related to the *medium motivated* class.

Chapter 5

Conclusion

During the past years, thanks to the increment of the level of competitiveness and the necessity to make comparison between countries, it has been possible to see a spread of many different surveys to measure the student skills (). The principal aim of these surveys is the country classification according to the student abilities, as mathematics, reading, science and technology. This purpose is noble because it is possible to describe the different actors involved in the school system, as teachers, family and students. Using the results and the evidences highlighted by these surveys, the institutions could take decisions on the policies of the educational system, considering the actual situation of their country. Often in these studies, students achievement receives more attention than other concepts. There are a lot of factors that are useful to describe the education system and that have an influence on the student skills. These should be used to measure the development level of a country. In our work, taking into account the psychological and educational literature, efforts have been concentrated to study the students motivation. Particular attention has been addressed to the aspects of reading related to motivation.

The importance of motivation has been discussed in paragraph 3.1. The different approaches proposed in the past years to measure and describe motivation have been presented. Student motivation is composed of three dimensions. Each of them measure different aspects. The *enjoyment of reading* dimension (connected with four items) investigates the student thoughts about reading, and if they enjoy in these activities. The *value of reading* dimension (connected with four items) measures the number of times in which a student dedicates time to reading outside of school. These items have been selected because of the importance of the dimension that we want to analyze. If a student decide to direct attention to reading in his/her own free time, it means that reading is an interesting and important activity for him/her.

The last analyzed dimension is the *self concept as reader* (connected to three items). This measures the student awareness as reader. In this life period students "'*have learned to read and are reading to learn*"'.

The theoretical statistical background used in this applicative work has been focused on the generalized latent variable modeling framework. Particular attention has been addressed to the multilevel mixture factor models that have been developed in recent years. This work is an attempt to unify and extend the latent variable models and to apply these models in the educational context. In the past literature on this topic, the models that have been proposed dealt with a single type of latent variables (continuous or categorical). In this applicative context, both the latent class models and the multilevel data structure have been combined. The factor analysis is referred to the models where all the latent variables - called factors - are continuous. In the mixture factor analysis models, both the continuous and the categorical latent variables have been considered. In this work, attention has been focused on the comparison between the results obtained with the mixture factor model and the multilevel mixture factor model. These two models differ for the classification purposes. The aim of the mixture factor model was the **classification of students** in different level of motivation according to their motivation (measured with three factors). In the multilevel mixture factor model the hierarchical data structure has been included and the mixture component has been shifted to the group level (teacher). The aim of this model specification is the **classification of teachers** according to the classrooms reading motivation. Attention has been addressed on the effects that teachers have on the student motivation.

The usefulness of the multilevel specification is the possibility to analyze the motivation contemporaneously at the student and teacher level. The focus of the multilevel mixture factor model application is especially on the teacher (group) level. In this way, it is possible to individuate the possible differences and similarities between teachers and their way of teaching. The models represent a very interesting instrument to classify subjects. To analyze and classify both the lower and the higher level observations, it is necessary to introduce the covariates in order to describe the parental and teacher role in the process of motivation development. The added value of this application for both the theoretical and applicative point of view, is the introduction of some explicative variables. Usually in the education literature student characteristics and home/school environments have been used to describe the impact on the achievement of students. We decide to use this kind of covariates to explain the effect on motivation. These covariates are useful to trace a profile of a motivate student and to outline a favorable home/school environment for students. In the mixture factor model, the covariates are related to the individual level. We want to explain the factor score mean differences between the latent classes. In this way, we outline the effects of the student characteristics and the home context. In the multilevel mixture factor model, the group level covariates have been inserted on the class membership. We want to individuate the influences that increase or decrease the probability of belonging to a motivated or unmotivated classes and to trace the teacher profile.

In chapter 1, we presented the use of the latent variables in the field of the applied science. The aim was to show the way in which many aspects in human life are unobservable and unmeasurable, and the important role covered by the latent variables. This topic has been investigated from the beginning of the 20^{th} century with Pearson, Galton and Spearman (1904). For these reasons, section 1.1 is dedicated to a brief historical introduction on the advanced research in the latent variables framework. In the last twenty years with the development of softwares, it has been possible to see a huge increment of the literature on this framework. Fields as psychology, education, marketing, biology and medicine are examples of fields where latent variables are often used.

In chapter 2 the technical details of the generalized latent variable modeling framework have been reported. The existent literature on the mixture factor model and the multilevel mixture factor model have been presented. The mixture factor model (Lubke and Muthén (2005)) and the multilevel mixture factor model (Vermunt (2007b)) have been introduced as part of the generalized latent variable models. Attention has been addressed to the models with continuous and categorical latent variables. The aim is the analysis of the differences and similarities between these models. The last part of the chapter is dedicated to the model formalization, the likelihood specification, the estimation procedures and the posterior analysis.

Chapter 3 and 4 are dedicated to the results of the models introduced in chapter 2. In chapter 3, the first section is dedicated to the results of the exploratory factor analysis (section 3.1.1). We choose the items to measure motivation in the best way. In chapter 3, the results of the mixture factor model have been reported. Paragraph 3.3.1 is dedicated to the simple mixture factor model, where no covariates have been introduced. We classify students according to the latent structure at lower level. The results confirmed the hypothesis on the population heterogeneity. The values of the fit indexes indicated as the best solution the model with four latent classes (*BIC=81071* and *AIC=80707*).

These results highlight the different role of items on the latent factor structure measuring reading motivation (table 3.8). For the *enjoyment of reading* dimension, the highest contribution is related to the item *Enjoy in* *reading*, for the Value of Reading is *Talk with family*, and for the Self Concept as Reader is *Not as well as other*. These models subdivide the population in four different subpopulations (latent classes). These subpopulations are different because of the values assumed by the factor score mean measuring motivation. We introduce the individual level covariates to describe the differences between the latent classes and how these differences are related to the students and the home environment. The covariates modify the classification obtained with the previous model. For the mixture factor model with covariates, the fit index indicate as the best solution the model with five latent classes. More details have been reported in section 3.3.2.

The individual level covariates describe the influences on motivation by parental behavior and attitudes towards reading. We following summarize the obtained results. An Italian girl who likes being at school, go to the library and that was able to read sentences and write some words before starting school, identify the profile of a motivated student. The best environment that positively influence the development of motivation is a family where the level of parental education is high, reading is seen as an important activity and they want to promote a challenging environment.

The covariates modify the variability and the classification of the model. A reduction of the internal differences for students in the same latent classes and an increment of the differences between groups is due to the introduction of the explicative variables. The covariates make the groups more homogeneous within them and more heterogeneous between them. The effect on the classification have been showed in paragraph 3.3.3. It corresponds to a reduction of the differences between the factor score means. Student differences are flattened out through the introduction of covariates.

The subsection 3.3.4 shows the role played by the reading motivation on the student ability. The plausible values describes the reading ability. We want to show the existent connection between the reading motivation and students ability. It has been highlighted that motivated report higher scores for the reading achievement than students classified as unmotivated.

Chapter 4 describes the results obtained with the multilevel mixture factor model (Varriale and Vermunt (2009)). The introduction is dedicated to the advantages of considering the multilevel structure in the model specification. Both the Intraclass Correlation Coefficients (ICC) and the Likelihood Ratio Test (LRT) values support the advantages of the hierarchical structure form this kind of data. This chapter is subdivided into three parts. The first is dedicated to the simple multilevel mixture factor model. The second and the third are dedicated to the description of the results obtained with the introduction of the individual and group level covariates. Each of these models have a different aim. The purpose of the multilevel mixture factor model is the classification of teachers according to the classroom motivation. This is possible because the mixture component has been shifted to the group level. With the lower level covariates, we described the factors affecting motivation. Teachers and the way of teaching have been described by the the group level covariates.

With the multilevel mixture factor model without covariates, we classify teachers according to the classes motivation. The fit indexes indicate as the best solution the model with *7* latent classes (this means that there are seven different subpopulations). In the multilevel context, some simulation studies (Lukočienė *et al.* (ress)) have shown that the best fit indexes is the BIC with sample size equal to the number of groups. More details have been reported in the paragraph 2.5 and in the work of Lukočienė and Vermunt (2009). The new specification takes into account the fact that students share the same environment. The classrooms classification has been reported in paragraph 4.2. The latent structure at group level is invariant after the introduction of the lower level covariates. The best solution is the model with *7* latent class.

One of the effects of the introduction of the lower level covariates is a decrement of both the fit indexes values and the misclassification error (-2%). The covariates increment the classification performance of the model. The aim is the description of the home context. One of the important results is sex difference. Girls result to be more motivated than boys during this period of life. There are essentially two reasons for this evidence. One is related to the benefit produced by activities concerning reading and learning and the concerns with the interactions with other children. The covariates illustrate the importance for a child to grow in a high educated context and to live in a challenging environment, because children appear to be more involved towards reading activities and learning. Another results underline the role of the pre-entrance abilities in the development of motivation. These have a positive effect on the actual abilities. They highlight the grade of parental involvement in children learning and this have an indirect effect on motivation. The introduction of the lower level covariates modify the variability structure. This means that the individual covariates reduce the variability within groups (making groups more homogeneous) and increase the classification power of the model.

Section 4.4 describes the results obtained with the multilevel mixture factor model. The teacher level covariates have been inserted. We want to describe teachers and the way of teaching. The fit indexes indicate as the best model the one with five latent classes. The high level covariates affect the teacher class membership and the latent class structure has been modified. The latent classes identify five different subpopulations. More details could be found in section 4.4.

The teacher covariates directly affect latent class membership. The goal of the introduction of these covariates, is to highlight the teacher practices which positively influence class motivation. The results outline the best teaching practice to increase student motivation. The results show the role played by the use of different kinds of textbooks during lessons. How often in which new words are taught and the number of oral evaluations are examples of these covariates. With the group level covariates, after regulating for family and students characteristics, we want to underline the natural teacher effect on the class motivation in order to compare different ways of teaching and find the best class practices.

Using textbook very often has a positive effect on the reading motivation because it is a way to develop student ability. Using books with long chapter and the type of books used during lessons has a positive effect. The reason is related to the fact that these practices could be seen as a challenge for students. The theoretical context suggests that challenge is a productive element in classroom. Positive effect has also been found for the teaching practices. In this way, teachers could involve students in reading, in order to make them more autonomous in reading, and encourage students to take seriously reading activities. The evaluation methods have a positive effect on motivation, because students often evaluated appear to be more motivated. The reasons are related to the fact that students need to be evaluated. In this way, they could give value to their efforts and their way of studying. They also demonstrate ability comparing their selves with the other students, influencing the motivation of the entire classroom. Teacher background and competences have a positive effect on motivation. The child language specialization of the teachers, has a positive effect on the student motivation.

Future perspectives for this work could be subdivided in two different areas: theoretical and applicative. The theoretical extensions of these models could be the creation of a sort of guideline to implement the multilevel mixture factor model. In order to simplify the classification procedures, or the inclusion of different structures of dependencies between observations in model implementation. As it is known, the multilevel mixture factor models are a particular case of the multilevel mixture regressions. These models are useful to investigate population heterogeneity in presence of hierarchical data structure. The question related to these models concerns with the effective ability of capturing the real data structure. This extension could investigate this aspect with some simulation studies where the data have been generated with a particular structure, and misspecifying the multilevel regression mixture by incorrectly fixing the random effects across clusters, we could examine the model results and to establish if the model fits the data or not.

With regards to the applicative perspective, the attention should be focused on the model comparison between Multilevel Mixture Factor and Mixture Factor model. In order to investigate using educational data, the class effect on motivation (what kind of effects are produced sharing the same environment or teachers) should appear clear in the contrast between these two models. Another extension could be focused on the intra-country comparison to understand which are the similarities between educational systems and the way of motivating students. Also which are the roles of families and teachers in different cultures.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, New York, NY.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–128.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Ames, C. and Ames, R. (1984). *Research on motivation in education: Vol. 1 Student motivation*. Academic Press, San Diego, CA.

Andrews, R. L. and Currim, I. S. (1974). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, **40**, 235–243.

Arminger, G., Stein, P., and Wittenberg, J. (1999). Mixtures of conditional mean and covariance structure models. *Psychometrika*, **64**, 475494.

Baker, L., Serpell, R., and Sonnenschein, S. (1995). Opportunities for literacy learning in the homes of urban preschoalers. *Family literacy: Connections m schools and communities, In L. M. Morrow (Ed)*, pages 236–252.

Baker, L., Scher, D., and K., M. (1997). Home and family influences on motivations for reading. *Educational Psychologist*, **32(2)**, 69 – 82.

Bartholomew, D. and Knott, M. (1999). *Latent variable models and factor analysis*. Arnold, London.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41**(1), 164–171.

Bemdt, E., Hall, B., Hall, R., and Hausman, J. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, **3**(4), 653–665.

Bijmolt, T., Paas, L., and Vermunt, J. (2004). Country and consumer segmentation: Multilevel latent class analysis of financial product ownership. *International Journal of Research in Marketing*, **21**, 323–340.

Bozdogan, H. (1987). Model selection and akaikes information criterion(aic): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. *Studies in Classification, Data Analysis,and Knowledge Organization*, pages 40–54.

Brophy, J. (1986). Teaching and learning mathematics: Where research should be going. *Journal for Research in Mathematics Education*, **17**, 323346.

Carmines, E. and Zeller, R. (1979). *Reliability and validity assessment*. Sage Publications, Inc.

Chapman, M., Skinner, E. A., and Baltes, P. B. (1990). Interpreting correlations between children's perceived control and cognitive performance: Control, agency, or means-ends beliefs? *Developmental Psychology*, **23**, 246–253.

Corno, L. and Rohrkemper, M. M. (1985). The intrinsic motivation to learn in classrooms. *Research on motivation in education*, **2**, 53–90.

Cothern, N. B. and Collins, M. D. (1992). An exploration: Attitude acquisition and reading instruction. *Reading Research and Instruction*, **31**, 84–97.

Csikszentrnihalyi, M. (1991). Literacy and intrinsic motivation. *Literacy: A overview by fourteen experts, S. R. Graubard (Ed.), New York: Hill and Wang*, pages 115–140.

Day, N. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, **56**(3), 463.

Deci, E. and Ryan, R. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39, 1**, 1–38.

Dias, J. (2004). Finite mixture models. review, applications, and computer intensive methods. *Research School Systems, Organisation and Management.*

Dweck, C. S. and Elliot, E. S. (1983). Achievement motivation. *Handbook of child psychology*, pages 643–691.

Fox, J. and Glas, C. (2001). Bayesian estimation of a multilevel irt model using gibbs sampling. *Psychometrika*, **66**, 269–286.

Fruhwirth-Schatter, S. (2006). *Finite Mixture and Markov Switching Models.* Springer, New York, NY.

Gambrell, L., Palmer, B., Codling, R., and S.A., M. (1995). Assessing motivation to read. *Instructional resources*, **14**.

Goldstein (2003). *Multilevel statistical models.* Arnold,Third Edition, London.

Goldstein, H. and Browne, W. (2002). Multilevel factor analysis modelling using markov chain monte carlo (mcmc) estimation. *Latent Variable and Latent Structure Models*, page 225243.

Goldstein, H. and McDonald, R. (1988). A general model for the analysis of multilevel data. *Psychometrika*, **53**, 455–467.

Greaney, V. and Hegarty, M. (1987). Correlates of leisure-time reading. *Journal of Research in Reading*, **10**, 3–27.

Guthrie, J. and Wigfield, A. (2000). Engagement and motivation in reading. *Handbook of reading research*, **3**.

Guthrie, J., Schafer, W., Von Secker, C., and Alban, T. (2000). Contributions of instructional practices to reading achievement in a statewide improvement program. *Journal of Educational Research*, **93(4)**, 211–225.

Guthrie, J. T. and Greaney, V. (1991). *Litexacy acts.* Handbook of reading research, Vol. 2, pp. 68-96, In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), New York: Longman.

Hagenaars, J. A. and McCutcheon, A. (2002). *Applied latent class analysis models.* Cambridge University Press.

Hansen, H. S. (1969). The impact of the home literary environment on reading attitude. *Elementary English*, **46**, 17–24.

Harter, S. (1981). A model of mastery motivation in children: Individual differences and developmental change. *The Minnesota Symposium on Child Psychology*, **14**, 215–255.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Sage Publications, Thousand Oakes, CA.

Hox (2002). *Multilevel analysis: Techniques and application*. Lawrence Erlbaum, Mahwah,NJ.

Juang, B. H. and R., R. L. (1991). Hidden markov models for speech recognition. *Technometrics*, **33**, 251272.

Keller, J. (1983). *Motivational design of instruction*. In C.M. Reigeluth (Ed.), Instructional design theories and models (pp. 383-433), New York: Lawrence Erlbaum Associates.

Lazarsfeld, P. and Henry, N. (1968). *Latent structure analysis*. Houghton Mifflin, Boston.

Longford, N. and Muthén, B. (1992). Factor analysis for clustered observations. *Psychometrika*, **57**, 581–597.

Lubke, G. and Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, **10**, 21–39.

Lukočienė, O. and Vermunt, J. (2010). Determining the number of components in mixture models for hierarchical data. *Advances in Data Analysis, Data Handling and Business Intelligence*, pages 241–249.

Lukočienė, O. and Vermunt, J. K. (2009). Determining the number of components in mixture models for hierarchical data. *Advances in data analysis, data handling and business intelligence*, pages 241–249.

Lukočienė, O., Varriale, R., and Vermunt, J. (in press). The simultaneous decision about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology.*

Magidson, J. and Vermunt, J. K. (2004). Latent class models. *Handbook of Quantitative Methodology for the Social Sciences*, pages 175–198.

McLachlan, G. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, **36**, 318–324.

McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions.* Wiley New York.

McLahan, G. and Peel, D. (2000). *Finite Mixture Models.* J. Wiley and Sons. Inc., New York.

Mclver, D., Stipek, D. J., and Daniels, D. H. (1991). Explaining within-semester changes in student effort in junior high school and senior high school courses. *Journal of Educational Psychology*, **83**, 201–211.

Morrow, L. and Young, J. (1997). Parent, teacher, and child participation in a collaborative family literacy program: The effects of attitude, motivation, and literacy achievement. *Journal of Educational Psychology*, **89**, 736–742.

Morrow, L. M. (1983). Home and school correlates of early interest in literature. *Journal of Educational Research*, **76**, 221–230.

Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, **28**, 338–354.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, **22**, 376–398.

Muthén, B. (2002). Beyond sem: general latent variable modelling. *Behaviormetrika*, **29, n. 1**, 81–117.

Muthén, B. and Muthén, L. (2000). Integrating personcentered and variable-centered analysis: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, **24**, 882891.

Muthén, B. O. (2001). Latent variable mixture modeling. *New Developments and Techniques in Structural Equation Modeling*, pages 1–33.

Muthén, B. O. (2004). *Latent variable analysis: Growth mixture modeling and related techniques.* Handbook of quantitative methodology for the social sciences, In D. Kaplan (Ed.), (pp. 345 368), Sage, Newbury Park, CA.

Muthén, B. O. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, **55**, 463469.

Muthén, L. K. and Muthén, B. O. (1998–2007). *Mplus User's Guide. Fourth Edition.* Muthén and Muthén, Los Angeles, CA.

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, **91**, 328–346.

Nylund, K., Asparouhov, T., and Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, **14**, 535–569.

Oldfather, P. and Dahl, K. (1994). Toward a social constructivist reconceptualization of intrinsic motivation for literacy learning. *Journal of Reading Behavior*, **26**, 139–158.

Palardy, G. and Vermunt, J. K. (2007). Multilevel growth mixture models for classifying group-level observations. *submitted*.

Pintrich, P. R. and De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom performance. *Journal of Educational Psychology*, **82**, 33–50.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004a). Generalized multilevel structural equation modeling. *Psychometrika*, **69**, 167–190.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004b). Generalized multilevel structural equation modeling. *Psychometrika*, **69**, 167–190.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004c). *GLLAMM Manual*. Berkeley Division of Biostatistics Working Paper Series, University of California, Berkeley.

Roeser, R., Midgley, C., and Urdan, T. (1996). Perceptions of the school psychological environment and early adolescents psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology*, **88(3)**, 408–422.

Rothenberg, T. J. (1971). Identification of parametric models. *Econometrica*, **39**, 577–591.

Saracho, . N. and Dayton, C. M. (1989). A factor analytic study of reading attitudes in young children. *Contemporary Educational Psychology*, **14**, 12–21.

Scher, D. and Baker, L. (1996). Attitudes toward readingand children's home literacy environments. *Meeting of the American Educational Research Association*.

Schiefele, U. (1991). Interest learning, and motivation. *Educational Psychologist*, **26**, 299–324.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, **26**, 207–232.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Skinner, E. and Belmont, M. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, **85**, 571–581.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Chapman and Hall, London.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. Sage, London.

Sorbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, **27**, 229239.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, **15**, 201–293.

Sweet, A., Guthrie, J., and Ng, M. (1998). Teachers perceptions and students reading motivations. *Journal of Educational Psychology*, **90**, 210–223.

Thurstone, L. (1947). *Multiple factor analysis*. University of Chicago Press, Chicago.

Varriale, R. and Vermunt, J. K. (2009). Multilevel factor mixture models.

Vermunt, J. (2003). Multilevel latent class models. *Sociological Methodology*, **33**, 213–239.

Vermunt, J. (2004). An em algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, **58**, 220233.

Vermunt, J. (2007a). A hierarchical mixture model for clustering three-way data sets. *Computational Statistics and Data Analysis*, **51**, 5368–5376.

Vermunt, J. (2008). Latent class and finite mixture models for multilevel dataset. *Statistical Methods in Medical research*, **17**, 33–51.

Vermunt, J. (2011). *Mixture models for multilevel data sets*. Handbook of Advanced Multilevel Analysis, Lawrence Erlbaum, J. Hox and J.K. Roberts (eds.), New York: Routledge.

Vermunt, J., Tran, B., and Magidson, J. (2008). *Latent class models in longitudinal research*. Handbook of longitudinal research: Design, measurement, and analysis, S. Menard (ed.), pp. 373-385., Burlington, MA.

Vermunt, J. K. (2007b). Multilevel mixture item response theory models: an application in education testing. *The Bulletin of the International Statistical Institute*, **56th Session**, 1–4.

Vermunt, J. K. and Magidson, J. (2005a). *Factor Analysis with categorical indicators: A comparison between traditional and latent class approaches.* New Developments in Categorical Data Analysis for the Social and Behavioral Sciences, 41-62, Mahwah.

Vermunt, J. K. and Magidson, J. (2005b). *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Statistical Innovations Inc., Belmont, Massachusetts.

Vermunt, J. K. and Magidson, J. (2007). *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*. Statistical Innovations Inc., Belmont, Massachussetts.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. Springer-Verlag, New York.

Weisz, J. and Cameron, A. (1985). Individual differences in the student's sense of control. *Research on motivation in education*, **2**, 93140.

Wentzel, K. R. (1997). Student motivation in middle school : The role of perceived pedagogical caring. *Journal of educational psychology*, **89(3)**, 411–419.

Wigfield, A. and Karpathian, M. (1991). Who am i and what can i do? children's self-concepts and motivation in academic situations. *Educational Psychologist*, **26**, 233–262.

Wolfe, J. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, **5**(3), 329–350.

Yung (1997). Finite mixture in confirmatory factor analysis models. *Psychometrica*, **62**, 297–330.