Università degli Studi di Firenze
Dipartimento di Statistica "G. Parenti"
Dottorato di Ricerca in Statistica Applicata
XX *ciclo* SECS-S/01

Multilevel Mixture Models for the analysis of the University Effectiveness

Roberta Varriale

Tutor: **Prof. Bruno Chiandotto**

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

Co-tutor: **Dott. Leonardo Grilli**

Coordinatore: **Prof. Guido Ferrari**

A chi mi ha regalato un sorriso
e, soprattutto,
alla mia famiglia

Ringrazio innanzitutto il prof. Bruno Chiandotto per avermi incoraggiata nell'intraprendere il dottorato e per avermi guidata nella mia ricerca, anche quando non ho seguito proprio tutti i suoi suggerimenti.

Ringrazio Leonardo per i suoi utili consigli e il prof. Vermunt per avermi dato l'opportunità di svolgere un periodo di studi all'estero. Ringrazio il prof. Ferrari per l'organizzazione, non sempre semplice, di questo dottorato e il Dipartimento di Statistica, per la gentilezza delle persone che ne fanno parte.

Ringrazio l'Olanda, per avermi dato tanto, senza chiedere niente in cambio.

Ringrazio Bruno, per essere stata una presenza comunque costante e per aver creduto in me, sempre.

Un ringraziamento speciale va ai miei colleghi: Bruno, Caterina, Eleonora, Federica, Giulia, Graziano, Vincenzo. Per i momenti che abbiamo condiviso, per quello che mi hanno insegnato, per le risate e per la nostra mensa. Se di questo periodo ho un buon ricordo lo devo soprattutto a loro.

Ringrazio Caterina, per essere stata un'efficientissima CapoCate e per aver condiviso con me questa esperienza e tutto quello che ne ha comportato, dagli esordi del lontano 2003...

Grazie, inoltre, ai vecchi amici e ai nuovi incontri che in questi tre anni mi sono stati vicini, attraverso piccoli gesti o, semplicemente, incoraggiandomi.

Infine, ringrazio Gianmarco, per essere entrato nella mia vita e per aver rappresentato una presenza fondamentale durante questo ultimo anno di dottorato. E per tutto quello che sarà.

E ringrazio i miei genitori e mio fratello per l'affetto di cui mi hanno circondata. Per avere creduto in me e nelle mie scelte cercando, sempre, di capirmi fino in fondo.

Roberta

Contents

List of Figures

List of Tables

Introduction

During recent years there has been a substantial increase in the size of the statistical literature on measuring the performance of public sector institutions (Bratti et al., 2004). This is naturally linked to the need to efficiently allocate scarce public resources to public institutions, for example in the context of education and health, increasing the emphasis of public policy on institutional auditing and surveillance. Every evaluation process has its own specific connotations depending on the context, but the aim is always the same: to trigger a system of actions and counteractions aimed at improving the general quality of the performed activities.

This thesis focuses on the evaluation of higher education (university) quality with the use of multilevel mixture factor models included in the generalized latent variable modeling framework. This framework has been developed in recent years and tries to unify and extend latent variable models, integrating specific methodologies with different traditions and application fields in a global theoretical context. The literature on the topic is not very developed and usually focuses on theoretical aspects of models with a single type of latent variables (all continuous or all categorical). This thesis describes a very general framework proper for models with both continuous *and* categorical latent variables, from the theoretical and applied point of view. The ultimate aim is to provide policy advice for universities and to highlight the flexibility of the modeling framework. In particular, the models used in the applications combine the features of factor models and latent class models in the multilevel framework.

The evaluation of the global performance of a university system, and generally of a public activity, can be divided into two phases: the first deals with how resources are spent to reach particular objectives, *efficiency analysis*; the second deals with the adherence of results to the planned objectives, *effectiveness analysis* (Lockheed and Hanushek, 1994). Moreover, the two phases can be analyzed from both internal and external points of view. Internal and external refer to the individuals or institutions that are effected by the processes that are evaluated; these may have different interests and expectations.

In this context, Chiandotto (2004) suggests to modify the scheme proposed by Lockheed and Hanushek (1994) including the users' subjective point of view in the evaluation of the performance of a university system (Table 1). He suggests evaluating the students' perception of the quality of the services provided by the institution, both at the time of completion of the degree (*internal effectiveness*) and at some later date (*external effectiveness*).

Table 1: Concepts of educational efficiency and effectiveness: the schema of Lockheed and Hanushek (Lockheed and Hanushek, 1994) modified by Chiandotto (2004).

	Internal to the system	**External to the system**
	Internal effectiveness	**External effectiveness**
Physical aspects	*effect of university or study programs on the student's learning ability*	*effect of university or study programs on the graduate's skills*
Satisfaction	student's satisfaction with the attended study program	graduate's satisfaction with the occupational condition
	Internal efficiency	**External efficiency**
Monetary aspects	*costs/returns analysis of the investments*	*economic return due to study programs attended*
Satisfaction	student's satisfaction with the employed resources	graduate's satisfaction with the economic condition

Martensen et al. (2000) and Harvey and Knight (1996) also underline that it is essential to measure students' perceived quality and satisfaction within higher education institutions to develop continuous improvement of teaching, staff, equipment and programs (the terms study programs, programs, degree programs are used intercheanglebly in this literature).

Other works assess students as the principal actors of the university system and use different methodologies to evaluate both the internal and external effectiveness of the university.

For instance, Chiandotto and Bacci (2006*b*) propose an indicator to evaluate the internal efficacy of university education using students' satisfaction, time to complete the degree and "statement of re-enrollment" in order to build a synthetic indicator.

Some "objective" indicators of external effectiveness are the first employment rate (Chiandotto and Bacci, 2004) and the time between graduation and employment. "Subjective" indicators include the evaluation of the usefulness of the qualification for the work, the degree to which graduates use the skills acquired at university at work (Grilli and Rampichini, 2007*b*), and so on. Chiandotto and Bacci (2006*a*) focus on the use of skills that graduates achieved at university, analysing the ability of study programs to create the competencies required by the labour market. They estimate logistic regression models, also using multilevel techniques, on students who graduated at the University of Florence in 2000.

Chiandotto et al. (2006) propose an index for measuring at the same time both internal and external effectiveness using individual characteristics, objective measurements of performance, subjective measurements of satisfaction with the university experience and variables relating to the occupation after the degree. The modeling approach, such as the modeling approach used by Martensen et al. (2000) and Eskildsen et al. (2000), is inspired by the European Customer Satisfaction Index (ECSI); the basic model is a structural equation model with latent variables linking students' satisfaction to its determinants and consequences.

In this thesis we evaluate the effectiveness of the university from the users' point of view with two separate analyses: we study the perceived quality of students on the global university experience at the completion of the degree to evaluate the internal effectiveness of the university and we analyse the job satisfaction of students who graduated one year before as an indicator of the university external effectiveness. Indeed, the more satisfied a student is with the university and his job, the more effectiveness the university is. In particular, we analyse "similar" items on the students' perceived quality expressed on the same scale in order to stress the interpretation of the models and we exclude other aspects of the quality of the university, such as grade at the degree, degree completion time and also employment status, time to get the first job, use of skills acquired at university, etc.

Actually, when we speak about university effectiveness we refer to the effectiveness of each study program, relative to each other, since each program has its own features and organization. Students attending the same study program share common environments, experiences, and interactions that can influence their perceived quality (internal or external) of the university. Instead of relying on indicators for each group calculated trough the mean of individual responses (Chiandotto and Bacci, 2006*b*), we use multilevel techniques that recognise the existence of data hierarchies by allowing for residual components at each level in the data structure (Snijders and Bosker, 1999), and allow correct inferences treating the units of analysis as dependent observations. Furthermore, through the multilevel models, we are able to analyse the phenomenon at the same time both at the individual level and the program level, "pushing" up the information obtained with the questionnaires filled in by students. Due to the aim of the thesis we will use only random effect methods.

In the thesis we use different specifications of the *multilevel mixture factor model*. One aim is to explain the correlation among observed random variables in terms of fewer unobserved random variables, called common *factors*; in our research, these represent the overall satisfaction and/or satisfaction with some specific features of the university or of the job. In particular, when we use all continuous latent variables we call the models multilevel factor models, while when we use a combination of continuous and categorical latent variables all different models specifications are called multilevel mixture factor models. Which model should be selected depends on the aims of the specific research and on the substantive reason to believe in the nature, continuous or categorical, of latent variables.

In the analysis of the university internal effectiveness, we first use a *multilevel factor model*, following the strategy used by Grilli and Rampichini (2007*a*). With continuous latent variables at both levels of the analysis we study the latent constructs underlying the phenomenon of satisfaction at the student and program level, highlighting the differences between the two structures. We also investigate the differences between programs in the students' satisfaction ranking the programs along a continuum.

Next, we apply a *multilevel mixture factor model* with continuous latent variables at the student level and a categorical variable at the program level. The aim is to classify the second level units (programs) into a small number of classes, which differ with respect to the item intercept of the specified factor analysis model. Our model is indeed a mixture factor analysis model in which we classify groups rather than individuals. In particular, the latent class approach is well known in the one-level framework (Hagenaars and McCutcheon, 2002); in the multilevel framework it was first proposed by Vermunt (2003).

At the end we merge the results of the two analyses relative to the program level. An example on the use of both continuous and categorical latent variables in one-level context is given by Muthén (2001); this thesis reproduces that analysis in the two-level context showing how different statistical techniques can be used together to better explain a phenomenon.

In the analysis of job satisfaction we use a *multilevel mixture factor model*. With continuous latent variables at the individual level we reduce the dimensionality of the phenomenon and with a categorical variable at program level we classify programs relatively to the obtained latent dimensions. The use of the categorical variable is different from the previous analysis: the aim is to determine whether the programs (or groups of programs) differ in the mean value of the latent variables at the individual level representing the job satisfaction, and not if they differ in the mean value of each indicator. Indeed, classifying programs in groups differing in 14 characteristics can be "confounding": the factor model at the individual level is used to reduce the dimensionality of the phenomenon and the classification of programs is based on the obtained latent dimensions.

Data come from two surveys of the consortium AlmaLaurea, which currently includes 51 Italian universities[1]. Because of our knowledge of the context, we focus only on data about the university of Florence.

Data used for the analysis of the internal effectiveness of the university come from the AlmaLaurea survey on profile of students who graduated in 2004. We focus on 1800 students who graduated (Bachelor degree) under the new Italian university system operating since 2001 since we want to provide policy advice to the university; 1473 students from 38 study programs replied to the survey. Unfortunately in the questionnaire there are no questions on the contents of the study programs; in the evaluation of the university quality this represents a big weakness of the data.

Data used for the analysis of the external effectiveness of the university come from the AlmaLaurea survey "Employment opportunities, 2005". In this survey AlmaLaurea collected information on students who had graduated 1, 3 and 5 years previously. We focus on students that graduated in 2004 working at the moment of the interview; of course this study represents only a first step in evaluating the evolution of job satisfaction over the time. In particular, AlmaLaurea only interviews students that graduated during the summer session; this may cause misleading results in comparing different study programs. For some study programs there may be differences in terms of characteristics of students who graduated during different periods of the academic year. For example, for administrative reasons, a student graduating before a specific date (varying slightly among Faculties) does not have to pay taxes for the next academic year, so in the spring term the quality of students graduating may be less than in other periods.

In the analysis of the external effectiveness, we focus on data relative to students who graduated with the old university system. First students enrolled with the new system could graduate in september 2004, so analysing students who graduated with the new university system in the summer session means analyse hybrid students that may have particular features. Usually, students that "change" system are involved in the educational process for many years and prefer to finish their studies quickly (for almost all study programs the new university system degree takes 3 years and the old univer-

[1]At 15/12/2007, AlmaLaurea has information on over 950.000 students from 51 Italian universities, out of 85.

sity system degree takes 4 or 5 years) even if the new degree is less prestigious than the old degree and offers fewer employment opportunities.

It should be noted that, although possible within the modeling framework, we do not use covariates in our models and we measure the students' satisfaction as it is experienced in the real world. Indeed our main aim is to provide police advice for university and it is difficult for university to act in different ways depending on students' characteristics (covariates at first level) or study programs' characteristics (covariates at second level). Obviously, the use of covariates let to evaluate the "net" effectiveness of the study programs, controlling for their composition and their features; from an applied point of view, it lead to a better knowledge of the phenomenon of satisfaction and as a result will lead to focus the university economical and political resources on particular aims. In this respect, our analysis can also be seen as an initial attempt to include more information in the model.

We use the syntax module (*Beta version* at 1st of April 2007) of Latent GOLD software, version 4.5 (Vermunt and Magidson, 2007) that allows the definition of models containing any combination of categorical and continuous latent variables at each level of the hierarchy. In this thesis, we will illustrate the characteristics of the software.

In Chapter 1 we show the theoretical framework of our analysis: the generalized latent variable modeling framework. This framework captures a wide variety of statistical concepts, including random effects, common factors, missing data, finite mixtures, latent classes, and clusters (Skrondal and Rabe-Hesketh (2004), Vermunt (2007), Muthén and Muthén (1998-2007)).

After describing the literature on the topic, the multilevel mixture factor model used in the thesis is introduced as a specification of the generalized latent variable model. Since the aim of this research is to study data with one level of aggregation (students nested in programs), only two-level models are illustrated.

Section 1.2 illustrates how to model different type of outcomes and the linear predictor of a multilevel mixture factor model. Section 1.3 shows the difference and similarities between the use of continuous or categorical latent variables and the nine typologies of models that can be obtained combining different specification of latent variables both at the within and between level of the analysis. Sections 1.3.1 and 1.3.2 show models with continuous latent variables at the individual level and, respectively, continuous and categorical latent variables at the group level. Sections 1.4.1 and 1.4.2 present technical aspects relating to the fitting and evaluation of a model. Section 1.5 is dedicated to "posterior analysis"; when using continuous latent factors, the aim is to locate units on the dimensions of the latent space (finding the *factor scores*) and when using categorical latent variables the aim is to classify units in different *classes* representing some typical profile (Bartholomew and Knott, 1999).

In Chapter 2 and Chapter 3 the empirical part of the work is described. Both Chapters are divided into two sections: the first deals with the analysis of the internal effectiveness of the university, the second deals with the analysis of its external effectiveness.

In Chapter 2, after a brief description of available data (sections 2.2 and 2.3), results of traditional analyses are shown. In both case studies, we first investigate the correlation between the items relating to specific aspects of satisfaction (sections 2.2 and 2.3). This analysis describes the phenomenon, but it does not allow an evaluation of the individual effect of each single aspect on global satisfaction and it does not allow an evaluation of the performance of the study programs. For the evaluation of the internal effectiveness of the university system, we next apply a multilevel regression model to global satisfaction (section 2.2.1) using as covariates the students responses to the items on specific aspects of satisfaction. The first level units are the students, the second level units are the programs that students attended. By using the responses on the various satisfaction items either as dependent and independent variables in a regression models, one considers these as perfect measures of a particular characteristic of an individual. It is however more suitable to treat item responses as imperfect measures of one or more latent constructs that cannot be observed directly.

Considering the nature of the phenomenon and the data of AlmaLaurea survey, the most suitable statistical methodology for the analysis of the global satisfaction is the multilevel mixture factor model; the results are shown in Chapter 3.

Section 3.1.1 shows the results of the application of the *multilevel factor model* for the analysis of the university internal effectiveness and section 3.1.2 shows the results of the application of the *multilevel mixture factor model*. In section 3.1.3 the results of both analyses are merged highlighting the information obtained with this procedure.

Section 3.2.1 describes the results of the *multilevel mixture factor model* for the analysis of the university external effectiveness from the graduates' point of view.

At the end of the dissertation, some concluding remarks summarize the main results of the work, relative to both the analysis of the university system effectiveness, and latent variable modeling. Furthermore, some limitations of the analyses are highlighted, together with some proposal for future research.

Chapter 1

Latent variables modeling, multilevel framework

In almost all fields of human science it is possible to recognise the presence of some *latent variables*. These are variables that are not directly observed but are rather inferred from other variables that are observed and directly measured. So, latent variables are concepts, hypothetical constructs that, within a statistical process, influence the observed realisation of a phenomenon. They are used in many disciplines, for examples in social sciences, economics, psychometrics, but it is sometimes difficult to recognize the "concept" of underlying latent variables in all situations.

The aim of this dissertation is to deeply analyse the use of latent variables in the well know framework of factor model, traditionally used to analyse the relationship between one or more latent construct(s) and the observed indicators.

The standard formulation of factor models concerns a set of latent variables measured on a set of independent units. The latter assumption may be inadequate in multilevel settings where units are nested in clusters, leading to within-cluster dependence and, as a result, the necessity to use multilevel techniques (Rabe-Hesketh et al., 2004*a*). In this context, this thesis deepens the analysis of factor models in the multilevel framework. Furthermore, the latent variables used in multilevel factor models are assumed to be both continuous and categorical; when a combination of continuous and categorical latent variables is used all different models specifications are called multilevel mixture factor models.

This Chapter illustrates some methodological aspects of the multilevel mixture factor model; since the aim of the research is to study an empirical topic, the focus is most on interpretational aspects.

Section 1.1 presents the generalized latent variable modeling framework, integrating in a global theoretical context several specific methodologies, such as multilevel and longitudinal models, generalized linear mixed models, random coefficient models, factor models, etc. After describing the literature on the topic, the multilevel mixture factor model is introduced as a specification of the generalized latent variable model. Since the aim of the research is to study data with one level of clustering, only two-level models are illustrated.

Conditional on the latent variables, the response model of a generalized latent variable model is a generalized linear model specified via a linear predictor, a link, and a distribution from the exponential family. Section 1.2 illustrates both how to model different types of outcomes and the linear predictor of a multilevel mixture factor model.

The latent variables can be assumed to be both continuous and categorical, depending on their substantive meaning, but also on the aim of the research. Section 1.3 shows the differences and the similarities between the two specifications and shows nine model typologies that can be obtained combining continuous and categorical latent variables at the within and between levels of the analysis. Sections 1.3.1 and 1.3.2 show models with continuous latent variables at the individual level and, respectively, continuous and categorical latent variables at the group level.

Sections 1.4.1 and 1.4.2 present technical aspects relating to the fitting of a model. In section 1.4.1 the maximum likelihood approach is described, together with likelihood maximization methods and methods to solve the multiple integrals involved in the likelihood expression. In section 1.4.2 some statistical aspects related to the evaluation of a factor model are illustrated.

Section 1.5 is dedicated to the "posterior analysis". Indeed, as underlined by Bartholomew and Knott (1999), the main aim of the researcher using factor models is in what can be known about the latent variables after the manifest variables have been observed.

In this work we focus on the interpretational point of view of the models in order to highlight the flexibility of the generalized latent variable modeling framework. Since the models will be estimated using the syntax module of Latent GOLD 4.5 (Vermunt and Magidson, 2005*b*), this Chapter focuses on the Latent GOLD framework, highlighting some technical aspects used by the software.

In order to illustrate and clarify the meaning of the models, common path diagrams are used.

1.1 Generalized latent variable model

A very general definition of latent variables is given by Skrondal and Rabe-Hesketh (2004), Ch. 1: a latent variables is a "*random variable whose realizations are hidden from us*".

Traditionally, latent variable models were used in psychometrics and have been concerned with measurement error and latent variable constructs measured with multiple indicators (factor analysis). Nowadays, latent variables are used to represent different phenomena, such as "true" variables measured with error, hypothetical constructs, unobserved heterogeneity, missing data, etc.

In the current literature, many authors propose a generalized latent variable modeling framework, integrating specific methodologies in a global theoretical context. One example is the Generalized Linear Latent and Mixed Models framework of Skrondal and Rabe-Hesketh (2004) that unify and extend latent variable modeling as multilevel, longitudinal and structural equation models as well as generalized linear mixed models, random coefficient models, item response models, factor models, etc. This framework is implemented in the GLLAMM software package (Rabe-Hesketh et al., 2004*b*), a Stata program to fit generalized linear latent and mixed models and includes models where the latent variables are all continuous or all discrete. Another example is the work of Vermunt (2007) that allows defining models with any combination of categorical and continuous latent variables at each level of the hierarchy; the framework is implemented in the syntax version of Latent GOLD software (Vermunt and Magidson, 2007). Also Muthén deals with the same topics, he developed the software M*plus* (Muthén and Muthén, 1998-2007) that also allows defining models with categorical and continuous latent variables at each level of the analysis.

An interesting esemplification of the global latent variable framework is given in Figure 1.1 (Muthén and Muthén, 1998-2007).

The rectangles represent observed variables that can be background variables (x) or continuous and censored outcome variables (y) or binary, ordinal, nominal, and count outcomes (u). The circles represent latent variables, both continuous (f) and categorical (c). The arrows represent regression relationships between variables; regression relationships that are allowed but not specifically shown in the figure include regressions among observed outcome variables, among continuous latent variables, and among categorical latent variables[1]. Of course, different regression models are used depending on the

[1] The framework implemented in Latent GOLD is less general than the framework suggested by the figure: the regression relationships for the latent variables is only partially implemented.

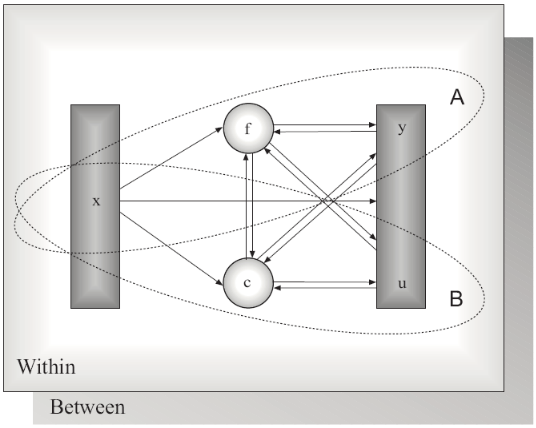

Figure 1.1: Latent Variable modeling, (Muthén and Muthén, 1998-2007, Introduction).

nature of the involved variables.

Ellipse A describes models with only continuous latent variables. Ellipse B describes models with only categorical latent variables. The full modeling framework describes models with a combination of continuous and categorical latent variables. The Within and Between parts of the figure indicate multilevel models that describe individual-level (within) and cluster-level (between) variation in the data.

The generalized latent variable model is formally described by two elements: the response model for the observed variables conditional on the latent variables and the model for the latent variables.

Using the index j to denote an independent observation corresponding to the highest level of the hierarchy (L), the model can be formulated with two equations (Vermunt, 2007) relative, respectively, to the measurement and the structural part of the complete model:

$$g[E(\boldsymbol{y}_j|\boldsymbol{\eta}_j)] = \boldsymbol{Z}_j\boldsymbol{\beta} + \boldsymbol{W}_j^{(1)}\boldsymbol{\Lambda}^{(1)}\boldsymbol{\eta}_j \tag{1.1}$$

$$h[E(\boldsymbol{\eta}_j^{(l)})] = \boldsymbol{X}_j\boldsymbol{\gamma} + \boldsymbol{W}_j^{(l)}\boldsymbol{\Lambda}^{(l)}\boldsymbol{\eta}_j^{(l+)} \qquad l = 2, \ldots, L \tag{1.2}$$

The vector $\boldsymbol{y}_j$ denotes the response vector with elements representing the responses of unit j. In a model with 3 levels of analysis each element of $\boldsymbol{y}_j$ is $y_{hij}, h = 1, \ldots, H, i = 1, \ldots, I, j = 1, \ldots, J$. For instance, given one level of clustering (students nested in study programs), y_{hij} represents the item h response of each student i in group j. The model allows for item nonresponse.

The vector $\boldsymbol{\eta}_j$ denotes the M_l random latent variables varying at all levels of the analysis, so $(\boldsymbol{\eta}_j^{(2)'}, \ldots, \boldsymbol{\eta}_j^{(L)'})'$; the latent variables may be either continuous or categorical[2]. The two vectors $\boldsymbol{\eta}_j^{(l)}$ and $\boldsymbol{\eta}_j^{(l+)}$ refer, respectively, to the latent variables at level l and l and higher $(l+)$; so, $\boldsymbol{\eta}_j^{(l)} = (\eta_{j1}^{(l)}, \ldots, \eta_{jM_l}^{(l)})'$ and $\boldsymbol{\eta}_j^{(l+)} = (\boldsymbol{\eta}_j^{(l)'}, \ldots, \boldsymbol{\eta}_j^{(L)'})'$. As underlined by Skrondal and Rabe-Hesketh (2004), to regress a higher level latent variable on a lower level latent variable would mean to force the higher level variable to vary at a lower level; in equation (1.2) the latent variables $\boldsymbol{\eta}_j^{(l)}$ are regressed on latent variables at the same level l and l or at a higher level $(l+)$. In the Latent GOLD framework the regression relationships for the latent variables is only partially implemented; in particular, on the right hand side of equation (1.2) only latent variables at higher level than l can appear.

The two matrices $\boldsymbol{Z}_j$ and $\boldsymbol{X}_j$ with the corresponding coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ denote the fixed part of the model affecting, respectively, the observed items and the latent structure at level l. Different links and distributions can be specified for different responses.

The matrices $\boldsymbol{W}_j$ and $\boldsymbol{\Lambda}$ denote, respectively, the design matrix and factor loading matrix[3] of a generalized latent variable model with dimension depending on the structure of the model. The two matrices are different in the two equations: the superscripts (1) and (l) in the general equations (1.1) and (1.2) indicate the level of analysis their are "affecting". In particular, $\boldsymbol{W}_j^{(1)}$ and $\boldsymbol{\Lambda}^{(1)}$ indicate the design matrix and factor loading matrix affecting directly the outcomes; $\boldsymbol{W}_j^{(l)}$ and $\boldsymbol{\Lambda}^{(l)}$ indicate matrices affecting level l latent variables. An example of the matrix $\boldsymbol{\Lambda}^{(1)}$ is given at the end of the section.

In particular, the product term $\boldsymbol{W}_j^{(1)}\boldsymbol{\Lambda}^{(1)}$ in equation (1.1) yields the generalization of both the factor analytic and the random coefficient model (Vermunt, 2007). By setting $\boldsymbol{W}_j^{(1)} = \boldsymbol{1} \otimes \boldsymbol{I}$, a factor analytic model is obtained and by setting $\boldsymbol{\Lambda}^{(1)} = \boldsymbol{I}$ a random coefficient model is obtained. So, depending on the definition of the two matrices, the latent variables $\boldsymbol{\eta}_j$ can be common factors in factor analysis or random coefficients in multilevel models.

When the latent variables are categorical, the vector $\boldsymbol{\eta}_j$ and the matrices $\boldsymbol{\Lambda}^{(1)}$ and $\boldsymbol{\Lambda}^{(l)}$ are, respectively, a "block" vector and matrix, where the block dimension depends on the number of classes of each latent variable (see the end of the section for an example).

The main aim of this thesis is to deeply analyse the use of the latent variables in the well know framework of factor modeling used to analyse

[2]The multilevel models with a combination of continuous and categorical latent variables are called multilevel mixture models.

[3]The elements of the matrix $\boldsymbol{\Lambda}$ (factor loadings) do no vary depending on j, the subscript j is not indicated.

the relationship between the latent constructs and the observed indicators. So, only factor models are illustrated, removing the matrices $\boldsymbol{W}_j$ from the equations (1.1) and (1.2).

With factor models, the correlation among observed random variables is explained in terms of fewer unobserved random variables, called common *factors*. These can represent an hypothetical construct or fallible measurements of a variable. One aim of using latent variables in a factor model is to reduce the dimensionality of data: a large number of observable variables are aggregated in a statistical model to represent an underlying concept, making easier to understand the data. Of course, other multivariate statistical methods can be used to explore the "dimensions" underlying the data, for example principal component analysis, canonical correlations, discriminant analysis and multidimensional scaling but, contrary to factor models, these methods merely represent transformations or geometric features of the data.

Since the factor model is used to study to data with one level of aggregation, only two-level models are illustrated[4]; the models with a combination of continuous and categorical latent variables are called two-level mixture factor models.

The two-level (mixture) factor model is expressed by[5]:

$$g[E(\boldsymbol{y}_j)] = \boldsymbol{Z}_j\boldsymbol{\beta} + \boldsymbol{\Lambda}^{(1)}\boldsymbol{\eta}_j \tag{1.3}$$
$$h[E(\boldsymbol{\eta}_j^{(2)})] = \boldsymbol{X}_j\boldsymbol{\gamma} + \boldsymbol{\Lambda}^{(2)}\boldsymbol{\eta}_j^{(3)} \tag{1.4}$$

where $\boldsymbol{y}_j$ denotes the response vector with element y_{hij} representing the response to indicator h of each individual i belonging to group j. The model allows for item nonresponse; that is, for each unit i of cluster j, y_{hij} may be missing for some h.

Following the conventions, these models are called *two*-level (mixture) factor models: the individual units i are the level-1 units, and the group level units j are the level-2 units. If the items are treated as level-1 units, the models become 3-level models with units at level 2 and groups at level 3. Table 1.1 shows the terms that are used interchangeably in this thesis.

In the two-level framework, the vector $\boldsymbol{\eta}_j$ in equation (1.3) denotes the latent variables varying both at the i-th and j-th level of the analysis affecting directly the observed responses. The latent variables varying at the i-th level of the analysis are denoted by $\boldsymbol{\eta}_j^{(2)}$ and the latent variables varying at the j-th level of the analysis are denoted by $\boldsymbol{\eta}_j^{(3)}$, so $\boldsymbol{\eta}_j = (\boldsymbol{\eta}_j^{(2)'}, \boldsymbol{\eta}_j^{(3)'})'$. The vector

[4]The extension to more than two levels is conceptually straightforward.

[5]The equations refer to Latent GOLD framework. Indeed, in the second equation only latent variables at higher level than 2 are used.

Table 1.1: Two-level generalized latent variable model, terminology.

	i	j
Terms	first level units	second level units
	lowest level units	highest level units
	level-1 units	level-2 units
	unit level	clusters
	individuals	groups

$\boldsymbol{\eta}_j^{(3)}$ in equation (1.4) denotes the latent variables at the j-th level affecting the latent variables at the i-th level of the analysis.

When all latent variables are continuous, the vector $\boldsymbol{\eta}_j$ has dimension $(M_2 + M_3) \times 1$ and the two factor loadings matrices $\boldsymbol{\Lambda}^{(1)}$ and $\boldsymbol{\Lambda}^{(2)}$ have dimension, respectively, $H \times (M_2 + M_3)$ and $M_2 \times M_3$.

For instance, assuming two ($M_2 = 2$) continuous latent variables η_{j1} and η_{j2} varying at level i of the analysis, omitting the covariates, for each i in group j the equation (1.3) is expressed by:

$$g \begin{pmatrix} y_{1ij} \\ y_{2ij} \\ \vdots \\ y_{Hij} \end{pmatrix} = \begin{pmatrix} \lambda_{11}^{(2)} & \lambda_{12}^{(2)} \\ \lambda_{21}^{(2)} & \lambda_{22}^{(2)} \\ \vdots & \vdots \\ \lambda_{H1} & \lambda_{H2} \end{pmatrix} \begin{pmatrix} \eta_{1ij}^{(2)} \\ \eta_{2ij}^{(2)} \end{pmatrix}$$

where the superscripts of the factor loadings λ refer to the level at which the random latent variables η_{j1} and η_{j2} vary.

If the latent variables are categorical the vector $\boldsymbol{\eta}_j$ and the matrices $\boldsymbol{\Lambda}^{(1)}$ and $\boldsymbol{\Lambda}^{(2)}$ are, respectively, a "block" vector and matrix, where the block dimension depends on the number of classes (K_m) of each m-th latent variable[6]. For example, in the two-level factor model represented in equation (1.3), $\boldsymbol{\eta}_j$ has dimension $\sum_{l=2}^{3}\sum_{m=1}^{M_l} K_m \times 1$, and $\boldsymbol{\Lambda}^{(1)}$ has dimension $H \times \sum_{l=2}^{3}\sum_{m=1}^{M_l} K_m$.

Assuming two ($M_2 = 2$) categorical variables $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ at the i-level of the analysis with, respectively, K_1 and K_2 categories, for each i in group j, omitting the subscripts i and j and the covariates, the latent variable model

[6]For simplicity the same notation $\boldsymbol{\eta}_j$ and $\boldsymbol{\Lambda}^{(1)}$ and $\boldsymbol{\Lambda}^{(2)}$ is used for both continuous and latent variables. The difference is in the dimension of the vector and matrices in the two cases.

is expressed by:

$$g\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_H \end{pmatrix} = \left(\begin{array}{ccc|ccc} \lambda_{111}^{(2)} & \dots & \lambda_{11K_1}^{(2)} & \lambda_{121}^{(2)} & \dots & \lambda_{12K_2}^{(2)} \\ \lambda_{211}^{(2)} & \dots & \lambda_{21K_1}^{(2)} & \lambda_{221}^{(2)} & \dots & \lambda_{22K_2}^{(2)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{H11}^{(2)} & \dots & \lambda_{H1K_1}^{(2)} & \lambda_{H21}^{(2)} & \dots & \lambda_{H2K_2}^{(2)} \end{array}\right) \left(\begin{array}{c} \eta_{11}^{(2)} \\ \eta_{12}^{(2)} \\ \vdots \\ \eta_{1K_1}^{(2)} \\ \hline \eta_{21}^{(2)} \\ \eta_{22}^{(2)} \\ \vdots \\ \eta_{2K_2}^{(2)} \end{array}\right)$$

where $\eta_{mk_m}^{(2)}, m = 1, 2, k_m = 1, \dots, K_m$ is an indicator variable taking the value 1 with probability π_{k_m} if unit i belongs to latent class k_m of the variable $\eta_m^{(2)}$ and 0 otherwise; the classes are mutually exclusive.

The models that are illustrated in the thesis do not contain covariates since no covariates are used in the applications.

The models can be represented by the figure 1.2.

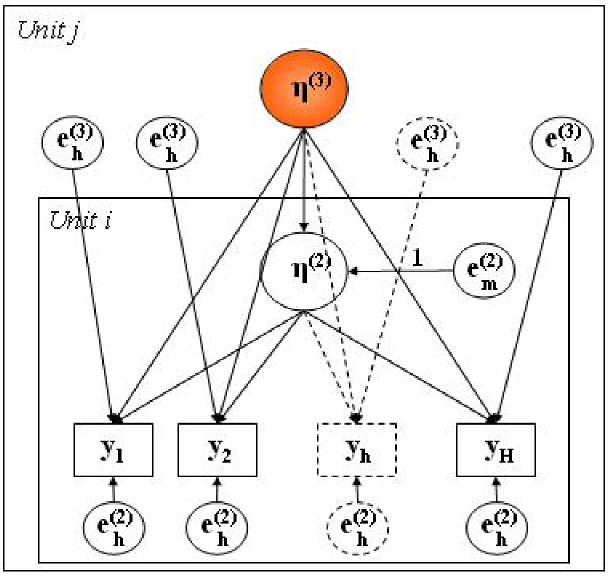

Figure 1.2: Two-level (mixture) factor model.

Following the conventions, circles represent latent variables and rectangles observed variables; latent class variables are indicated with filled circles. The nested frames represent the nested levels, for example, variables located within the outer frame labeled j vary between clusters and have a

j subscript (Rabe-Hesketh et al., 2004*a*). All the latent variables such as residuals, disturbances, or random effects are also enclosed in circles and the arrows connecting latent and/or observed variables not necessarily represent linear relations. Possible correlation among latent variables or among items are represented with dotted lines.

1.2 Response Model

Conditional on the latent variables, the response model for the observed variables is a generalized linear model specified via a linear predictor, a link, and a distribution from the exponential family.

Let y_{hij} denote the observed response on indicator h $(h = 1, \ldots, H)$ of individual i $(i = 1, \ldots, n_j)$ within group j $(j = 1, \ldots, J)$; the total number of individuals is N, where $\sum_{j=1}^{J} n_j = N$.

For a model with h items, 2 levels (i,j) of analysis and M_l latent variables at level l $(l = 2, 3)$, the linear predictor is[7]:

$$v_{hij} = \mu_h + \sum_{m=1}^{M_2} \sum_{k_m=1}^{K_m} \lambda_{mhk_m}^{(2)} \eta_{mijk_m}^{(2)} + \sum_{m=1}^{M_3} \sum_{k_m=1}^{K_m} \lambda_{mhk_m}^{(3)} \eta_{mjk_m}^{(3)} + e_{hj}^{(3)}. \tag{1.5}$$

If the m-th $(m = 1, \ldots, M_2$ and $m = 1, \ldots, M_3)$ latent variable is continuous $K_m = 1$.

The conditional expectation of the response y_{hij} given the latent variables at different levels is "linked" to the linear predictor v_{hij} via a link function:

$$g(E(y_{hij}|\boldsymbol{\eta}_j)) = v_{hij} \tag{1.6}$$

where $\boldsymbol{\eta}_j = (\boldsymbol{\eta}_j^{(2)\prime}, \ldots, \boldsymbol{\eta}_j^{(L)\prime})'$ and $\boldsymbol{\eta}_j^{(l)} = (\eta_{j1}^{(l)}, \ldots, \eta_{jM_l}^{(l)})'$.

The kinds of response variable that can be accommodated are (Vermunt and Magidson, 2005*c*):

- continuous responses,
- polythomous responses,
- ordinal responses,
- counts and durations in continuous time,

[7]Some constraints on the λ's are necessary for the identification of the model. In this Chapter the identification topic is not addressed and some details on the identification of specific models are in Chapter 2.

- rankings and pairwise comparisons.

Of course, the choice among different link functions follows naturally from the scale types of the observed variables. Furthermore, it is also possible to allow for different distributional form for each indicator.

Here, some examples of response variable that can be accommodated are briefly illustrated[8]. For simplicity, the subscripts are omitted.

With continuous responses, an identity link and a normal distribution are usually assumed, so:

$$y = v + e$$

with $f(e) \sim N(0, \sigma^2)$; the conditional density given the latent variables becomes:

$$f(y|\boldsymbol{\eta}) = \sigma^{-1}\phi(v\sigma^{-1})$$

where ϕ represents the standard normal density.

With polytomous responses the multinomial logit model is generally implemented. Let $a_s, (s = 1, \ldots, S)$ be the categories of the response y and $P(y = a_s|v)$ be the probability of each category conditional to the linear predictor v, the multinomial logit model is expressed by:

$$P(y = a_s|v) = \frac{exp(v^s)}{\sum_{t=1}^{S} exp(v^t)}$$

where v^s is the linear predictor for each category a_s that can include unit and category specific covariates:

$$v^s \quad = \quad m^s + \boldsymbol{x}^{'}\boldsymbol{\beta} + \boldsymbol{x}^{s'}\boldsymbol{\beta}.$$

With ordinal responses, more model specifications are possible, the more used being the proportional odds model and the adjacent-category logit model.

Let $s, s = 1, \ldots, S$ be the category of the ordinal response y, the proportional odds model is expressed by:

$$g[P(y \leq s|\boldsymbol{\eta})] = \alpha_s - v \quad s = 1, \ldots, S-1$$

where α_s with $0 < \alpha_1 < \ldots < \alpha_{S-1}$ are the thresholds to be estimated. The main feature of this model is that the effect of parameters are invariant to the choice of categories for y. Furthermore, each cumulative logit has its

[8]The identification topic of the models that are presented is not addressed.

own intercept. Typical choices of link function include the probit, logit and complementary log-log.

Alternatively, the categories of an ordinal variable can also be interpreted as realisations of an underlying continuous variable $y* = v + e$ related to y through the following relation:

$$y = s \quad if \quad \alpha_{s-1} < y* < \alpha_s \quad \alpha_0 = -\infty, \alpha_1 = 0, \alpha_s = \infty$$

so:

$$P(y \leq s|\boldsymbol{\eta}) = P(y* \leq \alpha_s|\boldsymbol{\eta}) = P(e \leq \alpha_s - v|\boldsymbol{\eta}).$$

The logit, probit, and complementary log-log links correspond to specifying, respectively, $f(e) = exp(-e)[1+exp(-e)]^{-2}$, $f(e) = (2\pi)^{-1/2}exp\left(-\frac{1}{2}e^2\right)$, and $f(e) = exp(e - exp(e))$ (Rabe-Hesketh et al., 2004*a*). If the variance of e is identifiable, scaled versions of these densities, corresponding to scaled links, can be used.

Alternative logit models can be used to model ordinal responses: one model is the adjacent-categories logit. Models using adjacent-categories logits can be expressed as baseline-category logit models (Agresti, 2002), with adjusted model matrix and a single parameter for each predictor. Let $s, s = 1, \ldots, S$ be the category of the ordinal response y, the odds for adjacent categories are expressed by:

$$\frac{P(y = s)}{P(y = s - 1)} = exp(m^s - m^{s-1} + \boldsymbol{x\beta})$$

and the linear predictor of the adjacent category logit is:

$$v^s = m^s + s\boldsymbol{x\beta}.$$

The adjacent category logit model assumes proportionality of the adjacent category odds, whereas the proportional odds model assumes proportionality of the cumulative odds.

As suggested by (Agresti, 2002, Ch. 7), "*the choice of model should depend less on goodness of fit than on whether one prefers effects to refer to individual response categories, as the adjacent-categories logits provide, or instead to groupings of categories using the entire scale or an underlying latent variable, which cumulative logits provides. Since effects in cumulative logit models refer to the entire scale, they are usually larger. The ratio of estimate to standard error, however, is usually similar for the two model types*".

1.3 Latent variables, continuous or categorical?

Latent variables are hypothetical constructs influencing in some way the observed realisation of a phenomenon. As for the observed variables, also the latent variables can have different "nature", in particular, they can be thought as continuous or categorical. Bartholomew and Knott (1999) proposed a four-fold classification of latent variable models based on the types of observed and latent variables[9], as shown in Table 1.2. The classical Factor analysis model is used to analyse phenomena characterised by continuous manifest and latent variables. When the manifest variables are categorical the Latent Class analysis and the Latent Trait analysis are obtained, depending on the nature of the latent variables.

Table 1.2: Classification of latent variable modeling.

	Manifest variables	
Latent variables	Continuous	Categorical
Continuous	Factor analysis	Latent Trait analysis
Categorical	Latent Profile analysis	Latent Class analysis

The origins of Factor analysis (or common factor analysis) can be found in the work of Galton and Pearson between the end of the 19th century and the beginning of the 20th century on the problem of inheritance of genetic traits (Kaplan, 2000). However, the work of Spearman (1904) on the underlying structure of mental abilities can be accounted for the development of the traditional factor model. Then, factor analysis gained popularity during the 1950s and 1960s, also, of course, for the development of statistical computing capacity; nowadays, it remains a popular methodology in quantitative social science research.

On the other hand, in the last decades, Latent Class analysis[10] has become a widely used technique in social and behavioral research.

In the standard formulation, Latent Class analysis investigates phenomena where the manifest variables are categorical and it is assumed the ex-

[9]Actually, Bartholomew and Knott (1999) classified the variables, both latent and manifest, in categorical variables or metrical variables. Categorical variables assign units to one of a set of categories, ordered or unordered; metrical variables have realized values in the set of real numbers and may be categorical or continuous. In this dissertation, the metrical variables will be assumed continuous.

[10]For a technical overview of the latent class analysis see for example Bartholomew and Knott (1999), Heinen (1996) and Goodman (1974). For some examples on the applied use of the latent class models see Hagenaars and McCutcheon (2002).

istence of a categorical latent variable. The levels of a categorical latent variable are called *classes*, representing a mixture of unobservable (latent) subpopulations where membership is not known but is inferred from the data. The goal of the analysis is to identify the nature and the number of latent classes. Cases within the same latent class are homogeneous on certain criteria, while cases in different latent classes are dissimilar from each other.

Latent Class analysis was originally introduced by Lazarsfeld in 1950s as a way of explaining respondent heterogeneity in survey response patterns involving dichotomous items (Vermunt and Magidson, 2005*b*). During the 1970s, LC methodology was formalized and extended to nominal variables by Goodman (1974), who developed the maximum likelihood algorithm. Over the same period, the related field of finite mixture models for multivariate normal distributions began to emerge, through the work of Day, Wolfe and others (Vermunt and Magidson, 2005*b*). Basically, finite mixture models have the same structure as the latent class; indeed, they seek to separate out data that are assumed to arise as a mixture from a finite number of distinctly different populations (McLahan and Peel, 2000). In recent years, the fields of latent class analysis and finite mixture modeling have come together and the terms latent class model and finite mixture model have become interchangeable.

As shown by Hagenaars and McCutcheon (2002), the applications of latent class models are numerous and various. For example, one interesting application refers to clustering of units, where each latent class represents a hidden cluster (Magidson and Vermunt (2002); Vermunt and Magidson (2003)), or dealing with measurement error in nominal and ordinal indicators.

On the difference between the use of continuous or categorical latent variables, Heinen (1996), p.27, writes "*It is, however, doubtful whether this difference between continuous and discrete latent traits is important from a more pragmatic point of view. First, there are a number of latent trait models that can be expressed as log-linear models. This means that the estimation of the parameters in the log-linear model will lead to the same results as estimation of the parameters in the (continuous) latent trait model. Second, when one tries to estimate the latent scores on the basis of the estimate parameters in some latent trait model and the observed response patterns (...) the "measurement" of latent variables is, in practice, always discrete. (...) Besides these practical considerations, an important question is whether latent class models and latent trait models will yield different results when applied to the same set of data.*".

Other authors showed the similitudes between the use of continuous and categorical latent variables.

Aitkin (1999) (see also Vermunt and Van Dijk (2001)) shows, in the context of variance component models, that a continuous latent distribution can be approximated by a nonparametric specification. In particular, he shows that a finite mixture distribution results from the discretization of the continuous latent variable distribution into K probability masses π_k at mass points z_k. The nonparametric specification is so represented by a finite mixture model with the maximum number of identifiable latent classes. An advantage of this approach is that it is not necessary to introduce possible inappropriate and unverifiable assumptions about the distribution of the random effects, so avoiding the bias in the item parameters estimates due to misspecification of the distribution of the continuous latent variables.

"*This suggests that the distinction between continuous and discrete latent variables is less fundamental than one might think, especially if the number of latent classes is increased*" (Vermunt and Magidson, 2005*a*, p. 4).

In the factor model framework, Muthén (2001) shows an interesting approach to the use of latent variables. He analyses the same dataset assuming first the existence of some continuous latent factors, then the existence of one categorical latent variable. Through this analysis, he shows the "empirical" connection between factor and latent class analysis, given by the different aims they are usually used for. Of course, the two models with the relative results can be combined in order to understand in a better way the phenomenon analysed.

In this thesis, the term *factor analysis* is used to refer to models with all continuous latent variables and the term *mixture factor analysis* is used to refer to models with both continuous and categorical latent variables, regardless of the nature of the observed variables. Of course, which model should be selected depends on the specific research, in particular on the substantive reason to believe in the nature, continuous or categorical, of latent variables.

The aim of the dissertation is to analyse the use of the latent variables in the multilevel framework, where there is a hierarchical structure of the data. The techniques that are illustrated refer to the well known traditional factor model with continuous dimensions, with the extension to the use of categorical latent variables.

Assuming two levels of latent variables and taking into account that the latent variables at each level may be continuous, categorical, or combination of these, the nine-fold classification (Vermunt, 2007) provided in Table 1.3 is obtained.

Model A1, in which both the lower and higher level latent variables are continuous, is represented by the multilevel factor model, as described by Goldstein and McDonald (1988) and Longford and Muthén (1992); its exten-

Table 1.3: Matrix of potential two-level models with underlying latent variables.

	Higher level latent variables		
Lower level latent variables	Continuous	categorical	Combination
Continuous	A1	A2	A3
categorical	B1	B2	B3
Combination	C1	C2	C3

sion to ordinal indicators is given by Skrondal and Rabe-Hesketh (2004) and Grilli and Rampichini (2007*a*). Model A1 contains also three-level regression models with continuous random effects. Model B2, in which both the lower and higher level latent variables are categorical, is the multilevel latent class model (Vermunt, 2003). In this case, lower level units are clustered based on their observed responses and higher level units (groups) are clustered based on the likelihood of their members to be in one of the unit level clusters. Vermunt (2003) also proposes a multilevel latent class model with continuous random effects at the group level (B1). Palardy and Vermunt (2007) used specification A3 to define a multilevel extension of the mixture growth model (Muthén, 2004), where two-level units are classified into homogeneous groups based on properties of their mean growth trajectories.

This brief and incomplete review of the literature shows how modeling using a combination of continuous and categorical latent variables provides an extremely flexible framework of analysis. Furthermore, different traditions such as growth modeling, multilevel modeling, latent class analysis are brought together using the unifying theme of latent variables. Which specification should be selected depends on the specific application; that is, whether it is more meaningful and/or practical to define the latent variables at a particular level to be continuous, categorical, or a combination of the two.

In the next section the models A1 and A2 are illustrated. At the first level of the analysis the standard factor model with continuous variables is implemented, while at the second level the latent variables are either continuous (model A1, section 1.3.1) or categorical (model A2, section 1.3.2).

Since the theoretical literature is well developed (section 1.1), in this dissertation the principal aim is to illustrate some specific models from their interpretational point of view in order to give some guidelines in their application. For each model, after a brief review of its analytical specification, some interpretational peculiarities are highlighted.

It should be noted that, in the implementation and estimation of a model, particular attention is necessary on its identification. Unfortunately this topic did not receive the proper attention in the literature, especially in the multilevel framework. Vermunt (2005), however, indicates that in multilevel mixture models identification is not a big problem as long as the number of level 1 units per level 2 units is not extremely small (larger than 3). For further details on and for a general discussion of this topic we refer to Skrondal and Rabe-Hesketh (2004).

1.3.1 Two-level factor model

The basic idea of the standard factor model is to find a set of latent factors, fewer in number than the observed variables, that contain essentially the same information of a given set of observed variables. In particular, the latent factors are supposed to account for the dependencies among the response variables: if the factors are held fixed, the observed variables would be independent.

Let y_{hij} denote the observed response on indicator h $(h = 1, \ldots, H)$ of individual i $(i = 1, \ldots, n_j)$ within group j $(j = 1, \ldots, J)$, and let υ_{hij} be the linear predictor of the response model.

As shown in section 1.2, the conditional expectation of the response y_{hij} given the latent variables at different levels is "linked" to the linear predictor υ_{hij} via a link function:

$$g(E(y_{hij}|\boldsymbol{\eta}_j)) = \upsilon_{hij}$$

where $\boldsymbol{\eta}_j = (\boldsymbol{\eta}_j^{(2)\prime}, \ldots, \boldsymbol{\eta}_j^{(L)\prime})'$, $\boldsymbol{\eta}_j^{(l)} = (\eta_{j1}^{(l)}, \ldots, \eta_{jM_l}^{(l)})'$ and M_l denotes the number of latent variables at level l. Again, different distributional forms for each indicator are allowed.

For each i in group j, the standard factor model is expressed by:

$$\upsilon_{hij} \;\; = \;\; \mu_{hj} + \sum_{m=1}^{M_2} \lambda_{mh}^{(2)} \eta_{mij}^{(2)} \qquad (1.7)$$

where $\eta_{mij}^{(2)}$ denotes the common factor(s), and $\lambda_{mh}^{(2)}$ represents the related factor loadings. The item-specific errors $e_{hij}^{(2)}$ implied by the response distribution are implicit[11]. In the standard factor model it is usually assumed that $\boldsymbol{\eta}_j^{(2)}$ are independent and identically distributed with:

$$\boldsymbol{\eta}_j^{(2)} \;\; \sim \;\; MN(\mathbf{0}, \boldsymbol{\Psi}^{(2)})$$

[11]The variance-covariance matrix of $e_{hij}^{(2)}$ depending on the link function is denoted by $\Omega^{(2)}$.

where $\mathbf{\Psi}^{(2)}$ is an $M_2 \times M_2$ matrix with elements $\psi^{(2)}_{mm'}$.

It is also usually assumed that the H observed variables are independent of each other given the latent variables $\eta^{(2)}_{mij}$, commonly referred to as the local independence assumption (Bartholomew and Knott, 1999). "*But it is misleading to think of it as an assumption of the kind that could be tested empirically because there is no way in which the latent variables can be held fixed and therefore no way in which the independence can be tested. It is better regarded as a definition of what we mean when we say that the set of the latent variables is complete*" (Bartholomew and Knott, 1999, Ch. 1).

The two traditional approaches to factor analysis are exploratory (EFA) and confirmatory (CFA) factor analysis. The first, more traditional, does not require a priori hypotheses about how indicators are related to underlying factors or even the number of factors. In this context different factor structures with different correlation structures, generated by different rotations in the factor space, can yield the same joint distribution of the observed variables and there are no reasons to prefer one rotation, and hence one solution, to another. On the contrary, confirmatory factor analysis analyses a priori models in which both the number of factors and their correspondence to the indicators are explicitly specified. When prior knowledge is not available, a common way to get information on the model can derive from a previous exploratory factor analysis. Usually, the loadings resulted to be close to zero in an EFA are tested to be exactly zero by applying a confirmatory analysis[12].

As mentioned previously, in some situations, the standard assumption of independence of the observations or the assumption of simple random sampling is not appropriate (for example, students usually observed within classrooms and schools, and employees observed within companies). In these cases, since the subjects share common environments, experiences, and interactions, it is reasonable to assume that the observations within a group are more similar than observations of different groups and it is more appropriate to use a multilevel factor analysis. This model allows to analyse the factor structures underlying the phenomenon both at the unit and group level, in order to study their characteristics, similarities and peculiarities.

The variance decomposition is:

$$\mathbf{\Sigma}_{TOT} = \mathbf{\Sigma}_W + \mathbf{\Sigma}_B \tag{1.8}$$

where $\mathbf{\Sigma}_{TOT}$ represents the total variation (variance and covariance matrix) of $\boldsymbol{y}$, $\mathbf{\Sigma}_W$ represents the variation (variance and covariance matrix) of $\boldsymbol{y}$

[12]As noted by Kline (2005), some authors do not agree with this procedure: low loadings in EFA often account for relatively high proportions of the variance, so constraining them to zero in CFA may be too conservative.

within individuals and $\mathbf{\Sigma}_B$ denotes the variation of $\boldsymbol{y}$ *between* individuals. The diagonal elements of $\mathbf{\Sigma}_{TOT}$, $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_B$ are, respectively, $\sigma^2_{T,h}$, $\sigma^2_{W,h}$ and $\sigma^2_{B,h}$.

The covariance structure modeling assumes that the population covariance matrix $\mathbf{\Sigma}_{TOT}$ can be described by separate models for the between groups and within groups structure (Hox, 1995).

Assuming that different latent factors at the two levels affect the observed variables, a multilevel extension of the standard factor model (equation (1.7)) would be:

$$\mu_{hj} \quad = \quad \mu_h + \sum_{m=1}^{M_3} \lambda^{(3)}_{mh} \eta^{(3)}_{mj} + e^{(3)}_{hj}$$

where μ_h is the item mean for each indicator h and $\eta^{(3)}_{mj}$ and $e^{(3)}_{hj}$ represent, respectively, the common factors and the item-specific errors (*specificities* or *unique factor*) at the highest level of the analysis, and it is usually assumed:

$$\begin{aligned} \boldsymbol{\eta}^{(3)} &\sim MN(\mathbf{0}, \mathbf{\Psi}^{(3)}) \\ \boldsymbol{e}^{(3)} &\sim MN(\mathbf{0}, \mathbf{\Omega}^{(3)}) \end{aligned}$$

where $\mathbf{\Psi}^{(3)}$ is an $M_3 \times M_3$ matrix with elements $\psi^{(3)}_{mm'}$ and $\mathbf{\Omega}^{(3)}$ is an $H \times H$ matrix with elements $\omega^{(3)}_{hh'}$. Usually, while $\mathbf{\Psi}^{(3)}$ is unconstrained, $\mathbf{\Omega}^{(3)}$ is assumed to be diagonal.

For each i in group j, the global model is then:

$$v_{hij} \quad = \quad \mu_h + \sum_{m=1}^{M_2} \lambda^{(2)}_{mh} \eta^{(2)}_{mij} + \sum_{m=1}^{M_3} \lambda^{(3)}_{mh} \eta^{(3)}_{mj} + e^{(3)}_{hj}.$$

The model is represented in Figure 1.3.

In particular, the latent variables $\eta^{(2)}_{mij}$ affect the phenomenon at the unit level, while the latent variables $\eta^{(3)}_{mj}$ underly the observed variables at the group level or, in other words, the mean level of each indicator is affected directly by the higher level latent factors.

With two factor models, one for the between covariance matrix and the other for the within covariance matrix, the covariance matrices in equation (1.8) are so decomposed (Muthén, 1994):

$$\mathbf{\Sigma}_B = \mathbf{\Lambda}^{(3)} \mathbf{\Psi}^{(3)} \mathbf{\Lambda}^{(3)'} + \mathbf{\Omega}^{(3)}$$

and

$$\mathbf{\Sigma}_W = \mathbf{\Lambda}^{(2)} \mathbf{\Psi}^{(2)} \mathbf{\Lambda}^{(2)'} + \mathbf{\Omega}^{(2)}.$$

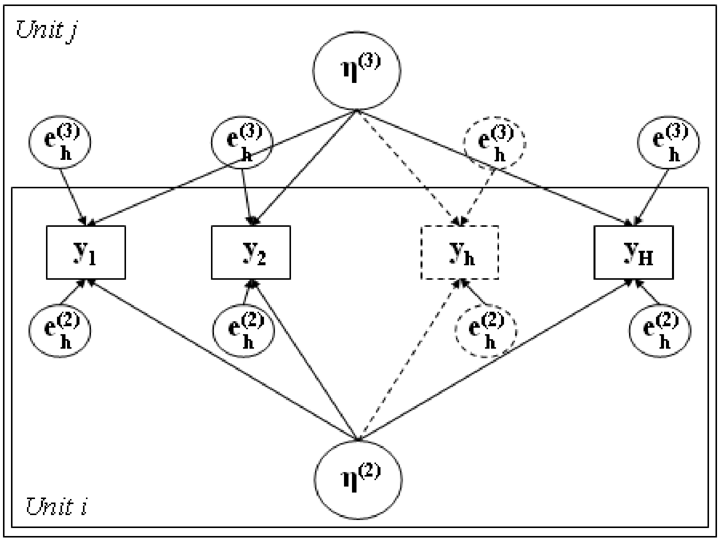

Figure 1.3: Two-level factor model.

A special case of the two-level factor model is the so called *variance component factor model* (Skrondal and Rabe-Hesketh, 2004) with the same number of common factors at both levels, no unique factors $e^{(3)}_{hj}$ and with factor loadings set equal across levels ($\lambda^{(2)}_{mh} = \lambda^{(3)}_{mh}$). The model is represented in figure 1.4.

In this case, the multilevel extension of equation (1.7) can be expressed as a null random intercept model for each latent variable $\eta^{(2)}_{mij}$ at the unit level:

$$\eta^{(2)}_{mij} \quad = \quad \eta^{(3)}_{mj} + e^{(2)}_{mij}$$

where

$$\boldsymbol{\eta}^{(3)} \quad \sim \quad MN(\mathbf{0}, \boldsymbol{\Psi}^{(3)}).$$

and μ_{hj} of equation (1.7) is equal to μ_h for every j.

The errors $e^{(2)}_{mij}$ represent the variability of each latent variables at the individual level. Indicating with $\boldsymbol{e}^{(2)}_{(m)}$ the error vector[13] with elements $e^{(2)}_{mij}$, it is usually assumed that $e^{(2)}_{mij}$ is uncorrelated with $\eta^{(3)}_{mj}$, and that it is multinormally distributed.

[13]For each i in group j, the vector $\boldsymbol{e}^{(2)}_{(m)}$ has dimension $M_2 \times 1$ and its elements are $e^{(2)}_{mij}$, while the vector $\boldsymbol{e}^{(2)}$ of the specificities at individual level has dimension $H \times 1$ and its elements are $e^{(2)}_{hij}$.

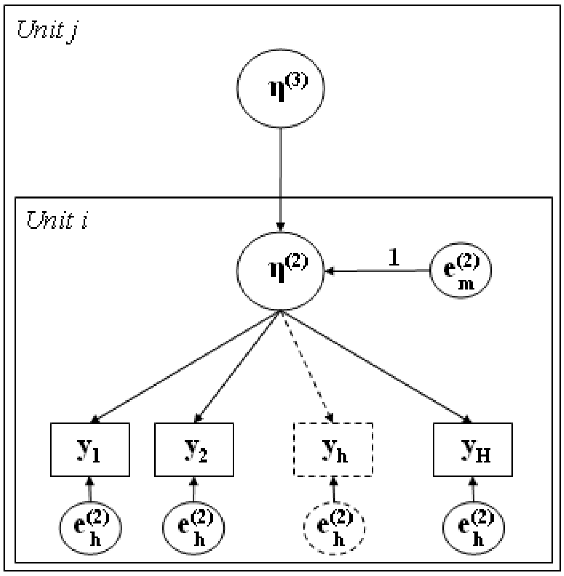

Figure 1.4: Two-level variance component factor model.

In the applications it is assumed that the mean level of the latent factors at the unit level varies between the clusters.

In factor models, the factor variances are not directly interpretable, as they represent contributions with respect to the arbitrary item that has the loading fixed to one (Grilli and Rampichini, 2007*a*). The interpretable quantity is the variance contribution expressed by the products: $\lambda^{(2)}_{mh}\psi^{(2)}_{m}$ and $\lambda^{(3)}_{mh}\psi^{(3)}_{m}$.

Other interpretable quantities are the so called *communalities* representing the proportion of the variance of a response explained by the factors. In L level models, the total communality of the h-th item is:

$$\frac{\sum_{l=2}^{L}\sum_{m=1}^{M_l}\sum_{m'=1}^{M_l}\lambda^{(l)}_{mh}\lambda^{(l)}_{m'h}\psi^{(l)}_{mm'}}{\sigma^2_{T,h}}$$

where $\psi^{(l)}_{mm}$ is the variance of the m-th factor at level l.

Then, the communality of the h-th item due to the l-level is:

$$\frac{\sum_{m=1}^{M_l}\sum_{m'=1}^{M_l}\lambda^{(l)}_{mh}\lambda^{(l)}_{m'h}\psi^{(l)}_{mm'}}{\sigma^2_{T,h}}$$

As illustrated by Grilli and Rampichini (2007*a*), in most two-level applications ($l = 2, 3$) the cluster-level item-specific errors $e^{(3)}_{hj}$ are constrained to zero, in order to save computational time and to avoid estimation problems, "*this simplification prevents a full variance decomposition and the computation of the related quantities, but it is expected to be of minor importance because the interest of the researcher centers on the factor structure*" (Grilli and Rampichini, 2007*a*, p. 13). In these situations, while the factor structure is unaffected, the variance decomposition (equation (1.8)) is no more feasible and the subject-level item-specific errors $e^{(2)}_{hij}$ represent the total item specificity. While the total communality of the h-th item and the communality of the h-th item due to the l-level can be still computed, other quantities cannot be computed, as the ICC_h and the item communalities at a given level of the analysis. The ICC_h is expressed by:

$$ICC_h = \frac{\sigma^2_{B,h}}{\sigma^2_{W,h}}$$

and the communality at subject level of h-th item is:

$$\frac{\sum_{m=1}^{M_2} \sum_{m'=1}^{M_2} \lambda^{(2)}_{mh} \lambda^{(2)}_{m'h} \psi^{(2)}_{mm'}}{\sigma^2_{W,h}}.$$

Then, all the estimable quantities in a multilevel factor model are scaled by $\omega^{(2)}_j$ (Grilli and Rampichini (2007*a*), see also Fielding (2004)). So, "*if cluster-level item-specific errors $e^{(3)}_{hj}$ are omitted, each scale factor ($\omega^{(2)}$) represents the square root of the item total specificity, leading to smaller estimable quantities. Nevertheless, the communalities are unaffected by the item scale, as they are ratios of parameters within the same item*" (Grilli and Rampichini, 2007*a*, p. 11).

1.3.2 Two-level mixture factor model

The multilevel mixture factor model combines elements of multilevel factor models and latent class models.

One limitation of factor models is that they assume that observations originated from a single population (Palardy and Vermunt, 2007); mixture models[14] are designed to examine this assumption and test whether unobserved subpopulations or latent classes are present.

[14]As mentioned in section 1.3, the terms mixture models and latent class models are used interchangeably. In this context the term mixture models is also used to refer to mixture factor models with more than one latent categorical variable.

The basic idea of a two-level mixture factor model with categorical latent variables at second level of the analysis[15] is that some parameters are allowed to differ across groups (classes) of second level units; in a random-effect approach the group-specific coefficients are assumed to come from a particular distribution, whose parameters should be estimated. Depending on whether the form of the mixing distribution is specified or not, either a parametric or nonparametric random-effect approach is obtained.

As described in section 1.1, the M_3 categorical latent variables at the group level are denoted by $\boldsymbol{\eta}_j^{(3)}$, each variable having, respectively, $K_1, K_2, \ldots, K_{M_3}$ categories.

At the individual level, there is the standard factor model, as expressed in equation (1.7):

$$v_{hij} = \mu_{hj} + \sum_{m=1}^{M_2} \lambda_{mh}^{(2)} \eta_{mij}^{(2)}.$$

One possible multilevel extension, with latent categorical variables at the highest level of analysis, is:

$$\mu_{hj} = \sum_{m=1}^{M_3} \sum_{k_m=1}^{K_m} \lambda_{mhk_m}^{(3)} \eta_{mjk_m}^{(3)} + e_{hj}^{(3)}. \tag{1.9}$$

The variable $\eta_{mjk_m}^{(3)}, m = 1, \ldots, M_3, k_m = 1, \ldots, K_m$ is an indicator variable taking the value 1 if unit i belongs to latent class k_m of the variable $\eta_{mj}^{(3)}$ and 0 otherwise[16]; the classes are mutually exclusive.

The variable $\boldsymbol{\eta}_j^{(3)} = (\eta_{1jk_1}^{(3)}, \ldots, \eta_{M_3jk_{M_3}}^{(3)})$ has a multivariate multinomial distribution, so:

$$\begin{aligned} \pi_{k_1,k_2,\ldots,k_{M_3}} &= P[\boldsymbol{\eta}_j^{(3)} = (k_1, k_2, \ldots, k_{M_3})] \\ &= P(\eta_{1j}^{(3)} = k_1, \ldots, \eta_{M_3j}^{(3)} = k_{M_3}) \end{aligned}$$

and

$$\pi_{k_1,k_2,\ldots,k_{M_3}} = \frac{\exp(\gamma_{k_1,k_2,\ldots,k_{M_3}})}{\sum_{k_1,k_2,\ldots,k_{M_3}} \exp(\gamma_{k_1,k_2,\ldots,k_{M_3}})} \tag{1.10}$$

[15] The categorical latent variables can be used at both level of the analysis. Since in the applications they are used only at the second level, only this "extension" of the two-level factor model is presented.

[16] As mentioned previously, some constraints on the $\lambda_{mhk_m}^{(3)}$ are necessary for the identification of the model. In this Chapter the identification topic is not addressed and some details on the identification of specific models are in Chapter 2.

with

$$\sum_{k_1,k_2,\ldots,k_{M_3}} \pi_{k_1,k_2,\ldots,k_{M_3}} = 1.$$

The γ term in equation 1.10 represents the linear predictor of the logit model for the expectation of the latent distribution ($\pi_{k_1,k_2,\ldots,k_{M_3}}$). In the case of multiple latent variables this linear term includes intercepts, bivariate associations, and possibly also higher-order interactions. Moreover, models with covariate effects on class membership can be defined by including covariate effects in this linear term.

The total number of classes of the joint distribution of $\boldsymbol{\eta}^{(3)}$ is $K_1 \times \ldots \times K_{M_3}$. In this model, the classes of each latent variable $\boldsymbol{\eta}_{mj}^{(3)}$ differ with respect to the item intercepts.

As for the multilevel factor model, it is possible to assume that the classes differ in the mean level of the latent factor(s) at the individual level, $\eta_{mij}^{(2)}$. Therefore, the alternative multilevel extension of equation (1.7) for each latent factor $\eta_{mij}^{(2)}$, $m = 1, \ldots, M_2$, is specified by:

$$E(\eta_{mij}^{(2)}) = \sum_{m'=1}^{M_3} \sum_{k_{m'}=1}^{K_{m'}} \lambda_{mm'k_{m'}}^{(3)} \eta_{m'jk_{m'}}^{(3)}. \tag{1.11}$$

and μ_{hj} of equation (1.7) is equal to μ_h for every j.

In the standard formulation, latent class models assume the existence of one categorical latent variable[17], whose levels are the classes representing a mixture of unobservable (latent) subpopulations.

In this case, the two-level extensions of equation (1.7) presented in equations (1.9) and (1.11) are, respectively:

$$\mu_{hj} = \sum_{k=1}^{K} \lambda_{hk}^{(3)} \eta_{jk}^{(3)} + e_{hj}^{(3)}$$

and, for each latent factor $\eta_{mij}^{(2)}$, $m = 1, \ldots, M_2$:

$$E(\eta_{mij}^{(2)}) = \sum_{k=1}^{K} \lambda_{mk}^{(3)} \eta_{jk}^{(3)}$$

with μ_{hj} of equation (1.7) equal to μ_h for every j.

[17] An interesting and clear discussion about the statistical differences between using one latent categorical variable with k classes, and more latent categorical variables is given by Magidson and Vermunt (2001), see also Magidson and Vermunt (2004).

The probability π_k is equal to:

$$\begin{aligned}\pi_k &= P(\eta_j^{(3)} = k) = P(\eta_{jk}^{(3)} = 1) \\ &= \frac{\exp(\gamma_k)}{\sum_{t=1}^{K} \exp(\gamma_t)}\end{aligned}$$

with

$$\sum_{k=1}^{K} \pi_k = 1.$$

In the traditional latent class analysis, the usual approach is to begin by fitting a 1-class model (independence) to the data and increasing the number of classes until an adequate fit is reached. An exploratory latent class analysis utilising latent class factor models can be used to determine the number of dimensions underlying the observed responses (Magidson and Vermunt, 2001).

Of course, the number of variables and the related number of classes depend on the specific peculiarities and aims of the research.

In the applications, the traditional approach to latent class analysis (with one categorical latent variable) is used.

1.4 Model fitting

1.4.1 Likelihood and estimation

Several estimation methods have been proposed for latent variable models. Skrondal and Rabe-Hesketh (2004) propose a classification of these methods according to the hypotheses on the randomness of the latent variables and/or the parameters. There are three general "classes": random latent variables and fixed parameters, fixed latent variables and parameters, random latent variables and parameters. The last includes the Bayesian approach, the second the well known "fixed effects" approach (one example being the classical ANOVA), the first the "classical approach". In this thesis, only some specific aspects of the first framework will be analysed.

As far as the classical framework with random latent variables and fixed parameters is concerned, recent computational developments have improved the applicability of maximum likelihood (ML) estimation. The importance of this development comes from the fact that maximum likelihood estimators have a number of nice theoretical properties under suitable regularity conditions. For example, the maximum likelihood estimators are consistent, asymptotically normal, and asymptotically efficient. Then, they are invariant under transformations and retain some properties under the assumption of Missing At Random (MAR), namely when "*the probability that a response is missing does not depend on the value of the response had it been observed, although it may depend on covariates included in the model and other responses*" (Skrondal and Rabe-Hesketh, 2004, Ch. 8).

When the latent variables are treated as random and parameters as fixed inference is usually based on the marginal likelihood, that is the likelihood of the observed data marginal to all latent variables.

In multilevel models with L levels, the total marginal likelihood equals:

$$L(\boldsymbol{\theta}) = \prod f^{(L)}(\boldsymbol{y}_{(L)}|\boldsymbol{\theta})$$

where the product is over all top-level clusters and $\boldsymbol{\theta}$ represents the complete set of unknown parameters to be estimated.

The marginal likelihood is constructed recursively. Defining $\boldsymbol{\eta}^{(l+)} = (\boldsymbol{\eta}^{(l)}, \boldsymbol{\eta}^{(l+1)}, \ldots, \boldsymbol{\eta}^{(L)})'$, the conditional density of a level-l unit, conditional on the latent variables at levels $l+1$ and above, is equal to:

$$\begin{aligned} f^{(l)}(\boldsymbol{y}_{(l)}|\boldsymbol{\eta}^{((l+1)+)};\boldsymbol{\theta}) &= \\ \int_{\boldsymbol{\eta}^{(l+)}} f^{(l)}(\boldsymbol{\eta}^{(l)}|\boldsymbol{\eta}^{((l+1)+)}) \quad \prod \quad & f^{(l-1)}(\boldsymbol{y}_{(l-1)}|\boldsymbol{\theta}, \boldsymbol{\eta}^{(l+)}) d\boldsymbol{\eta}^{(l)}. \end{aligned} \tag{1.12}$$

This recursive relationship can be used to build up the likelihood, increasing l from 2 to $L-1$. The conditional distribution of the responses of a level L unit then is:

$$f^{(L)}(\boldsymbol{y}_{(L)}|\boldsymbol{\theta}) = \int_{\boldsymbol{\eta}^{(L)}} f^{(L)}(\boldsymbol{\eta}^{L}) \prod f^{(L-1)}(\boldsymbol{y}_{(L-1)}|\boldsymbol{\theta}, \boldsymbol{\eta}^{(L)}) d\boldsymbol{\eta}^{(L)}. \tag{1.13}$$

The conditional density (or probability) of a response of a level-1 unit (denoted as $f(\boldsymbol{y}_{(1)}|\boldsymbol{\eta}^{(2+)};\boldsymbol{\theta})$), depends on the response process (section 1.2). When the latent variables are categorical the multiple integrals in equations (1.12) and (1.13) are replaced by multiple sums.

The likelihood of multivariate two-level models with continuous variables at the first level and continuous or categorical variables at the second level is now illustrated.

In two-level models, the total marginal likelihood equals:

$$L(\boldsymbol{\theta}) \;=\; \prod_{j=1}^{J} L_j(\boldsymbol{\theta}) = \prod_{j=1}^{J} f^{(j)}(\boldsymbol{y}_{(j)}|\boldsymbol{\theta})$$

where L_j indicates the likelihood of group j, the groups are assumed to be independent and $\boldsymbol{\theta}$ represents the complete set of unknown parameters to be estimated. The complete likelihood can be derived in more steps.

In a model with $\boldsymbol{\eta}^{(2)}$ and $\boldsymbol{\eta}^{(3)}$ being, respectively, continuous latent variables at the first and second level of the analysis, the likelihood for each group j is given by:

$$L_j(\boldsymbol{\theta}) = \int_{\boldsymbol{\eta}^{(3)}} \prod_{i=1}^{n_j} L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(3)}) f(\boldsymbol{\eta}^{(3)}) d\boldsymbol{\eta}^{(3)}$$

where the n_j level-1 units within level-2 units are assumed to be independent given the random coefficients $\boldsymbol{\eta}^{(3)}$.

Then, for each first-level unit, controlling for the effect of the latent variables at the highest level, the likelihood is expressed by:

$$L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(3)}) = \int_{\boldsymbol{\eta}(2)} L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}) f(\boldsymbol{\eta}^{(2)}|\boldsymbol{\eta}^{(3)}) d\boldsymbol{\eta}^{(2)}.$$

Finally, considering the *local independence assumption*, the observed indicators are assumed to be independent given the latent variables, so:

$$L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}) = \prod_{h=1}^{H} f(y_{hij}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)})$$

where $f(y_{hij}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)})$ indicates the distribution of the response variables (section 1.2).

When the latent variables are categorical the multiple integrals are replaced by multiple sums. In a model with $\boldsymbol{\eta}^{(3)}$ and $\boldsymbol{\eta}^{(2)}$ being, respectively, M_3 categorical and M_2 continuous latent variables, the likelihood is expressed by:

$$\begin{aligned} L_j(\boldsymbol{\theta}) &= \sum_{k_1} \dots \sum_{k_{M_3}} [P(\eta_{1j}^{(3)} = k_1, \dots, \eta_{M_3 j}^{(3)} = k_{M_3})] \\ &\qquad \prod_{i=1}^{n_j} L_{ij}(\boldsymbol{\theta}|\eta_{1j}^{(3)} = k_1, \dots, \eta_{M_3 j}^{(3)} = k_{M_3}) \\ L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(3)}) &= \int_{\boldsymbol{\eta}(2)} L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}) f(\boldsymbol{\eta}^{(2)}|\boldsymbol{\eta}^{(3)}) d\boldsymbol{\eta}^{(2)} \\ L_{ij}(\boldsymbol{\theta}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}) &= \prod_{h=1}^{H} f(y_{hij}|\boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}) \end{aligned}$$

In a model with only one categorical latent variable $\boldsymbol{\eta}^{(3)}$ at the highest level of the analysis with K categories, the likelihood for each j is expressed by:

$$L_j(\boldsymbol{\theta}) \;=\; \sum_{k=1}^{K} P(\eta^{(3)} = k) \prod_{i=1}^{n_j} L_{ij}(\boldsymbol{\theta}|\eta^{(3)} = k)$$

Maximum Likelihood estimation involves finding the estimates for $\boldsymbol{\theta}$ that maximize the marginal likelihood function (or the log-likelihood function). In maximizing the likelihood, two separated problems must be considered: solving the integrals involved in the likelihood and maximizing the likelihood function.

Relating to the first aspect, there are in general no closed forms for the multidimensional integral involved in the equations (1.12) and (1.13). A closed form expression for these integrals is available when all response and latent variables are continuous and normally distributed (Vermunt and Magidson, 2005c). In the other cases, as shown by Skrondal and Rabe-Hesketh (2004), there are several approaches to approximating the integrals, as Laplace approximation, numerical integration using quadrature or adaptive quadrature, Monte Carlo integration.

Relating to the second aspect, several methods were proposed for maximizing the likelihood, the most common being the Expectation-Maximization (EM) algorithm and Newton-Raphson or Fisher scoring algorithms; another algorithm is the so called BHHH. Each integration method may be combined with some maximization method(s).

In this thesis, only the numerical integration and some methods used to maximize the likelihood are briefly illustrated, referring to Skrondal and Rabe-Hesketh (2004) for a detailed discussion of the other techniques.

Numerical integration, also known as "quadrature", approximates an integral by a weighted sum of the integrand function evaluated at a set of values of the variable being integrated out. The weights are usually called *nodes* and the points at which the function is evaluated *quadrature points*.

With the numerical integration the integral is approximated by interpolating functions which are easy to integrate and a large class of quadrature rules can be derived by constructing different interpolating functions. Typically these interpolating functions are polynomials and, depending on the degree of the polynomial and the space of the interpolation points (equally spaced or not), different "rules" for the numerical integration are obtained.

Gaussian quadrature formulas are an example of numerical integration when the integration domain is the entire axis and the intervals between the interpolation points are allowed to vary.

For one-dimensional integral, the Gauss-Hermite quadrature states:

$$\int_{-\infty}^{+\infty} f(x)\phi(x)dx \approx \sum_{r=1}^{R} f(x_r)p_{x_r}$$

where $f(x)$ is an arbitrary function, $\phi(x)$ is the density of a standard normal distribution, R is the number of the quadrature points and $(x_r, p_{x_r}), r = 1, \ldots, R$, are respectively the quadrature points and their weights. For the standard normal univariate density, optimal points and weights are given in Stroud and Sechrest (1966). If the function $f(x)$ is well approximated by a polynomial of order $2R - 1$, then a quadrature with R nodes suffices for a good estimate of the integral (Skrondal and Rabe-Hesketh, 2004). The fundamental choice is relative to the value of R: the higher this value the higher the goodness of approximation, but also the computational efforts. In many cases, values between 5 and 10 represent a good compromise.

In case of a multidimensional integral (with q dimensions), each quadrature point becomes a q-dimensional vector:

$$\boldsymbol{x}_r = (x_{r1}, \ldots, x_{rq})$$

which weight is given by the corresponding unidimensional weights:

$$p_{\boldsymbol{x}_r} = \prod_{h=1}^{q} p_{x_{rh}}.$$

As the number of dimensions q is increased, the terms in the summation needed to approximate the integral solution increase exponentially (R^q). This is a disadvantage of this procedure: the computational efforts become too much large even with "small" values of q.

"*Fortunately, the number of points in each dimension can be reduced as the dimensionality is increased without impairing the accuracy of the approximations. Thus, factor analysis with five factors can be performed with good accuracy with as few as three points per dimension*" (Bock et al., 1988, p. 263). For example, in an application with $q = 5$ a good solution can be obtained with $R = 3$, obtaining a total number of quadrature points equal to $3^5 = 243$ (instead of $5^5 = 3125$ with $R = 5$).

Because of the multiple sums necessary to solve multidimensional integrals of equations (1.12) and (1.13), the likelihood function of a model with continuous latent variables is very similar to the likelihood function of a latent class model with multiple latent variables.

Each integration method may be combined with some maximization methods in order to maximize the total marginal likelihood.

There are several methods for maximizing the likelihood, the most common being the Expectation-Maximization (EM) algorithm and Fisher scoring or Newton-Raphson algorithms.

The EM approach has been introduced for the first time by Dempster et al. (1977) as an iterative algorithm based on the maximum likelihood estimation for models with missing information; in the latent variable framework the missing pieces of information are the values of the latent variables. Let $\boldsymbol{C} = (\boldsymbol{y}, \boldsymbol{\eta})$ be the complete data, with $\boldsymbol{y}$ being the uncomplete observed data and $\boldsymbol{\eta}$ the unobservable or latent data. The complete data log-likelihood, imaging that the latent data were observed, is denoted $\log L^c = \log L(\boldsymbol{\theta}|\boldsymbol{C})$.

The EM algorithm includes two steps (Skrondal and Rabe-Hesketh, 2004):

E-step: Evaluate the posterior expectation $\boldsymbol{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^k) = E_{\boldsymbol{\eta}}[\log L^c|\boldsymbol{y}; \theta^k]$.

M-step: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$ with respect to $\boldsymbol{\theta}$ to produce an updated estimate $\boldsymbol{\theta}^{k+1}$.

Sometimes closed form solutions are available in the M-step. In other cases, standard iterative methods can be used (see later). Latent GOLD uses iterative proportional fitting and unidimensional Newton in the M-step (Vermunt and Magidson, 2005*c*).

A review of the EM algorithm adapted to the multilevel framework is given in Vermunt (2003).

The advantage of the EM algorithm is that each iteration increases the likelihood and if it converges it converges to a local maximum or saddle point (Skrondal and Rabe-Hesketh, 2004). On the other side, the convergence of the EM algorithm can be very slow and it does not produce an estimates of the standard errors for the maximum likelihood estimate of $\boldsymbol{\theta}$.

The Fisher scoring and the Newton-Raphson algorithms are iterative methods used to maximize the log-likelihood. The two algorithms can be derived by considering an approximation of the derivatives of the log-likelihood using a first order Taylor series expansion around the current parameter estimates $\boldsymbol{\theta}^m$.

Let $\boldsymbol{\theta}$ be the vector of all parameters, and $\boldsymbol{\theta}_m$ the value of $\boldsymbol{\theta}$ at the m-th iteration. With the Newton-Raphson algorithm the parameters are updated through:

$$\boldsymbol{\theta}^{m+1} = \boldsymbol{\theta}^m - \boldsymbol{H}(\boldsymbol{\theta}^m)^{-1}\boldsymbol{g}(\boldsymbol{\theta}^m)$$

where $\boldsymbol{g}$ and $\boldsymbol{H}$ are, respectively, the gradient vector with the first-order derivatives and the Hessian matrix with the second-order derivatives of the log-likelihood function, evaluated at $\boldsymbol{\theta}^m$.

The Fisher scoring algorithm works in a very similar way, using the negative of Fisher's information matrix $\boldsymbol{I}(\boldsymbol{\theta}^m)$ in the place of the Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta}^m)$:

$$\boldsymbol{\theta}^{m+1} = \boldsymbol{\theta}^m + \boldsymbol{I}(\boldsymbol{\theta}^m)^{-1}\boldsymbol{g}(\boldsymbol{\theta}^m)$$

where $\boldsymbol{I}(\boldsymbol{\theta}^m) = -E(\boldsymbol{H}(\boldsymbol{\theta}^m))$.

An advantage of Newton-Raphson and Fisher scoring algorithms compared with EM is that they provide estimates of the standard errors[18] for the maximum likelihood estimate of $\boldsymbol{\theta}$. Indeed, the matrix $-\boldsymbol{H}$, usually referred to as the "observed information" matrix and used as an approximation of the expected information matrix, evaluated at the final $\hat{\boldsymbol{\theta}}$, gives $\hat{\Sigma}_{std} = -\boldsymbol{H}(\hat{\boldsymbol{\theta}})^{-1}$.

Both the Newton-Raphson and Fisher scoring algorithms use second order derivatives of the log-likelihood with respect to the parameters. Computing these derivatives analytically can be difficult and computing them numerically can be very slow, so other algorithms (*quasi-Newton algorithms*) have been proposed to overview these problems. For example, the BHHH algorithm (described by Berndt, Hall. B.H., Hall R.E. and Hausman in 1974) updates the parameters values through:

$$\boldsymbol{\theta}^{m+1} = \boldsymbol{\theta}^m + \boldsymbol{I}_{BHHH}(\boldsymbol{\theta}^m)^{-1}\boldsymbol{g}(\boldsymbol{\theta}^m)$$

where $\boldsymbol{I}_{BHHH}(\boldsymbol{\theta}^m)$ is a consistent estimator of the covariance matrix of the gradient vector $E(\boldsymbol{g}(\boldsymbol{\theta}^m)\boldsymbol{g}(\boldsymbol{\theta}^m)')$ and under correct model specification:

$$E(\boldsymbol{g}(\boldsymbol{\theta}^m)\boldsymbol{g}(\boldsymbol{\theta}^m)') = -E(\boldsymbol{H}(\boldsymbol{\theta}^m) = \boldsymbol{I}(\boldsymbol{\theta}^m).$$

The software Latent GOLD uses both a combination of EM and Newton-Raphson algorithm to find the Maximum Likelihood or Posterior Mode estimates (section 1.5) for the model parameters. In particular, the estimation process starts with a number of EM iterations and, when close enough to the final solution, the program switches to the Newton-Raphson algorithm. In solving the integrals of the likelihood function, Latent GOLD approximates the conditional density by means of Gauss-Hermite numerical integration.

The algorithm used to maximize the loglikelihood and the number of integration points used to approximate the multidimensional integrals effect the computational time necessary to estimate a generalized latent variable model. This depends also on other factors, for example the nature and number of the response variables and the number of latent variables.

[18] Another approach to the computation of an estimate of the standard errors for the maximum likelihood estimate of $\boldsymbol{\theta}$ is the so-called robust, sandwich, or Huber-White estimator (Vermunt and Magidson, 2005*c*). The advantage of this method is that, contrary to the other two, it does not rely on the assumption that the model is correct.

1.4.2 Model evaluation

As underlighted by Kline (2005), potential mistakes can be done relating to the (mis)interpretation of statistical models. Some of these are:

- look only at indexes of overall model fit ignoring other types of information about model fit that can show some problems in specific portion of the model,
- rely solely on statistical criteria or tests in model evaluation,
- fail to consider equivalent or alternative models.

In evaluating a model, substantive and statistical aspects have to be accounted for; in this section some statistical aspects are illustrated.

A number of overall and individual statistical measures of fit has been proposed in order to evaluate a specified model on the basis of empirical data and nowadays the evaluation of the statistical properties of the fit indexes in computer simulation studies is an active topic (Kline, 2005).

As a first evaluation of a model, overall fit indexes (with the related tests) are used in order to determine how well a proposed model fits a particular data set. Unfortunately, the use of these indexes has some limitations, as illustrated by Kline (2005). For example values of fit indexes indicate only the average or overall fit of a model, so it is possible that some specific parts of the model may poorly fit the data; usually a single index reflects only a particular aspect of model fit so a favorable value of an index does not indicate by itself a good overall fit; fit indexes do not indicate whether the results are theoretically meaningful; etc.

Maybe more interesting than the overall evaluation of a model is the comparison of nested and not nested models evaluated with the same data. This is assessed with the use of specific tests (nested models) or with comparative indexes (non nested models).

For the overall evaluation of a model the procedure is to first compute a test statistic based on the deviation of the model (with the parameter estimates) from the data and then compare the statistic to a theoretical or empirical distribution based on the assumption that the model is true. A rough probability of observing the particular data set, given the model is true, can then be determined: if the probability of observing the data is too low, the model is rejected.

Two types of goodness of fit tests have been commonly employed: Chi-squared type tests and tests based on the empirical distribution function. Chi-square tests are used when data are grouped into discrete classes, and observed frequencies are compared to expected frequencies based on a model.

Tests based on the empirical distribution function are used most often with continuous data.

For a particular model, $L(\boldsymbol{\theta})$ denotes the likelihood function expressed in terms of the parameters $\boldsymbol{\theta}$ and $L(\hat{\boldsymbol{\theta}})$ denotes the maximum of the likelihood for the model (section 1.4.1). The maximum achievable likelihood is $L(\boldsymbol{y})$: this occurs for the *saturated* model, having the maximum number of parameters that can be estimated (Dobson, 2002). The loglikelihood ratio statistic for testing the null hypothesis that the model holds against the general alternative (saturated model) is (Agresti, 2002):

$$-2\log\frac{\text{maximum likelihood for the model}}{\text{maximum likelihood for the saturated model}} = D(L(\hat{\boldsymbol{\theta}}), L(\boldsymbol{y}))$$

Large values of the statistic suggest that the model of interest is a poor description of the data relative to the saturated model, in other words the statistic describes the lack of fit of the model.

The statistic $D(L(\hat{\boldsymbol{\theta}}), L(\boldsymbol{y}))$ is usually called *deviance* $D(\boldsymbol{\theta})$ and its sampling distribution is, approximately:

$$D(\boldsymbol{\theta}) \sim \chi^2_{(N-p,v)}$$

where v is the noncentrality parameter, p is the number of model parameters and N is the number of total observations.

There are some problems with the χ^2 statistics (Kline, 2005), the main one is that it is sensitive to the sample size. If the sample size is large, which is required in order to interpret the index as a test statistic, the value of χ^2 may lead to rejection of the model even though differences between the model and the data are slight. To reduce the sensivity of the Chi-squared statistic to sample size, some researchers divide its value by the degree of freedom. However, there is no clear-cut guideline about what value of this statistic should be acceptable. Furthermore, in factor models, the χ^2 statistics is sensitive to the size of the correlations among observed responses: bigger correlation generally lead to higher value of χ^2. This happens because larger correlations tend to allow for the possibility of greater differences between the observed and the model implied correlations. Then, if the distributions of the continuous indicators are non normal, the value of χ^2 tends to be too high: the true model is rejected too often.

When observed data are categorical, a usual way to assess the model fitting is to rely[19] on the Pearson X^2 or the likelihood ratio statistic G^2 based on the comparison between theoretical and observed frequencies.

[19]Latent GOLD reports chi-squared and related statistics; the three reported chi-squared measures are the likelihood-ratio chi-squared statistic G^2, the Pearson chi-squared statistic and the Cressie-Read chi-squared statistic. Furthermore, it reports the values of the log-likelihood, the log-prior, and log-posterior (Vermunt and Magidson, 2005*c*).

The Pearson X^2 is expressed by (Bartholomew and Tzamurani, 1999):

$$X^2 = \sum_i \frac{(Obs - Exp)^2}{Exp}$$

and the likelihood ratio G^2 (sometimes referred to as the deviance statistic) is:

$$G^2 = 2 \sum_i Obs \log \left(\frac{Obs}{Exp} \right)$$

where Obs and Exp are, respectively, the observed and expected frequencies of the i-th response pattern. When the null hypothesis holds, the Pearson X^2 and the likelihood ratio G^2 (sometimes referred to as the deviance statistic) both have asymptotic Chi-squared distributions with degree of freedom equal to $c-1$ minus the number of estimated parameters, where c represents the number of all possible response patterns (Cagnone, 2003) and the two statistics are asymptotically equivalent; when H_0 is false, they tend to grow proportionally to n.

When the number of observed variables is large, the average expected frequency become too small for the Chi-squared approximation of the two statistics sampling distribution to be valid. Furthermore, for sparse tables no general results concerning the distribution of G^2 or X^2 are available.

The problem may be overcome by pooling response patterns so that the expected values for the groups thus formed are large enough (bigger than 5) to justify the Chi-squared distribution. However, as described by Bartholomew and Tzamurani (1999) this way of proceed has some disadvantages.

In presence of small sample size or sparseness of data and in models containing order restrictions (see, respectively, Langeheine et al. (1996) and Galindo-Garre and Vermunt (2005)), it can be useful to estimate the p-value of the deviance statistic by means of a parametric bootstrap rather than relying on its asymptotic p-value. The model of interest is not only estimated for the sample under investigation, but also for some p replication samples (Vermunt and Magidson, 2005c) generated from the probability distribution defined by the maximum likelihood estimates. The estimated bootstrap p-value, $p_{boot.}$, is then defined as the proportion of bootstrap samples with a larger deviance than the original sample; the standard error of p_{boot} equals $\left[\frac{(p_{boot})(1-p_{boot})}{B}\right]^{-\frac{1}{2}}$.

When there is a small sample size or sparseness of data, another approach to the evaluation of the model fitting is the computation of bivariate residual (BVR) statistics or, in other words, the computation of the fit statistics G^2

or X^2 for pairs and triplets of responses. Of course, the analysis of these statistics cannot solve the problem of the evaluation of the entire model, but can suggest if the model lack in explaining the association of two variables[20]. Furthermore, an interesting approach suggests that the bivariate "*measures can be interpreted as lower bound estimates for the improvement in fit if corresponding local independence constraints were relaxed*" (Vermunt and Magidson, 2005*c*, Ch. 7), so they indicate whether the local independence assumption[21] is met.

For two categorical variables g and k with, respectively, categories $s = 1, \ldots, S$ and $r = 1, \ldots, R$, the X^2 and G^2 statistics are defined as (Jöreskog and Moustaki, 2001):

$$(X^2)^{(r,s)}_{(g,k)} = \frac{(Obs^{(r,s)}_{(g,k)} - Exp^{(r,s)}_{(g,k)})^2}{Exp^{(r,s)}_{(g,k)}}$$

$$(G^2)^{(r,s)}_{(g,k)} = 2Obs^{(r,s)}_{(g,k)} \log\left(\frac{Obs^{(r,s)}_{(g,k)}}{Exp^{(r,s)}_{(g,k)}}\right)$$

By summing these measures across the categories of each pair of variables the information of the goodness of fit of each cell of the two-way marginal table is obtained. Since these fit measures are based on different number of categories for different variables and different pairs of variables, Jöreskog and Moustaki (2001) suggest to divide them by the number of categories to make them comparable across variables and pairs of variables and to assume values larger than 4 as indicative of poor fit[22].

Latent GOLD reports bivariate residuals similar to Lagrange-multiplier tests that can be computed for all outcome variable types. In general, bivariate residual statistics larger than 3.84 identify correlations between the associated variable pairs that have not been adequately explained by the model (Vermunt and Magidson, 2005*c*).

[20]The importance of bivariate relationship is quite clear in traditional factor analysis where, through the assumption of multivariate normality, higher order relationships are assumed not to exist. In models with categorical latent variables, the bivariate associations are generally the most prominent, and the ability to individuate specific two-way tables in which lack of fit may be concentrated can be useful in suggesting alternative models (Magidson and Vermunt, 2004).

[21]A fundamental assumption of factor models is that the manifest variables are locally independent (section 1.3.1): that is, given the latent factors, the manifest variables should be statistically independent from each other.

[22]This rule of thumb has been obtained by referring to a χ^2 distribution with one degree of freedom.

More interesting then the evaluation of a model global fit, is the comparison of that model with others. In order to choose between different models, more techniques are available. Here some tests based on the likelihood theory and some information criteria are briefly introduced.

Let M_0 and M_1 be two models with, respectively, p_0 and p_1 ($p_0 > p_1$) parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Assume that M_1 is nested in M_0, so that $p_0 - p_1$ restrictions are imposed on the structural parameters[23].

The likelihood ratio test can be performed through the statistic:

$$D_{(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)} = -2[\log L(\hat{\boldsymbol{\theta}}_{M_1}) - \log L(\hat{\boldsymbol{\theta}}_{M_0})]$$

which, under regularity condition and under the restricted model M_1, is asymptotically Chi-squared distributed with $p_0 - p_1$ degree of freedom.

Other tests to compare nested models are the Wald test and Lagrange Multiplier[24]. These tests only require the estimation of one model (Wald test uses M_0 and Lagrange Multiplier uses M_1), and can be regarded as quadratic approximations to the likelihood ratio test statistic; the Wald test evaluates whether restrictions can be imposed on the estimated model, whereas Lagrange Multiplier tests whether restrictions can be removed. Even if the three tests (Likelihood Ratio, Wald and Lagrange Multiplier test) are asymptotically equivalent, in finite sample they have different behaviours. In particular, the Wald test performs poorly in the neighborhood of the parameters estimates and is not invariant to nonlinear transformations parameters. If the Wald and likelihood tests yield different results, the likelihood ratio test is preferable (Skrondal and Rabe-Hesketh, 2004).

Standard asymptotic results for the three tests do not hold if the null hypothesis is on the boundary of the parameter space since regularity conditions would be violated (Self and Liang, 1987). A well known example is testing the null hypothesis regarding random effect[25].

[23]Skrondal and Rabe-Hesketh (2004) in Chapter 4 define the *structural parameters* as parameters "presumed to have generated the observed data". In multilevel factor models, the vector of structural parameters contains $\boldsymbol{\beta}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\gamma}$ and the residual variances and covariances between latent variables and between dependent variables.

[24]For some details on these tests, see Pawitan (2001).

[25]In variance component models, when there is only one variance being set to zero in the reduced model, the asymptotic distribution of the likelihood-ratio test statistic is a 50:50 mixture of a χ^2_k and χ^2_{k+1} distribution, where k is the number of other restricted parameters in the reduced model that are unaffected by boundary conditions. In this case, a rule of thumb is to divide by two the asymptotic p-value of the Chi-squared likelihood ratio test statistic distribution (Rabe-Hesketh et al., 2004*a*). The asymptotic distribution of the likelihood ratio statistic may become considerably more complicated once more then one boundary parameter is tested (Self and Liang, 1987).

Also in the latent class models framework the likelihood ratio statistic cannot be used to compare two nested models, one with k_0 classes and one with k_1 classes ($k_0 < k_1$). Indeed, under the null hypothesis of k_0 groups some of the parameters of the model with k_1 classes lie on the boundary of the parameter space so that regularity conditions for likelihood ratio statistic to be asymptotically Chi-squared are not fulfilled. In particular, fixing the probability of one class to be zero makes the corresponding location nonidentified and setting the locations of two classes equal to one another implies that only the sum of the corresponding probabilities becomes identified. Therefore, the correct null distribution of the likelihood ratio statistic is unknown (Everitt, 1988) but a lot of conjectures and simulations have been published on this topic (McLahan and Peel, 2000). For example, Lo et al. (2001) developed, in one-level framework, an exact parametric likelihood ratio test[26] to determine the number of components in a normal mixture model, developing the results of Vuong (1989) in the multiple regression framework. Nylund et al. (2007) use the work of Jeffries (2003) that points out a flaw in the mathematical proof of the Lo Mendell Rubin test for normal outcomes (Lo et al., 2001) and explore with simulation studies some available tools for determining number of classes in latent class models in several ways. As an alternative, McLahan and Peel (2000) describe how to derive the empirical distributions of the likelihood ratio statistic over bootstrap replications in order to obtain approximate significance probabilities[27].

Another approach for comparing models is based on the computation of some indexes representing a penalized form of the likelihood: as the likelihood increases with the addiction of some parameters, it is penalized by the subtraction of a term that "penalizes" the model likelihood for the number of the parameters used. These information criteria are generally expressed in terms of:

$$-2\log L(\boldsymbol{\theta}) + 2C$$

where the first term measures the lack of fit of the model and C is the penalty term that measures the complexity of the model. The intent is therefore to choose a model to minimize this criterion.

[26]The software Latent GOLD does not provide the Lo Mendell Rubin test.

[27]The software Latent GOLD provides a bootstrap estimate of the p-value corresponding to the difference in log-likelihood value between two nested models, such as two models with different numbers of latent classes (Vermunt and Magidson, 2005*c*). Note that in multilevel framework this procedure is really computer demanding.

To date, there is no common acceptance of the best criteria for determining the number of classes; a number of indexes have been proposed[28]. These indexes are not treated here because they are outside of the scope of this thesis.

A variety of textbooks and articles suggest the use of the Bayesian Information Criterion (BIC) (Schwarz, 1978) as a good indicator for class enumeration (Nylund et al., 2007). BIC, also known as Schwarz's information criterion, is:

$$BIC = -2 \log L + Npar \times \log(N)$$

where $\log L$ is the loglikelihood value, $Npar$ is the number of parameters and N is the number of observations for the fitted model. BIC weights the model log-likelihood favoring more parsimonious models and smaller samples.

In the context of multilevel modeling, the number of observations to be used in BIC formula can be measured in various ways. In two level models the number of observations can refer to both within and between level; this distinction can make a substantial difference when determining the number of classes of a multilevel mixture model. When comparing multilevel models differing only at the between-level of analysis, as is the case of this thesis (see Chapter 3), using the number of between-level observations is better applicable (Palardy and Vermunt, 2007).

1.5 Posterior analysis

The main aim of the researcher using factor models is in what can be known about the latent variables after the manifest variables have been observed (Bartholomew and Knott, 1999). At each level of the analysis, this information is represented by the conditional density:

$$h(\boldsymbol{\eta}|\boldsymbol{y}) = h(\boldsymbol{\eta})g(\boldsymbol{y}|\boldsymbol{\eta})/f(\boldsymbol{y}).$$

From the point of view of social behavioral scientists, this means locating units on the dimensions of the latent space (finding the *factor scores*), or classifying units in different *classes* representing some typical profile. Obviously, units with the same response pattern will be assigned the same factor score or class.

When the latent structure assumes continuous latent variables, the aims that can be reached through assigning factor scores to l-level units are various (Skrondal and Rabe-Hesketh, 2004). For example, the researcher can

[28] For a detailed review of the indices available to determine the optimal number of classes in mixture models see McLahan and Peel (2000).

be interested in finding some scoring procedure, in order to not use latent variable modeling for other studies or he could be interested in classifying units respect to their level of the latent variable (for example intelligence, or illness status measured by multiple indicators in order to decide the dose of a specific drug). Then, factor scores are also useful in adaptive testing, where the scores are updates sequentially as new item responses are obtained in order to choose for the next item difficulty level or factor scores can be used as vehicle for further analysis.

Latent variables can also be assumed to be categorical; in this case the aim of the research is to assign each l-level unit to a latent class. The assumption on the categorical nature of the latent variable may derive from substantive reasons (for example in marketing segmentation studies (Bassi, 2007) or in studies relative to determine the presence of a certain illness (Magidson and Vermunt, 2004)), or may come from the necessity to make a decision (for example in test ability) or to use latent classifications as "observed" variables in subsequent studies.

In this thesis, some scoring methods based on the empirical Bayesian posterior distribution and a maximum likelihood method[29] are briefly illustrated[30]. Usually the Bayesian posterior distribution are the most used methods; the maximum likelihood methods is sometimes used since, in contrast to Empirical Bayes, the scores are conditionally unbiased. However, the maximum likelihood approach is not consistent with the modeling assumptions since it requires that the latent variables are considered fixed parameters and does not yield predictions for clusters with insufficient information.

With the empirical Bayesian approach, according to Bayes' theorem, the conditional posterior distribution of the latent variables given the observed variables is expressed by:

$$f(\boldsymbol{\eta}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \frac{f(\boldsymbol{y}, \boldsymbol{\eta}|\hat{\boldsymbol{\theta}})}{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})} \tag{1.14}$$

where $\hat{\boldsymbol{\theta}}$ represent the estimated parameters, $f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ is the distribution of the observed variables and $f(\boldsymbol{y}, \boldsymbol{\eta}|\hat{\boldsymbol{\theta}})$ is the joint distribution of the observed and latent variables. This approach uses the term "Bayesian" since both the latent and observed variables are treated as random variables. Actually, the full Bayesian approach would assume a prior distribution for $\boldsymbol{\theta}$ in addition to the distribution for $\boldsymbol{\eta}$ and the $\boldsymbol{\theta}$ in equation (1.14) would be treated as fixed constants.

[29]Latent GOLD only use scoring methods based on empirical Bayesian posterior distribution.

[30]For a detailed discussion of the topic see Skrondal and Rabe-Hesketh (2004).

Developing the equation (1.14) the posterior distribution is obtained:

$$f(\boldsymbol{\eta}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \frac{f(\boldsymbol{y}|\boldsymbol{\eta}, \hat{\boldsymbol{\theta}})f(\boldsymbol{\eta}|\hat{\boldsymbol{\theta}})}{\int_{\boldsymbol{\eta}} f(\boldsymbol{y}|\boldsymbol{\eta}, \hat{\boldsymbol{\theta}})f(\boldsymbol{\eta}|\hat{\boldsymbol{\theta}})}.$$

The computation of the posterior distribution is strictly related to the specification of the prior distribution of the latent variables. Usually, the posterior distribution cannot be expressed in closed form and heavy numerical integration is required (section 1.4.1).

In factor models with continuous random variables, it follows from standard results on conditional multivariate normal densities that the posterior density is multivariate normal. For other response types, the posterior density tends to multinormality as the number of units in the clusters increases (Skrondal and Rabe-Hesketh, 2004).

After estimating the empirical Bayesian posterior distribution, two approaches can be used to estimate the factor scores (or latent class) associated to each unit: the prediction using empirical Bayes (also called *a posteriori*) and prediction using empirical Bayes modal (also known as *modal a posteriori*).

The empirical Bayes prediction is the most widely used method for scoring. The predictors are represented by the mean of the posterior empirical Bayesian latent variables distribution in equation (1.14), so:

$$\boldsymbol{\eta}^{EB} = E(\boldsymbol{\eta}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}).$$

With continuous normal latent variables, the empirical Bayes predictor is the best linear unbiased predictor BLUP (Skrondal and Rabe-Hesketh, 2004).

The prediction using empirical Bayes modal uses the posterior mode instead of the poster mean for the prediction of the factor scores:

$$\boldsymbol{\eta}^{EBM} = ^{\text{max arg}}\boldsymbol{\eta}\ (\boldsymbol{\eta}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}).$$

The posterior mode is the solution of:

$$\frac{\partial}{\partial \boldsymbol{\eta}} lnf(\boldsymbol{\eta}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \mathbf{0}$$

that can be expressed by:

$$\frac{\partial}{\partial \boldsymbol{\eta}} lnf(\boldsymbol{\eta}|\hat{\boldsymbol{\theta}}) + \frac{\partial}{\partial \boldsymbol{\eta}} ln \prod f(y_{hij}|\boldsymbol{\eta}) = \mathbf{0}.$$

This method does not require numerical integration, so when the posterior density is approximately multivariate normal it is often used as an approximation of the empirical Bayes solutions. In particular, this method represents the standard classification method in latent class modeling since it minimize the expected misclassification rate (Skrondal and Rabe-Hesketh, 2004). Obviously, the predictors obtained with the empirical Bayes and empirical Bayes modal coincide in standard factor model.

Maximum likelihood estimation of the latent variables requires that the latent variables are considered fixed parameters. Since this assumption is inconsistent with the used model framework and the marginal likelihood method of parameter estimation, Skrondal and Rabe-Hesketh (2004) do not recommend the maximum likelihood scoring method. Furthermore, this approach does not yield predictions for clusters with insufficient information, since the prior distribution of the latent variables is not used. However, it may be useful for assessing the normality assumption for the latent variables.

Maximum likelihood approach estimates the latent variables $\boldsymbol{\eta}$ given $\hat{\boldsymbol{\theta}}$ maximising the conditional distribution of the responses with respect to the unknown latent variables:

$$\frac{\partial}{\partial \boldsymbol{\eta}} ln \prod f(y_{hij}|\boldsymbol{\eta}, \hat{\boldsymbol{\theta}}).$$

In multilevel factor models with continuous multinormal latent variables, the maximum likelihood estimator is conditionally unbiased, given the values of $\boldsymbol{\eta}$ and $\hat{\boldsymbol{\theta}}$; furthermore, as the number of units in a cluster tends to infinity the estimates for the clusters are asymptotically unbiased (Skrondal and Rabe-Hesketh, 2004).

In addition to the estimation of the latent scores, also their variances and covariances can be estimated with different methods[31]. These are not treated here because they are outside of the scope of this work; a good review of the methods and their relationship in the factor model can be found in (Skrondal and Rabe-Hesketh, 2004).

When dealing with latent categorical variables, the analysis of the so called *classification statistics* (Vermunt and Magidson, 2005*c*) can be useful.

This set of statistics contains information on how well the latent classes are separated. As already shown, usually classification is based on the empirical Bayesian posterior distribution of equation (1.14); then, the proportion of classification errors is defined as:

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

[31]Latent GOLD does not provide this analysis.

where i represent different response pattern.

Other classification statistics can be computed, the so called R^2-type measures for nominal variables: the proportional reduction of classification errors $R^2_{\boldsymbol{\eta},errors}$, a measure based on entropy $R^2_{\boldsymbol{\eta},entropy}$ and a measure based on qualitative variance $R^2_{\boldsymbol{\eta},variance}$. Each of the three $R^2_{\boldsymbol{\eta}}$ measures is based on the same type of reduction of error structure:

$$R^2_{\boldsymbol{\eta}} = \frac{Error(\boldsymbol{\eta}) - Error(\boldsymbol{\eta}|\boldsymbol{y})}{Error(\boldsymbol{\eta})}$$

where $Error(\boldsymbol{\eta})$ is the total error when predicting $\boldsymbol{\eta}$ without using information on $\boldsymbol{y}$, and $Error(\boldsymbol{\eta}|\boldsymbol{y})$ is the prediction error if all observed information are used[32].

Other information on the goodness of the classification come from the so called "Classification Table" (Vermunt and Magidson, 2005*c*). This table cross-tabulates the two assignments based on the empirical Bayesian posterior distribution: empirical Bayes and empirical Bayes modal. The entry $(\eta_{mk_m}, \eta'_{mk_m})$ contains the sum of the class η_{mk_m} posterior membership probabilities for the cases allocated to modal class η'_{mk_m}. The marginal distributions of the Classification Table show the distribution of cases across classes under the two classification methods[33]. The diagonal elements are the numbers of "correct" classifications per latent class and the off-diagonal elements are the corresponding numbers of misclassification. From the analysis of the classification table it is possible to investigate how many cases are misclassified, and which are the most common types of misclassification.

[32]For technical details on these measures see Vermunt and Magidson (2005*c*).

[33]Except for rare situations, these marginal distributions are not equal to one another, showing that the modal class assignment do not conserve the estimated latent class distribution.

Chapter 2

The evaluation of university performance, traditional analyses

The aim of this thesis is to evaluate the effectiveness of the university from the users' point of view.

Data used for the analysis come from two surveys of the consortium Almalaurea. Because of our knowledge of the context, we focus on the university of Florence. After a brief description of the AlmaLaurea surveys and the Italian university reform (section 2.1), the Chapter is divided into two sections: the first deals with the analysis of the internal effectiveness of the university in terms of the perceived quality of students on the global university experience at the completion of the degree, the second deals with the analysis of the external effectiveness of the university in terms of the job satisfaction of students who graduated one year before.

In each section, after a brief description of available data, results of traditional analyses are shown.

In both case studies, we first investigate the item distributions and the correlation between the items relating to specific aspects of satisfaction (sections 2.2 and 2.3). This analysis describes the phenomenon, but it does not allow an evaluation of the effect of each single aspect on global satisfaction.

For the evaluation of the internal effectiveness of the university system we next apply a multilevel regression model to global satisfaction (section 2.2.1) using as covariates the students' responses to the items on specific aspects of satisfaction. The first level units are the students, the second level units are the programs that students attended.

2.1 Data and the Italian university reform

Data come from two surveys carried out by the consortium Almalaurea. AlmaLaurea was founded in 1994 by the Statistical Observatory of the University of Bologna and currently includes 51 Italian universities (out of 85). It is managed by a consortium of Italian universities with the support of the *Ministry of Education, Universities and Research* and its main purpose is to be a reference point within the university system for all the organizations involved in university education and employment of the young people.

Collecting information directly from the students through two kinds of survey, one aim of Almalaurea and the involved universities is to obtain information about the Italian university system. One survey typology collects information about graduates profile, the other asks for their employment status after some years (1, 3 or 5) after the degree.

Data used to analyse the internal effectiveness of the university in terms of the students' perceived quality on the university experience, come from the AlmaLaurea survey on students who graduated in 2004. Data used to analyse the external effectiveness of the university in terms of the graduates' perceived quality on their job, come from the AlmaLaurea survey on employment opportunities of students who graduated during the summer period 2004. In the two case studies, two different groups of individuals are analysed.

During the last years, the Italian university system was involved in the university reform operating since 2001. The reform was introduced in 1999 (Ministerial Decree no. 509/99), and it was enforced for the first time at the University of Florence during the academic year 2001/02. It introduced some important innovations in the organization of the academic degrees. It established the organization of studies in 3 cycles: the first cycle has a 3-year duration and leads to a Bachelor of Science equivalent degree (UK), the second cycle (Laurea Specialistica / Laurea Magistrale) has a 2-year duration and it leads to a Master of Science equivalent degree (UK), the third cycle (Dottorato) has a 3-year duration and it leads to a PhD equivalent. Some programs have a different duration (5 years) and, of course, different organization: *architettura e ingegneria edile*, *farmacia e farmacia industriale*, *medicina e chirurgia*; these programs are called LSCU (lauree specialistiche a ciclo unico).

The aim of the thesis is to provide some useful tools for the future local university policy so it would be proper to analyse data on students involved in the new university system. Unfortunately, this is not possibile in the second case study, indeed Almalaurea only interviews students that graduated during the summer session. First students enrolled with the new system could graduate in september 2004, so analysing students who graduated with the new university system during the summer session means analyse hybrid students that may have particular features. Usually, students that "change" system are involved in the educational process for many years and prefer to finish their studies quickly (for almost all study programs the new university system degree takes 3 years and the old university system degree takes 4 or 5 years) even if the new degree is less prestigious than the old degree and offers fewer employment opportunities.

2.2 University internal effectiveness

Data analysed in this section concern the profile of students who graduated from the University of Florence during the solar year 2004. Data are collected by the consortium AlmaLaurea submitting to students a questionnaire one month before the end of their studies. The questionnaires are filled in on Internet; in 2004 the response rate for the national survey was equal to 84.1%.

Students who graduated at the University of Florence during the year 2004 are 6966. As shown in table 2.1, the most of students (70.8%) graduated with the old university system, while only 25.8% of students took a Bachelor degree. Since one aim of the thesis is to supply some useful tools for the future policy of the university, only data relating to the 1800 students who graduated with the new system are analysed. In particular, the data about Master degrees are not included in the analysis, since their organization is quite different from the organization of the Bachelor degree. For the same reason, also the programs LSCU with a duration of 5 years instead of 3 years are excluded from the analysis (section 2.1).

Table 2.1: Students graduated at the University of Florence, year 2004.

	Frequency	%
Old system degree (4 or 5 years)	4930	70.8
New system degree (5 years)	179	2.6
New system degree - Bachelor Degree (3 years)	1800	25.8
New system degree - Master degree (2 years)	57	0.8
Total	6966	100

The questionnaire used for the survey about "graduates profile" concerns many aspects of students curricula:

- personal information,

- studies before university,
- information on the university experience,
- information on the student family,
- future perspective.

In the third section there are questions about the opinion of respondents on the university system. One aim of this dissertation is to provide information that can be useful for the internal policy of the degree programs: all items relating to services provided by external organizations, such as canteen, transports and sport services are not analysed.

The items that will be used in the analysis are:

- Are you satisfied about the relationship you had with professors?
- Are you satisfied about the relationship you had with your supervisor?
- Are you satisfied about the relationship you had with professors' assistants?
- Are you satisfied about the relationship you had with technical and administrative staff?
- Are you satisfied about the relationship you had with other students?
- What do you think about the lecture rooms?
- What do you think about the computers?
- What do you think about the laboratories and facilities for the didactic activities?
- What do you think about the libraries?
- What do you think about the rooms used for the individual study?
- Are you globally satisfied of your course of study?

On the 1800 students who graduated in 2004, 1474 (81.9%) filled in the questionnaires, and 1 student did not reply at any item considered in the analysis. The students attended 38 different study programs[1], as shown in Table 2.2 and 2.3.

Table 2.2: Students graduated (new system degree) at the University of Florence, year 2004: study programs attended.

FACULTY - Program	% on Faculty Total	% on Total	N resp.
AGRICULTURE	*100*	*4.7*	*69*
sc. tecn. agrarie	100	4.7	69
ARCHITECTURE	*100*	*2.0*	*29*
sc. architettura ing. edile	48.3	1.0	14
urbanistica	20.7	0.4	6
disegno industriale	31.0	0.6	9
ECONOMICS	*100*	*19.82*	*292*
sc. economia e gest. az.	76.4	15.1	223
sc. economiche	9.9	2.0	29
sc. sociali per cooperaz.	4.1	0.8	12
sc. statistiche	9.6	1.9	28
EDUCATION SCIENCE	*100*	*4.2*	*62*
sc. educ. e a formaz.	100	4.2	62
ENGINEERING	*100*	*11.2*	*165*
sc. architettura ing. edile	3.0	0.3	5
ing. civile e ambientale	20.6	2.3	34
ing. informazione	34.6	3.9	57
ing. industriale	41.8	4.7	69
INTER-FACULTY	*100*	*0.88*	*13*
biotecnologie	100	0.88	13
LAW	*100*	*4.28*	*63*
sc. giuridiche	100	4.28	63
LETTERS AND PHILOSOPHY	*100*	*17.7*	*261*
lettere	9.2	1.6	24
lingue e culture moderne	28.4	5.0	74
sc. beni culturali	18.0	3.2	47
sc. comunicazione	10.7	1.9	28
sc. e tecn. arti	18.8	3.3	49
filosofia	2.3	0.4	6
sc. geografiche	2.7	0.5	7
sc. storiche	10.0	1.8	26

Table 2.3: Students graduated (new system degree) at the University of Florence, year 2004: study programs attended (*continued*).

FACULTY - Program	% on Faculty Total	% on Total	N resp.
MEDICINE	*100*	*17.5*	*257*
sc. e attività motorie	9.3	1.6	24
prof. inferm. e ostetrica	60.3	10.5	155
prof. sanitarie a riab.	15.2	2.7	39
prof. sanitarie tecniche	15.2	2.7	39
PHARMACY	*100*	*1.0*	*14*
sc. e tecn. farm.	100	1.0	14
POLITICAL SCIENCE	*100*	*5.5*	*81*
sc. servizio sociale	24.7	1.4	20
sc. comunicazione	30.9	1.7	25
sc. pol. e rel. intern.	25.9	1.4	21
sc. amministrazione	12.4	0.7	10
sc. sociologiche	6.2	0.3	5
PSYCHOLOGY	*100*	*5.4*	*80*
sc. e tecn. psicologiche	100	5.43	80
SCIENCE (MPNS)	*100*	*5.9*	*87*
sc. biologiche	6.9	0.4	6
sc. e tecn. chimiche	20.7	1.2	18
sc. e tecn. fisiche	8.1	0.5	7
sc. e tecn. infor.	36.8	2.2	32
sc. matematiche	17.2	1.0	15
tecn. conservaz. beni cult.	10.3	0.6	9
TOTAL UNIVERSITY	100	100	1473

Satisfaction is expressed on an ordinal Likert scale with 4 categories, the only exception is the satisfaction for the *Computers* and *Individual spaces* that is measured by an ordinal Likert scale with 3 categories. The scale for the items relating to the satisfaction with the relationship with "persons" involved in the university system (such as professors, professor assistants, students, etc.) is composed by 4 categories: *definitively not satisfied*, *more not satisfied than satisfied*, *more satisfied than not satisfied*, *definitively satisfied*. Satisfaction with the *Lecture rooms* and the *Laboratories* is expressed choosing between the categories *never suitable*, *rarely suitable*, *often suitable*, *always suitable*, and the opinion on the *Library* is expressed on the scale: *definitively bad*, *quite bad*, *quite good*, *definitively good*. Satisfaction with the *Computers* and the *Individual spaces* is expressed choosing between the categories *not present*, *no suitable number*, *suitable number*[2].

Frequency distribution of almost all responses is strongly asymmetric (Tab. 2.4), so the two last categories are collapsed[3]: *definitively not satisfied* and *more not satisfied than satisfied* are collapsed in *not satisfied*, *never suitable* and *rarely suitable* in *not suitable*, and *definitively bad* and *quite bad* in *bad*.

The two items *Relationship with supervisor* and *Relationship with students* present a really high percentage of students that replied *definitively yes*; indeed, students usually choose their supervisor and, in some situations (for example when attending programs with a big number of students), also their "colleagues".

All items are expressed on an ordinal scale, so polychoric correlations are computed in order to study the association between the variables (Bollen, 1989). Polychoric correlations represent the correlation between the continuous variables $y_{hij}*$ underlying the observed indicators via a threshold model (section 1.2):

$$y = s \ if \ \alpha_{s-1} < y_{hij}* < \alpha_s \quad \alpha_0 = -\infty, \alpha_1 = 0, \alpha_s = \infty$$

where $s, s = 1, \ldots, S$ are the category of the ordinal response y, S is the

[1]Some programs belong to different Faculties, for example *sc. architettura ing. edile* belongs both to Architecture and Engegneering and *sc. comunicazione* belongs both to Letters and Phylosopy and Political Science. Since the focus of the thesis is on the programs attended by students and these have very similar charachteristics even if they are organized by different Faculties, they will be considered as one program.

[2]For these two items the N in table 2.4 represents the number of students that replied to the questionnaire less the number of nonrespondents for the specific item added to the number of students that never used that service; the category *never used* is not considered as a "measure" of the students' opinion.

[3]This is also useful in order to reduce the computational time for the estimation of a generalized model with latent variables (section 1.4.1).

Table 2.4: Item frequency distributions (percentages). Students graduated (new system degree) at the University of Florence, year 2004.

Item	Level of satisfaction				N resp.
	Definitively no	More no than yes	More yes than no	Definitively yes	
Rel. professors	1.7	15.0	65.6	17.8	1463
Rel. prof. assistants	3.7	17.0	58.3	21.1	1433
Rel. technical staff	13.1	30.3	43.0	13.6	1460
Rel. supervisor	2.5	6.0	26.0	65.5	1402
Rel. students	0.9	4.3	35.8	59.0	1464
Global satisfaction	1.8	13.5	57.2	27.5	1468
	Never suitable	Rarely suitable	Often suitable	Always suitable	
Lecture rooms	6.5	38.1	39.2	16.3	1451
Laboratories	10.1	43.9	35.6	10.4	1143
	Definitively bad	Quite bad	Quite good	Definitively good	
Library	5.0	13.1	61.5	20.4	1351
	Not present	No suitable number	Suitable number		
Computers	14.8	66.0	19.3		1360
Ind. spaces	14.6	55.3	30.1		1262

total number of categories for y and $\alpha_s (s = 1, 2, ..., S-1)$ are the category thresholds.

The polychoric correlations of the 11 items used in the analysis are shown in Table 2.5.

There are two groups of items where the correlations are quite high (higher than 0.40). The first group contains the first 4 items about the satisfaction of students with their relationship with the university "personnel". The second is composed by the last 4 items, with information about students' satisfaction with the university "physical services".

The two items *Relationship with supervisor* and *Relationship with students* have an anomalous behavior: they have a quite low correlation with all other items. Also this peculiarity, beyond the strongly asymmetric frequency distribution different from the other item distributions (Tab. 2.4) and the fact that the university cannot act directly on these two aspects, lead to not include these two variables in the subsequent analyses.

On the contrary, the item *Relationship with technical staff* has a strong correlation both with the items of the personnel group, and with the items *Lecture rooms*, *Laboratories* and *Library*. This is quite understandable, since the perceived quality on the physical services depends also on the quality of some secondary services offered by the university, such as the competence and organization of the technical personnel.

Global satisfaction is positively correlated with all other items; it has the highest correlation with the satisfaction with *Relationship with professors* and *Relationship with professors' assistants* and the lowest with *Computers* and *Individual spaces*.

In order to reach an improvement of students' satisfaction, the university should stress more on the characteristic of the personnel, for example paying more attention on the recruitment of the professors or giving them the opportunity to participate to "refresher" courses, and training periodically the technical staff. Then, the university should improve the quality of the laboratories and the lecture rooms.

This analysis describes the phenomenon of satisfaction, but it does not allow an evaluation of the effect of each single aspect on global satisfaction "controlling for the others"; as well known, the correlations take into account only the association between pair of variables. Furthermore, the analysis of polychoric correlations does not allow an evaluation of the university (programs) performance, since it does not consider that students attended different programs that can have a really different organization and features. For these reasons, a multilevel regression model is applied on global satisfaction. The first level units are the students, the second level units are the programs that student attended. Then, some covariates are added to the model.

Table 2.5: Item polychoric correlations. Students graduated (new system degree) at the University of Florence, year 2004.

Item	1	2	3	4	5	6	7	8	9	10	11
1. Global satisfaction	1										
2. Rel. prof. assistant	0.42	1									
3. Rel. professors	0.58	0.74	1								
4. Rel. technical staff	0.36	0.47	0.55	1							
5. Rel. supervisor	0.28	0.29	0.35	0.22	1						
6. Rel. students	0.32	0.37	0.33	0.16	0.15	1					
7. Lecture rooms	0.34	0.24	0.33	0.41	0.07	0.13	1				
8. Library	0.25	0.26	0.33	0.36	0.14	0.13	0.47	1			
9. Laboratories	0.39	0.35	0.38	0.43	0.09	0.16	0.72	0.41	1		
10. Computers	0.20	0.17	0.18	0.28	-0.02	0.09	0.56	0.42	0.50	1	
11. Ind. spaces	0.16	0.19	0.19	0.22	0.09	0.18	0.39	0.43	0.28	0.33	1

2.2.1 University internal effectiveness and the regression model

A multilevel regression model is applied on global satisfaction[4]. The response variable is expressed on an ordinal scale with 3 categories, and the response model is the cumulative ordinal model, also known as proportional odds model.

Using the traditional multilevel terminology, the null model is expressed by:

$$
\begin{aligned}
logit[P(y_{ij} \leq s)] &= \alpha_s + v_{ij} &\quad (2.1)\\
v_{ij} &= \mu_i + e_j &\quad (2.2)
\end{aligned}
$$

where i and j represent, respectively, first and second level units: students and the programs they attended. Then, $s = 1, \ldots, S$ represents the categories of the response variable y and α_s are the thresholds increasing in s.

The multilevel regression model is part of the generalized latent variable modeling framework (section 1.1). Using the terminology of the framework presented in Chapter 1, the student level would be indicated with the index h and the programs with i (Figure 2.1).

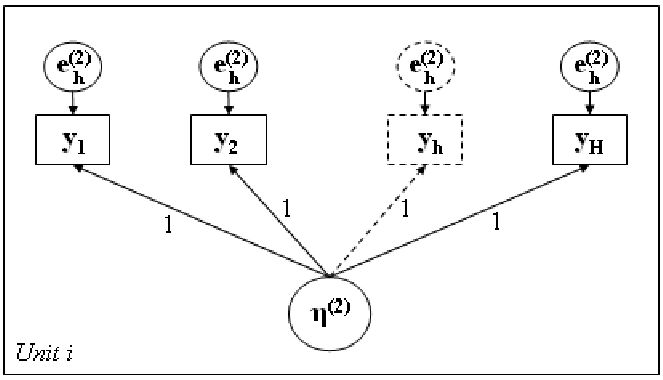

Figure 2.1: Two-level regression model.

The errors e_{ij}, implied by the binomial distribution in equation (2.1), represent "item" level errors $e_{hi}^{(2)}$ with variance assumed to be equal for each item ($\omega_{hh'}^{(2)} = \omega^{(2)}$), and e_j in equation (2.2) is a latent variable $\eta_i^{(2)}$, with variance $\psi^{(2)}$, representing the heterogeneity at the i-th level of the analysis. As usually assumed in multilevel models, the errors at different levels are

[4]The analysis is implemented with the software Latent GOLD, version 4.5 (Vermunt and Magidson, 2005*b*).

independent. Next, the multilevel symbology is used, so the variances of e_{ij} and e_j will be indicated, respectively, with σ^2 and τ_{00}.

For the identification of the model, μ_i is constrained to 0, so there are $S-1$ (two in the application) thresholds. Since the link logit is used, σ is equal to $\pi/\sqrt{3}$.

The variance at the program level is equal to 0.21. To test the null hypothesis that the variance is zero, the Deviance test is used; in particular, since the alternative hypothesis is one-sided (positive variance), the p-value is halved (Snijders and Bosker, 1999). The variance parameter is statistically significant[5]. The intraclass correlation coefficient is $ICC = (\tau_{00}/[\tau_{00}+\sigma^2] = 0.06)$, about 7% of the total variance of global satisfaction is due to differences between clusters (study programs).

As a second step of the analysis, students' satisfaction with specific aspects of the university system are introduced as covariates in the model. Note that covariates in a regression equation are typically assumed to be perfect measures of specific characteristics of individuals, *sex*, *age*, *residence*. Since this assumption is unrealistic for the responses on the satisfaction items, the estimated regression weights will be somewhat attenuated, that is, somewhat lower that would have been the case with a perfect measurement. The reference category of each covariate is *not satisfied*, so the "baseline student" is a student with a negative approach to the university.

The model is now expressed by:

$$\begin{aligned} logit[P(y_{ij} \leq s|\boldsymbol{X})] &= \alpha_c + v_{ij} \\ v_{ij} &= \boldsymbol{x}_{ij}\boldsymbol{\beta} + e_j \end{aligned}$$

Table 2.6 reports the estimates of the fixed parameters of the model.

The estimates with sign plus indicate that the logit of the cumulative probability $(P(y_{ij} \leq s))$ is increasing, so there is an increment of the probability of being less satisfied. All the estimates that are statistically significant have a positive effect (sign minus) on global satisfaction: the more satisfied a student is on specifical aspects of the university system, the more globally satisfied he is.

The Wald Tests (Vermunt and Magidson, 2005*c*) in Table 2.6 indicate that the satisfaction with *Computers*, *Library* and *Individual spaces* do not have a globally significant effect on global satisfaction. This confirms the low polychoric correlations with global satisfaction (section 2.2).

Again, the covariate with the highest effect on global satisfaction is *Relationship with professors*, so the university should stress more on the charac-

[5]The deviance of the multilevel null model is equal to 2795.818 and the deviance of the ordinary regression model is 2823.799.

Table 2.6: Multilevel logistic model: fixed parameters. Students graduated (new system degree) at the University of Florence, year 2004.

Variable	Estimate	Std.Error	Wald test p value	Odds
Global satisfaction (1)	0.44	0.28	0.00	1.55
Global satisfaction (2)	3.90	0.31		49.39
Rel. professors (2)	-1.80	0.18	0.00	0.17
Rel. professors (3)	-2.78	0.26		0.06
Rel. prof. assistant (2)	-0.10	0.16	0.04	0.90
Rel. prof. assistant (3)	-0.50	0.22		0.60
Rel. technical staff (2)	0.06	0.13	0.00	1.06
Rel. technical staff (3)	-0.62	0.21		0.54
Lecture rooms (2)	-0.32	0.14	0.01	0.72
Lecture rooms (3)	-0.61	0.21		0.54
Computers (2)	-0.40	0.18	0.05	0.67
Computers (3)	-0.22	0.23		0.80
Laboratories (2)	-0.31	0.15	0.04	0.73
Laboratories (3)	-0.51	0.24		0.60
Library (2)	0.00	0.17	0.27	1.00
Library (3)	-0.25	0.22		0.78
Ind. spaces (2)	0.13	0.18	0.28	1.14
Ind. spaces (3)	-0.09	0.20		0.92

teristics of the professors in order to have more satisfied students. Furthermore, the university should improve the quality of the lecture rooms and the laboratories.

The variance at the program level, which equals 0.16, is statistically significant. The intraclass correlation coefficient is = 0.045, which is lower than the ICC in the model without covariates: the reduction of the differences between programs is caused by the fact that their (student) composition is controlled for.

The regression model gives some indications about the effects of the perceived quality on the specific aspects of the university on global satisfaction. It also let to quantify the "program effect" and so the homogeneity between students attending the same program.

However, the regression model is not the most suitable statistical model to study the phenomenon of satisfaction. As mentioned previously, using the satisfaction item responses as either independent or dependent variables means to assume, unrealistically, that they are perfect measures of specific characteristics of individuals. Furthermore, using the satisfaction item responses as independent variables means that their effects may be attenuated.

Considering the nature of the phenomenon of satisfaction and the data of AlmaLaurea survey, the most suitable statistical methodology for the analysis of global satisfaction is the factor model that allows modeling the relationship between the latent constructs and the observed indicators (Spearman, 1904). Furthermore, data have a hierarchical structure, so multilevel techniques are necessary in order to correctly interpret the phenomenon.

2.3 University external effectiveness

Every year AlmaLaurea interviews students who graduated from the universities participating to the consortium about their employment opportunities. The interviews concern 3 different groups of students who graduated, respectively, 1, 3 and 5 years before. Data are collected through Computer Assisted Telephone Interviewing (C.A.T.I.) and they refer to students who graduated during the summer period.

The thesis focuses only on the analysis of the University of Florence with the ultimate aim of providing policy advice for the local university policy. In this section data collected during the survey in 2005 are analysed[6]. The focus is on students who graduated in 2004: of course this study represents only a first step in evaluating the evolution of job satisfaction over the time.

[6]The national response rate is equal to 86% for students who graduated in 2004, 81% for students who graduated in 2002 and 76% for students who graduated in 2000.

Data concern the employment status of students who graduated in 2004 at the University of Florence with the old university system. Data on LSCU programs are not used since they have particular features (section 2.1)[7].

As mentioned previously, as an indicator of the university external effectiveness we analyse the job satisfaction (perceived quality) of graduates, so we exclude from the analysis the graduates who are unemployed at the time of the interview. On the 1369 students who graduated from the University of Florence in 2004 that replied to the question about their employment status[8], the 61.1% (837 units) works[9] at the moment of the interview, the 15.2% (208 units) worked after the degree but do not work at the moment of the interview and the 23.7% (324 units) never worked.

The 837 graduates working at the interview attended 42 different programs; because of the low number of students some programs are grouped[10]. The final number of distinct programs is 23, with 826 students[11] (Table 2.7).

The questionnaire used for the survey about the "employment status of graduates" is organized in 4 main sections. After some questions about post degree education (Ph.D., Master, etc.), the other questions depend on the respondent employment status.

Section A contains questions for graduates working at the moment of the interview; the questions are relative to job features: time to get the job, ways of searching the job, employment contract, sector (public, private, area, etc.), job satisfaction, earnings, etc. In section B there are questions directed to graduates working at the degree but not working at the interview. The questions are about: the ways of searching the job, the time spent to get the first job and the reasons of interrupting the last job. Section C is dedicated to graduates that never worked after the degree; it contains questions about the possibility of searching a new job (reasons, feature of the desired job, etc.). Finally, the questions in section D, directed to all respondents, concern the possibility of re-enrolling at the university and the opinion about the university reform. Last section is dedicated to information

[7]Students who graduated in *architettura e ingegneria edile* are 2, students who graduated in *farmacia e farmacia industriale* are 10 and students who graduated in *medicina e chirurgia* are 56.

[8]The total number of students that graduated at the University of Florence in 2004 during the summer session is 1596.

[9]In the survey, training courses, PhD courses, Masters are not considered as working activities. On the contrary, non permanent jobs are included in the working activities.

[10]Two or more programs are aggregated if they (*i*) have less than 8 students, (*ii*) are offered by the same Faculty and belong to the same group of study.

[11]Programs *chimica* (3 students), *farmacia* (3 students), *medicina e chirurgia* (2 students), *odontoiatria e protesi dentaria* (3 students) are excluded because they cannot be aggregated.

Table 2.7: Students graduated (old system degree) at the University of Florence, summer session year 2004: study programs attended.

FACULTY - Program	% on Faculty Total	% on Total	N resp.
AGRICULTURE	*100*	*2.3*	*19*
scienze forestali ed ambientali	52.6	1.2	10
altro agraria	47.4	1.1	9
ARCHITECTURE	*100*	*18.8*	*155*
architettura	100	18.8	155
EDUCATION SCIENCE	*100*	*12.3*	*102*
scienze dell educazione	73.5	9.1	75
scienze della formazione primaria	18.6	2.3	19
altro scienze della formazione	7.8	1.0	8
ECONOMICS	*100*	*9.9*	*82*
scienze statistiche	9.8	1.0	8
economia aziendale	48.8	4.8	40
economia e commercio	41.5	4.1	34
ENGENEERING	*100*	*10.8*	*89*
ingegneria civile	12.4	1.3	11
ingegneria elettronica	21.3	2.3	19
ingegneria meccanica	29.2	3.1	26
ingegneria informatica	11.2	1.2	10
ingegneria per ambiente e territ.	12.4	1.3	11
altro ingegneria	13.5	1.5	12
LAW	*100*	*4.5*	*37*
giurisprudenza	100	4.5	37
LETTERS	*100*	*13.9*	*115*
filosofia	12.2	1.7	14
lettere	54.8	7.6	63
lingue e letterature straniere	33.0	4.6	38
POLITICAL SCIENCE	*100*	*8*	*66*
scienze politiche	100	8.0	66
PSYCHOLOGY	*100*	*15.9*	*131*
psicologia	100	15.9	131
SCIENCE (MPNS)	*100*	*3.6*	*30*
gruppo geo-biologico	56.7	2.1	17
gruppo scientifico	43.3	1.6	13
TOTAL UNIVERSITY	100	100	826

about respondents, such as family, children, etc.

The variables that will be analysed are included in section A of the questionnaire and are relative to the graduates' opinion on some features of their job:

- steadiness, stability
- coherence with studies
- competence-professionalism
- prestige
- cultural interests
- social utility
- independence
- involvement in the working activity and in the decisional processes
- schedule flexibility
- free time
- job place
- relationship with colleagues
- salary
- carrier.

Other items could be used as indicators of job satisfaction (for example the amount of earnings, the search of a new job, etc.), but the chosen items are a direct measure of some aspects of job satisfaction and are expressed on the same scale, making the model better interpretable. The direct question on global satisfaction is also used.

All items are expressed on an ordinal scale with 10 categories. The response frequency distributions are quite similar for all items (Figures 2.2 and 2.3). The most frequent scores are usually 7 and 8 and the score 1 has an higher percentage than the score 2; this feature is quite understandable: score 1 represents the highest grade of disagreement. The distribution of the two items *Coherence with studies* and job *Steadiness* is quite different from the others. These two items and the item *Carrier* have also the highest percentage of graduates that are completely not satisfied: the 11.5%, 7.8%

and 7.5% of graduated replied 1 respectively for the *Coherence with studies*, *Carrier* and *Steadiness* of the job. The graduates' responses to these three items reveal some critical features of the university system. Indeed, one year after the degree a high percentage of graduates consider their job not stable and with low carrier opportunities; furthermore, the job is not coherent with their studies. This may be naturally linked to the necessity of some graduates to continue their educational and training programs in order to do the job they studied for, and the connected necessity to do occasional jobs. Generally, this indicates that the university is not able to reply to the market requests, at least in the short period.

In subsequent analyses, the items are considered as continuous variables because of the number of the categories.

The highest correlation (Table 2.8) is between the items *Salary* and *Carrier* (0.80), followed by the correlation between *Carrier* and *Prestige* (0,64), *Prestige* and *Professionalism* (0.64) and by the correlation between *Independence* and *Involvement* in working activity and in decisional processes (0.63), and the correlation between *Coherence* with studies and *Cultural interests* (0.63). The item *Prestige* is correlated with almost all other items. The highest correlations with the *Global satisfaction* are for the items *Coherence with studies*, *Professionalism*, *Prestige*, *Cultural Interests*, *Involvement* in working activity and in decisional processes, *Salary* and *Carrier*. On the contrary, the item *Free time* is not correlated with any other item except with *Schedule flexibility*: the graduates' opinion on the free time is not associated with the other aspect of job satisfaction. The item *Free time* is not included in the subsequent analyses.

The correlations show a complex structure of the observed association between items; indeed, the satisfaction is a complex process that is naturally considered a latent construct, not directly observable.

As for the students' satisfaction about the university system, considering the nature of the phenomenon and the features of the data, the most suitable statistical methodology for the the analysis of job satisfaction is the factor model. Data have a hierarchical structure (first level units are the graduates and second level units are the programs that they attended) and some preliminary analyses show that the hierarchy is relevant in terms of the phenomenon analysed; multilevel tools of analysis will be used.

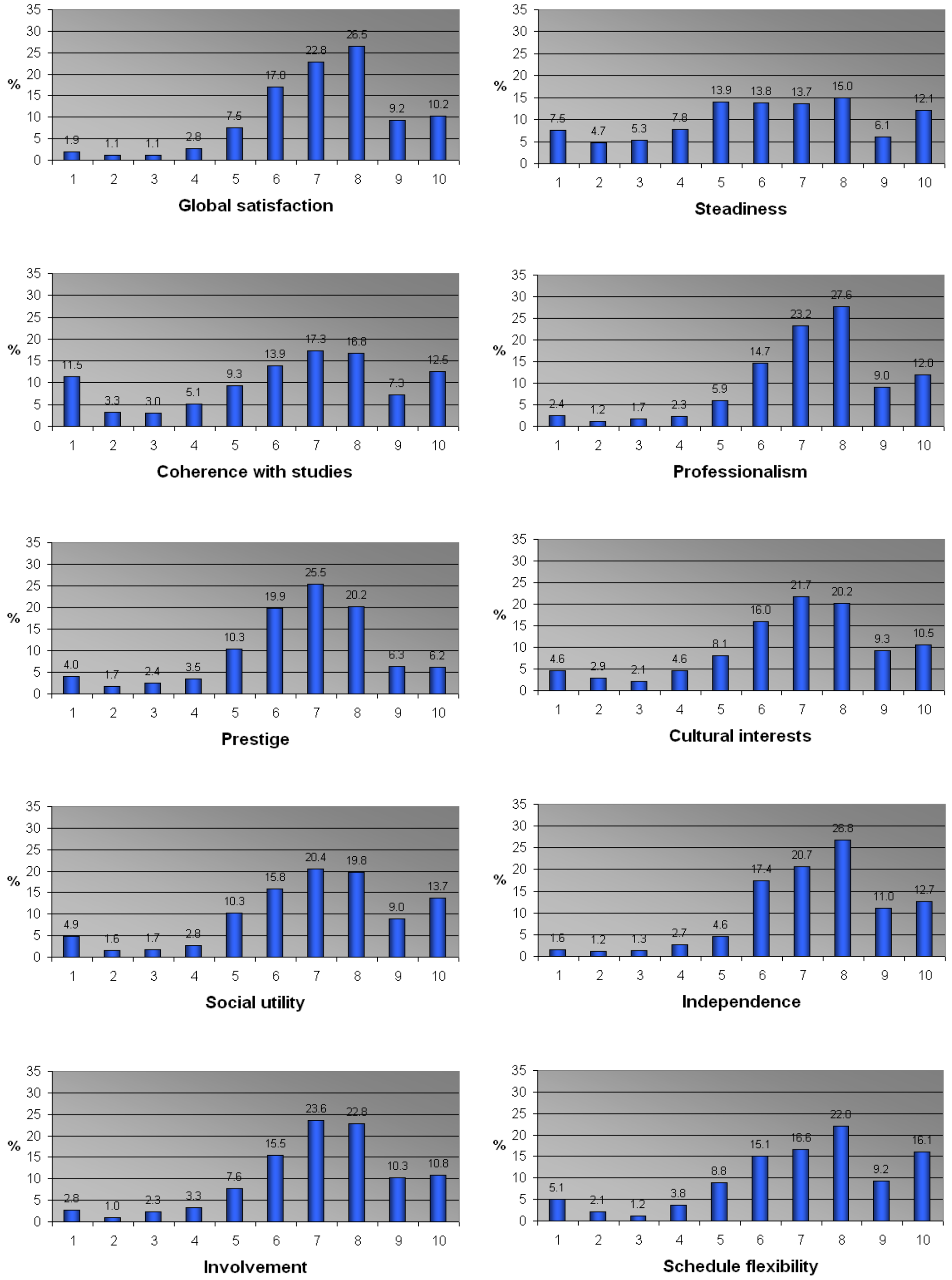

Figure 2.2: Frequency distribution of students' responses. Students graduated (old system degree) at the University of Florence.

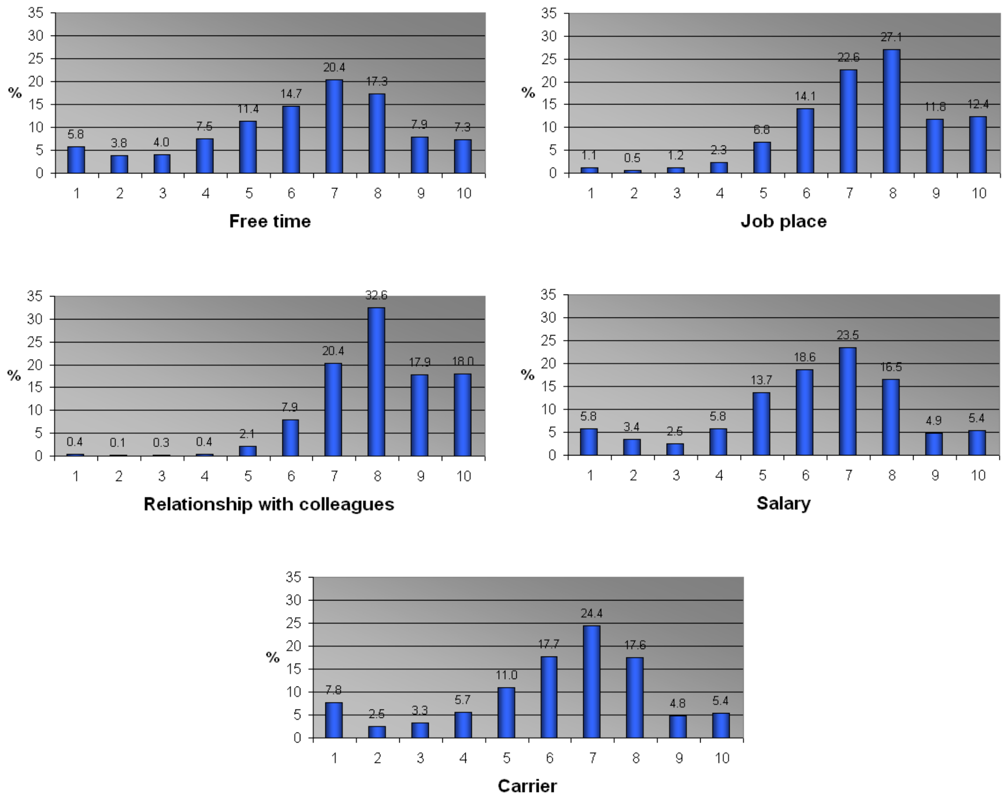

Figure 2.3: Frequency distribution of students' responses (*continued*). Students graduated (old system degree) at the University of Florence.

Table 2.8: Item correlations. Students graduated (old system degree) at the University of Florence, summer session year 2004.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Global satisf.	1														
2. Steadiness	0.27	1													
3. Coher. studies	0.51	0.10	1												
4. Professionalism	0.62	0.20	0.53	1											
5. Prestige	0.67	0.28	0.46	0.64	1										
6. Cultural int.	0.66	0.09	0.63	0.57	0.58	1									
7. Social utility	0.41	0.08	0.37	0.36	0.34	0.50	1								
8. Independence	0.46	0.20	0.27	0.39	0.45	0.41	0.34	1							
9. Involvement	0.54	0.22	0.35	0.49	0.54	0.49	0.39	0.63	1						
10. Sched. flexib.	0.29	0.04	0.17	0.26	0.33	0.26	0.19	0.44	0.39	1					
11. Free time	0.02	0.10	-0.05	-0.02	0.00	0.01	0.14	0.12	0.01	0.28	1				
12. Job place	0.25	0.16	0.09	0.17	0.23	0.15	0.10	0.24	0.22	0.22	0.17	1			
13. Rel. colleagues	0.29	0.12	0.12	0.24	0.27	0.17	0.13	0.21	0.21	0.17	0.07	0.31	1		
14. Salary	0.49	0.37	0.31	0.40	0.57	0.35	0.16	0.34	0.40	0.25	-0.02	0.15	0.16	1	
15. Carrier	0.54	0.33	0.43	0.52	0.64	0.49	0.27	0.36	0.44	0.27	-0.04	0.13	0.16	0.80	1

Chapter 3

Multilevel mixture models and the evaluation of university performance

In this Chapter we show the results of the multilevel mixture factor models used to evaluate the effectiveness of the university system from the users' point of view, both at the time of completion of the degree and one year later.

Data used for the analyses are described in Chapter 2. We focus on the university of Florence because of our knowledge of the context.

The Chapter is divided in two sections: the first deals with the analysis of the university internal effectiveness, the second deals with the analysis of the university external effectiveness.

In both analyses we use models of the generalized latent variable modeling framework (Skrondal and Rabe-Hesketh (2004), Vermunt (2007), Muthén and Muthén (1998-2007)), described in Chapter 1. Since data have a hierarchical structure, we use different specifications of the *multilevel mixture factor model*.

To evaluate the university internal effectiveness we analyse the students' global satisfaction with the university experience. We first implement a *multilevel factor model* (section 3.1.1): using continuous latent variables at both levels of the analysis we study the latent constructs underlying the phenomenon of students' satisfaction both at the student and program level, highlighting the differences between the two structures. In particular, we investigate the differences between programs in the students' satisfaction, ranking the programs along a continuum. Next, we apply to the same data a *multilevel mixture factor model* with continuous latent variables at the student level and a categorical latent variable at the program level (section 3.1.2); we investigate the differences and similarities between study programs in the students' satisfaction classifying programs in different groups of satisfaction.

At the end of the two analyses, we merge the results relative to the program level (section 3.1.3).

As indicator of the university external effectiveness we analyse the job satisfaction (perceived quality) of graduated students. In the analysis of job satisfaction we implement a *multilevel mixture factor model*, with continuous latent variables at the individual level and a categorical variable at program level (section 3.2.1). The aim is to see if the programs (or groups of programs) differ in the mean values of the latent variables at the individual level representing the job satisfaction. In particular, with continuous latent variables at the individual level we reduce the dimensionality of the phenomenon and with a categorical variable at program level we classify programs relatively to the obtained latent dimensions.

In the analyses we focus on the interpretational features of the models, showing the extreme flexibility of the generalized latent variable modeling framework. The focus is on the second level of the hierarchy; when we speak about university effectiveness we refer to the effectiveness of each study program, since each study program has its own features and organization. Furthermore, although possible within the modeling framework, we do not use covariates in our models since our main aim is to measure the student's satisfaction as it is experienced in the real world. Indeed, it is difficult for university to act in different ways depending on students' characteristics (covariates at first level) or study programs' characteristics (covariates at second level).

3.1 University internal effectiveness

In this section we show the results of the analysis of the internal effectiveness of the university, evaluated from the students' point of view.

Data come from the AlmaLaurea survey on students who graduated in 2004. We focus on 1473 students who graduated (Bachelor degree) at the University of Florence under the new Italian university system and replied to the interview.

Data are collected by the consortium AlmaLaurea submitting to students a questionnaire one month before the end of their studies. After some preliminary analyses presented in sections 2.2 and 2.2.1, the items that are included in the models are nine and are relative to: relationship with professors, relationship with professors' assistants, relationship with technical and administrative staff, opinion on the lecture rooms, opinion on the computers, opinion on the laboratories and facilities for the didactic activities, opinion on the libraries, opinion on the rooms used for the individual study, global satisfaction.

3.1.1 Two-level factor model

In this section the multilevel factor model for the analysis of data concerning students' satisfaction is illustrated. The aim of the multilevel factor model is to understand which are the factors underlying the phenomenon at both level of the analysis and to describe the relationships between them, and to analyse the differences between study programs depending on the level of students' satisfaction.

The strategy described by Grilli and Rampichini (2007a), adapted from Muthén (1994), is used to implement the multilevel factor model.

The strategy consists of 4 steps:

1. univariate two-level models,
2. exploratory non-hierarchical factor analysis,
3. exploratory between and within factor analyses,
4. confirmatory two-level factor analysis.

The first step consists in the implementation of an univariate two-level model for each indicator h; since the indicators are all expressed on an ordinal scale, the response model is the cumulative ordinal model (section 2.2.1). In the application there are 3 categories per each item and two thresholds are estimated. The errors at program and student level are assumed to be independent; since the link used is logit, the variance of the errors at student level is equal to $(\pi/\sqrt{3})^2$.

The intraclass correlation coefficients ICC_h are all quite high and the likelihood ratio tests (LRT) comparing models with and without a multilevel structure (section 1.4.2) are all significant (Tab. 3.1): the proportion of variance of the responses due to differences between clusters (programs) is quite high and this confirms that a two-level analysis is worthwhile. The highest ICC_h values are for the items relating to the *Lecture rooms*, *Computers* and *Laboratories*, but also the ICC_h of the item *Relationship with technical staff* is quite high. While students' opinions on human aspects offset in different programs, they are quite different in mean level when technical features are concerned.

The second step of the analysis consists in the implementation of the non hierarchical exploratory factor analysis (EFA).

Table 3.1: Univariate ordinal logit random intercept models. Students graduated (new system degree) at the University of Florence, year 2004.

Item	ICC	Thresholds		LRT	
	(%)	α_1	α_2	Statistic	p-value
Rel. professors	6.5	-1.82	1.44	27.8	< 0.0001
Rel. prof. assistants	6.0	-1.54	1.25	38.5	< 0.0001
Rel. technical staff	11.0	-0.36	1.90	84.6	< 0.0001
Lecture rooms	23.3	-0.54	1.62	209.2	< 0.0001
Laboratories	24.1	-0.10	2.14	107.5	< 0.0001
Library	10.4	-1.86	1.29	103.1	< 0.0001
Computers	23.9	-2.29	1.43	177.1	< 0.0001
Ind. spaces	6.4	-1.97	0.78	117.9	< 0.0001
Global satisfaction	6.3	-1.86	0.92	28.2	< 0.0001

The debate between exploratory and confirmatory factor analysis is still open and some authors explicitly critic the use of exploratory factor analysis (Kline, 2005) as a way to get information on the model and, specifically, to choose the loadings that are tested to be exactly zero by applying a confirmatory analysis (section 1.3.1). In this thesis the exploratory factor analysis is used preliminary to other analyses because of the complexity of the models. The aim is to explore the presence of some underlying latent factors that can explain the associations between the observed variables and define their meaning. The analysis of the polychoric correlation briefly described in section 2.2 shows that there are two groups of items with strong associations; so, it is possible to assume the presence of two underlying factors. With an exploratory factor analysis the data examination is deepened.

Table 3.2 reports the results of the fit statistics of the EFA[1]. Since the Chi-Square test is sensitive to sample size (here 1473 units) and non-normality in the input variables (Kline, 2005), other two descriptive fit statistics are used (Muthén and Muthén, 1998-2007): the Root Mean Square Error of Approximation (RMSEA) and the Root Mean Square Residual (RMSR). According to Hu and Bentler (1999), RMSEA values below 0.06 indicate satisfactory model fit and RMR values should be below 0.08, with lower values indicating better model fit (Muthén and Muthén, 1998-2007). Given the fit statistics and the other descriptive analyses, the two-factor solution is chosen.

[1]The EFA has be implemented with the software Mplus, version 4; with ordinal indicators the default estimator is a kind of weighted least square estimator called WLSMV (Muthén and Muthén, 1998-2007).

Table 3.2: EFA: fit statistics. Students graduated (new system degree) at the University of Florence, year 2004.

	2 Factors	3 Factors
Chi-square value	99.267	16.105
Degrees of freedom	17	11
Probability value	0.000	0.137
RMSEA	0.057	0.018
RMSR	0.038	0.013

In Table 3.3 the factor loadings obtained with the promax rotated solution are shown. The correlation between the two factors is equal to 0.521. In Figure 3.1 the correlations between the items and the factors are plotted. Both factors are correlated quite strongly with global satisfaction. Then, there are two groups of variables. The first is composed by the items relating to the satisfaction of students on their relationship with the university personnel. The second is composed by the items with information about students' satisfaction on the university physical services. Again, the item *Relationship with technical staff* has a quite strong correlation with both factors.

The exploratory factor analysis indicates the presence of two underlying factors: one factor represents the satisfaction related to the *Human Environment*, the other the satisfaction with the *Physical Environment*.

Table 3.3: EFA: promax rotated factor loadings. Students graduated (new system degree) at the University of Florence, year 2004.

	Factor 1	Factor 2
Global satisfaction	0.51	0.14
Rel. prof. assistants	0.77	-0.04
Rel. professors	1.03	-0.10
Rel. technical staff	0.46	0.27
Lecture rooms	-0.07	0.90
Laboratories	0.08	0.75
Library	0.08	0.56
Computers	-0.14	0.72
Ind. spaces	-0.02	0.49

The third step of the analysis consists in the decomposition of the covariance matrix in Between and Within components, in order to carry out separated factor analyses. Since the decomposition is computationally not feasible when the items are ordinal, an approximate solution is used, as sug-

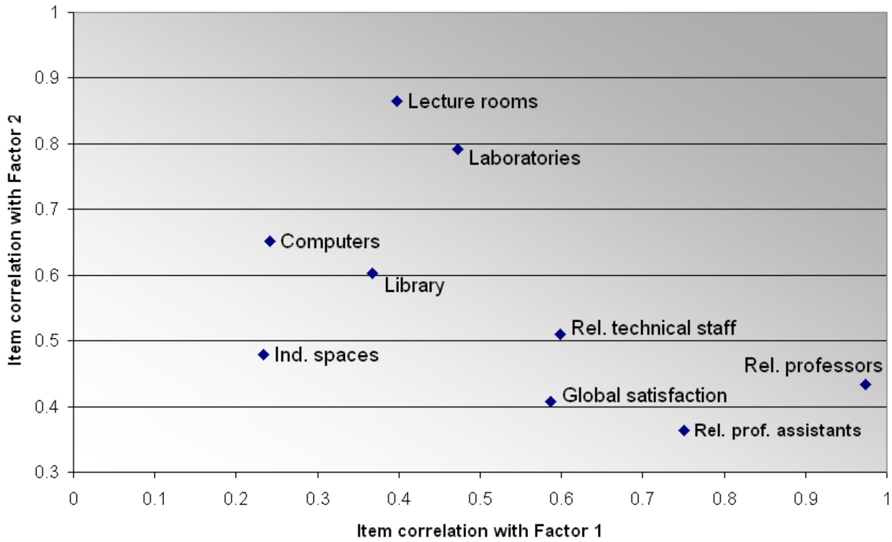

Figure 3.1: EFA: promax rotated solution, correlations between items and factors. Students graduated (new system degree) at the University of Florence, year 2004.

gested by Grilli and Rampichini (2007*a*): scores 1, 2 and 3 are assigned to the item categories and a multivariate two-level model for continuous responses is fitted[2]; results are reported in Table 3.4.

Table 3.4: Two-level multivariate model on item scores: between percentage of variance (**bold**) and covariance, and total variance of items (*italics*). Students graduated (new system degree) at the University of Florence, year 2004.

Item	1	2	3	4	5	6	7	8	9
Global satisfaction	**0.06**								
Rel. prof. collaborators	0.07	**0.06**							
Rel. professors	0.08	0.08	**0.05**						
Rel. technical staff	0.12	0.13	0.09	**0.10**					
Lecture rooms	0.20	0.17	0.09	0.24	**0.23**				
Library	0.08	0.14	0.1	0.07	0.20	**0.09**			
Laboratories	0.23	0.22	0.19	0.26	0.35	0.26	**0.22**		
Computers	0.32	0.21	0.2	0.34	0.38	0.23	0.33	**0.18**	
Ind. spaces	0.06	0.11	0.09	0.09	0.16	0.16	0.17	0.26	**0.04**
TOTAL VARIANCE	*0.41*	*0.42*	*0.35*	*0.48*	*0.53*	*0.38*	*0.45*	*0.34*	*0.42*

Confirming the previous analysis (Table 3.1), items with the highest per-

[2]The analysis is implemented using Mplus, version 4.2.

centage of the between component on the total variance are related to the structural features of the programs: *Lecture rooms*, *Laboratories* and *Computers*; the item referred to the *Library* has a lower between percentage, probably because many programs share the same library. The item *Relationship with technical staff* has the highest percentage of the between component compared to the other three items linked to the human relationships. Furthermore, the between percentages tend to be higher for covariances than for variances.

After decomposing the covariance matrix in within and between components, two exploratory analyses are carried out.

For the within component, the two-factor solution is chosen. Table 3.5 shows the factor loadings obtained with the promax rotated solution; the correlation between the two factors is equal to 0.542.

Table 3.5: EFA, within component: promax rotated factor loadings. Students graduated (new system degree) at the University of Florence, year 2004.

	Factor 1	Factor 2
Global satisfaction	0.43	0.15
Rel. prof. collaborators	0.69	-0.03
Rel. professors	0.93	-0.09
Rel. technical staff	0.37	0.27
Lecture rooms	-0.07	0.78
Library	0.07	0.46
Laboratories	0.02	0.64
Computers	-0.11	0.55
Ind. spaces	-0.02	0.41

More difficult is the interpretation of the EFA on the between component: the fit statistics Root Mean Square Residual is not proper neither for the 2 and 3 factors solution, being equal, respectively, to 0.465 and 0.110.

The univariate multilevel analysis (see the beginning of this section) shows that the between component is more “important” for the items linked with the university physical services. An hypothesis is that at the between level the human aspects are not measuring any underlying factor and the EFA using all items does not give any useful (interpretable) solution.

As the last step, a two-level confirmatory factor analysis is implemented in order to confirm the presence of the hypothetical constructs individuated with both the preliminary standard analysis, the exploratory factor analysis and the substantive knowledge of the phenomenon.

To select the final model the criteria illustrated in section 1.4.2 are used. When testing a factor loading to be equal to 0, the likelihood ratio test is used, while the BIC is used to compare models with different number of factors. Furthermore, the bivariate residual statistics (BVR) give indirect information on the global fit of the model.

The two-level factor model is expressed by (section 1.3.1):

$$v_{hij} \quad = \quad \mu_h + \sum_{m=1}^{M_2} \lambda^{(2)}_{mh}\eta^{(2)}_{mij} + \sum_{m=1}^{M_3} \lambda^{(3)}_{mh}\eta^{(3)}_{mj} + e^{(3)}_{hj}. \tag{3.1}$$

Since the indicators are all expressed on an ordinal scale, the conditional expectation of the response y_{hij} given the latent variables at different levels of the analysis is modeled with the cumulative ordinal model:

$$logit[P(y_{hij} \leq s|\boldsymbol{\eta^{(2)}}, \boldsymbol{\eta^{(3)}})] \quad = \quad \alpha_s + v_{hij}$$

where s represents the categories of the response variable y and α_s are the thresholds increasing in s.

In the questionnaire there is also a direct question on global satisfaction; this is used in the model in order to estimate the effect of each underlying factor on global satisfaction.

Some authors suggest to use the answer to the "global" question as a "*the gold standard with which scores derived from the remaining items are validated*" (Skrondal and Rabe-Hesketh, 2004, Ch. 7). Due to the aim of the thesis, this approach is not used.

In the model the global satisfaction is the reference item[3] ($h = 1$), i.e. $\lambda^{(l)}_{m1} = 1$, $m = 1, \ldots, M_l$, $l = 2, 3$. Therefore, the effect of each factor on the students' satisfaction is "measured" through its variance. Indeed:

$$\boldsymbol{\Sigma}(\boldsymbol{y}) = \boldsymbol{\Lambda\Psi\Lambda} + \boldsymbol{\Omega}$$

For the identification of the model the intercepts μ_h are constrained to 0.

Furthermore, at the student level, the loadings resulted to be close to zero with the EFA are constrained to 0, at the program level the model assumes one latent factor. The cluster-level item-specific errors $e^{(3)}_{hj}$ in equation (3.1) are constrained to zero (Grilli and Rampichini, 2007*a*) to reduce the computational burden.

[3]As shown very clearly by Millsap (2001) in one-level context, the choice of uniqueness constraints in confirmatory factor analysis is not trivial and different sets of uniqueness constraints may lead to different fit results when applied to the same data. In the model for the study of global satisfaction $\lambda^{(l)}$ relative to the global satisfaction item is constrained to 1 for interpretational reasons.

The multilevel factor model is thus:

$$v_{hij} \;=\; \sum_{m=1}^{M_2} \lambda_{mh}^{(2)} \eta_{mij}^{(2)} + \sum_{m=1}^{M_3} \lambda_{mh}^{(3)} \eta_{mj}^{(3)}$$

The final model has different structures at within and between level[4]. At the individual level there are two factors: one factor represents the satisfaction related to *Human Environment*, the other the satisfaction with *Physical Environment.* At the program level the structure with one factor is the most appropriate. The obtained model is illustrated in Figure 3.2.

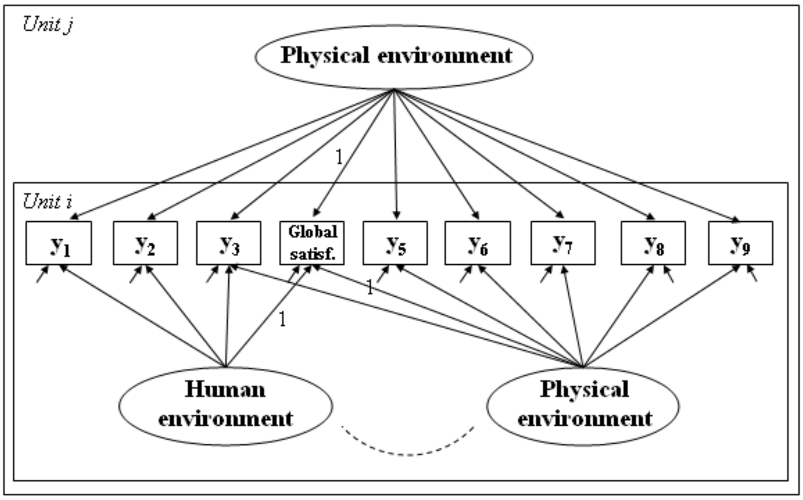

Figure 3.2: Two-level factor model. Students graduated (new system degree) at the University of Florence, year 2004.

Loglikelihood is equal to -10195.14, BIC^5 is 20532.15, the number of parameters is 39. The bivariate residual statistics of the model are reported in Table 3.6: each value is less than 3.84 (χ^2 with one degree of freedom, and significance level equal to 0.05), so the local independence assumption seems to hold and the model fit is considered good.

Model parameters and communalities are shown, respectively, in Table 3.7 and Table 3.8.

At the second level, only the loadings related to the items measuring the satisfaction with *Physical Environment* (*Lecture rooms*, *Library*, *Laboratories*, *Computers*, *Ind. spaces*) and the item *Rel. technical staff* are

[4]The model has been estimated with the software Latent GOLD. The software approximates the conditional density by means of Gauss-Hermite numerical integration, in this case study 10 quadrature nodes are used. The number of starting sets is 50 and 50 iterations are performed per set.

[5]BIC is calculated with $N = 38$ (section 1.4.2).

Table 3.6: Two-level factor model: bivariate residual statistics. Students graduated (new system degree) at the University of Florence, year 2004.

Item	1	2	3	4	5	6	7	8	9
1. Rel. prof. assistant	.								
2. Rel. professors	0.131	.							
3. Rel. technical staff	0.118	0.121	.						
4. Lecture rooms	0.120	0.010	0.283	.					
5. Library	0.319	0.069	0.329	0.095	.				
6. Laboratories	0.324	0.023	0.815	0.214	0.471				
7. Computers	0.013	0.021	0.095	0.051	0.265	0.027	.		
8. Ind. spaces	0.212	0.038	0.106	0.096	0.876	2.083	0.445	.	
9. Global satisfaction	0.307	0.144	0.187	0.067	0.106	0.794	0.433	0.306	.

significant. So, the programs differ only in the satisfaction with the *Physical Environment*.

Table 3.7: Two-level factor model: parameter estimates. Students graduated (new system degree) at the University of Florence, year 2004.

Item	Loadings			Thresholds	
	Within		Between		
	Human environment	Physical environment	Physical environment	α_1	α_2
Rel. prof. assistants	1.94		0.37[a]	-2.21	2.21
Rel. professors	6.18		0.72[a]	-5.95	5.86
Rel. techn. staff	0.87	1.60	1.90	-0.50	2.35
Lecture rooms		5.74	10.21	-1.13	2.69
Library		2.73	4.03	-2.18	1.56
Laborat.		4.97	5.99	-0.14	3.31
Computers		2.75	7.31	-2.94	1.59
Ind. spaces		1.92	2.65	-2.21	0.81
Global satisfaction	1	1	1	-2.28	1.25
Factor variance	1.24	0.18	0.02[a]		
Factor correlation	0.48				

[a] Coefficient not statistically significant at $\alpha = 0.05$

All the loadings have the same sign. Since the sign of the latent factor is arbitrary (satisfaction or dissatisfaction with specific aspects), only the absolute values of the coefficients and their sign can be interpreted.

The factor *Human environment* has an higher variance then the factor *Physical environment*: the satisfaction with the human relationship has a stronger effect on global satisfaction. At the within level, the most important aspect relating to the *Human environment* is the *Relationship with professor* and the highest coefficient of the *Physical environment* is related to the opinion of respondents on the *Lecture rooms.* Given the mean (program) level of satisfaction, or in other words inside each study programs, the student opinion on the university personnel is affected mostly by the *Relationship with professor* and the student opinion on the physical services is affected mostly by the *Lecture rooms.*

The communalities representing the proportion of the variance of an indicator explained by the factors (section 1.3.1) are shown in Table 3.8.

Table 3.8: Two-level factor model: communalities. Students graduated (new system degree) at the University of Florence, year 2004.

Item	Communality			
	Within	Between	TOTAL	Between on TOTAL
Rel. prof. assistants	58.51	0.03	58.54	0.04
Rel. professors	93.47	0.02	93.48	0.02
Rel. technical staff	37.88	1.03	38.91	2.65
Lecture rooms	55.47	14.61	70.07	20.84
Computers	28.07	5.08	33.16	15.33
Laboratories	54.27	6.57	60.84	10.80
Library	25.30	14.96	40.26	37.15
Ind. spaces	16.76	2.65	19.41	13.65
Global satisfaction	36.27	0.30	36.56	0.81

Item *Relationship with professors* is really well explained by the factor structure, mostly due to the within level of the analysis. Also the items *Lecture rooms*, *Laboratories* and *Rel. prof. assistant* have high communalities, while the communality values of the items *Rel. technical staff*, *Computers*, *Library* and *Ind. spaces* indicate that the latent structure is not so proper to explain the phenomenon. Indicators with the highest between percentage of the communalities are *Computers*, *Lecture rooms* and *Laboratories.* The communality of the indicator *Global satisfaction* is not so high, and the between percentage is really low. Probably more latent dimensions that are not included in the model (such as the satisfaction with the program contents) are necessary to better explain the variability of this item.

The empirical Bayes prediction (section 1.5) allows to rank the 38 programs[6]. The complete ranking is shown in Table 3.9.

[6]As is always the case, the latent dimension underlying the global satisfaction at the program level has an arbitrary scale, which means that factor scores must be interpreted relatively to each other.

Table 3.9: Two-level mixture factor model: study program ranking based on the empirical Bayesian posterior distribution. Students graduated (new system degree) at the University of Florence, year 2004.

	Program	Faculty	Latent factor (Between level)	N resp.
1.	sc. statistiche	ECONOMICS	2.85	28
2.	sc. e tecn. chimiche	SCIENCE (MPNS)	2.60	18
3.	sc. matematiche	SCIENCE (MPNS)	1.74	15
4.	sc. e tecn. fisiche	SCIENCE (MPNS)	1.54	7
5.	sc. pol. e rel. intern.	POLITICAL SCIENCE	1.47	21
6.	sc. e tecn. farm.	PHARMACY	1.36	14
7.	sc. economia e gest. az.	ECONOMICS	1.36	223
8.	ing. industriale	ENGINEERING	1.35	69
9.	sc. giuridiche	LAW	1.35	63
10.	sc. economiche	ECONOMICS	1.34	29
11.	ing. civile e ambientale	ENGINEERING	1.33	34
12.	urbanistica	ARCHITECTURE	1.18	6
13.	sc. amministrazione	POLITICAL SCIENCE	1.17	10
14.	sc. biologiche	SCIENCE (MPNS)	1.12	6
15.	disegno industriale	ARCHITECTURE	0.79	9
16.	biotecnologie	INTER-FACULTY	0.62	13
17.	sc. sociali per cooperaz.	ECONOMICS	0.53	12
18.	filosofia	LETTERS AND PHILOSOPHY	0.42	6
19.	sc. servizio sociale	POLITICAL SCIENCE	0.11	20
20.	ing. informazione	ENGINEERING	0.04	57
21.	sc. e attività motorie	MEDICINE	0.02	24
22.	sc. e tecn. arti	LETTERS AND PHILOSOPHY	0.01	49
23.	sc. tecn. agrarie	AGRICULTURE	0.00	69
24.	sc. storiche	LETTERS AND PHILOSOPHY	-0.02	26
25.	lettere	LETTERS AND PHILOSOPHY	-0.04	24
26.	sc. geografiche	LETTERS AND PHILOSOPHY	-0.08	7
27.	sc. sociologiche	POLITICAL SCIENCE	-0.13	5
28.	sc. architettura ing. edile	ARCHIT.-ENGINEERING	-0.38	19
29.	sc. e tecn. infor.	SCIENCE (MPNS)	-1.14	32
30.	sc. educ. e a formaz.	EDUCATION SCIENCE	-1.33	62
31.	prof. sanitarie a riab.	MEDICINE	-1.35	39
32.	sc. e tecn. psicologiche	PSYCHOLOGY	-1.35	80
33.	sc. comunicazione	LETT. AND PHIL. - POLIT. SC.	-1.35	53
34.	lingue e culture moderne	LETTERS AND PHILOSOPHY	-1.36	74
35.	prof. inferm. e ostetrica	MEDICINE	-1.36	155
36.	sc. beni culturali	LETTERS AND PHILOSOPHY	-1.36	47
37.	tecn. conservaz. beni cult.	SCIENCE (MPNS)	-1.40	9
38.	prof. sanitarie tecniche	MEDICINE	-1.44	39

The programs with the highest students' satisfaction belong to the scientific area: the first program is *sc. statistiche*, followed by 3 programs of the Faculty of Science. On the contrary, at the end of the ranking there are programs of the Faculties of Medicine, of Science and of Letters and Philosophy. At the program level only the items related to the satisfaction with *Physical environment* and the *Relationship with technical staff* are significant, so the programs with a low rank should improve the lecture rooms, the computers, the laboratories and the library. They should also pay attention on the recruitment of the technical staff, being the relationship with technical staff the only important aspect in defining the satisfaction at program level among the items relative to the university personnel.

3.1.2 Two-level mixture factor model

With the multilevel factor model the latent structure underlying the phenomenon of students' satisfaction is analysed at both individual and study program level.

The aim of the thesis is also to classify second level units (programs) into groups of similar between structure; in particular the groups differ with respect to the item intercepts.

So, the multilevel mixture factor model described in section 1.3.2 is implemented. In particular, the traditional approach to Latent Class analysis is implemented using one categorical latent variable in order to classify the second level units (programs) in some classes and to determine the number of dimensions underlying the observed responses (Magidson and Vermunt, 2001). The model is represented in Figure 3.3.

At the first level there is a standard factor model (equation (1.7)). Since the focus of the dissertation is on the program level, the factor structure at the student level implemented in the multilevel factor model is retained. Then, it is assumed that data originated from different populations of programs introducing in the model a latent categorical variable at the group level.

The complete model is (section 1.3.2):

$$\begin{aligned} \upsilon_{hij} &= \mu_{hj} + \sum_{m=1}^{M_2} \lambda_{mh}^{(2)} \eta_{mij}^{(2)} \\ \mu_{hj} &= \sum_{k=1}^{K} \lambda_{hk}^{(3)} \eta_{jk}^{(3)} + e_{hj}^{(3)} \end{aligned} \tag{3.2}$$

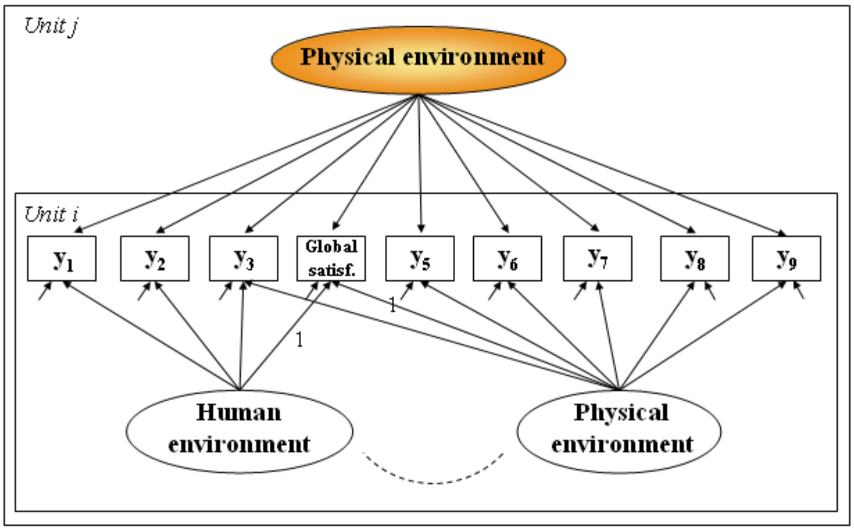

Figure 3.3: Two-level mixture factor model. Students graduated (new system degree) at the University of Florence, year 2004.

where

$$\begin{aligned} \boldsymbol{\eta}_{ij}^{(2)} &\sim MN(\mathbf{0}, \boldsymbol{\Psi}^{(2)}) \\ \pi_k &= P(\eta_j^{(3)} = k) = P(\eta_{jk}^{(3)} = 1) \end{aligned}$$

with

$$\sum_{k=1}^{K} \pi_k = 1.$$

Cluster-level item-specific errors $e_{hj}^{(3)}$ in equation (3.2) are constrained to zero and the additional constraint $\sum_{k=1}^{K} \lambda_{hk}^{(3)} = 0$ is imposed for each item h.

In order to choose between models with different number of classes at group level, the BIC index[7] calculated with N equal to the number of groups is used (section 1.4.2). Table 3.10 shows BIC values for models composed of 2 to 8 classes. The selected model has 5 classes.

[7]As described in Chapter 1, there are many methods to compare models with different numbers of latent classes. The software Latent GOLD reports some indexes, such as the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC), the Akaike Information Criterion 3 (AIC3), and the Consistent Akaike Information Criterion (CAIC) (Vermunt and Magidson, 2005*c*); it also provides a bootstrap method, and it does not calculate the Lo Mendell Rubin test. In this case study and in the evaluation of job satisfaction (section 3.2.1), the BIC is used since it is generally considered as a good indicator for class enumeration (Nylund et al., 2007).

Table 3.10: Two-level mixture factor model: loglikelihood and fit indexes. Students graduated (new system degree) at the University of Florence, year 2004.

Classes N	Param. N	LogLikelihood	BIC (N obs.)	BIC (N groups)	AIC
1	30	-10357.05	20932.96	20823.24	20774.11
2	40	-10216.14	20724.08	20577.78	20512.28
3	50	-10126.95	20618.65	20435.78	20353.90
4	60	-10091.85	20621.41	20401.96	20303.70
5	70	-10064.88	20640.42	20384.40	20269.77
6	80	-10049.74	20683.09	20390.49	20259.48
7	90	-10037.73	20732.01	20402.83	20255.45
8	100	-10018.40	20766.30	20400.55	20236.79

As in the previous analysis, the bivariate residuals reported in Table 3.11 show that the local independence assumption is reasonable and that in that aspect the model fits the data well.

Table 3.11: Two-level mixture factor model: bivariate residual statistics. Students graduated (new system degree) at the University of Florence, year 2004.

Item	1	2	3	4	5	6	7	8	9
1. Rel. prof. assistant	.								
2. Rel. professors	0.151	.							
3. Rel. technical staff	0.109	0.115	.						
4. Lecture rooms	0.116	0.009	0.358	.					
5. Library	0.290	0.095	0.317	0.120	.				
6. Laboratories	0.239	0.070	0.412	0.164	0.296	.			
7. Computers	0.011	0.081	0.153	0.038	0.233	0.036	.		
8. Ind. spaces	0.170	0.011	0.081	0.027	0.835	1.048	0.463	.	
9. Global satisfaction	0.287	0.151	0.184	0.080	0.116	0.735	0.487	0.289	.

The estimated factor loadings at the student level (Tab. 3.12) are very similar to the one estimated through the multilevel factor model: the factor *Human environment* has the highest loading on global satisfaction, the most important aspect relating to the *Human environment* is the *Relationship with professor* and the most important aspect relating to the *Physical environment* is the students opinion on the *Lecture rooms.*

The model classifies second level units in five classes with homogeneous between structure. Probabilities of the classes are quite different (Tab. 3.13): a group k has a probability of 41% to belong to the first class, of 24% to belong

Table 3.12: Two-level mixture factor model: parameter estimates. Students graduated (new system degree) at the University of Florence, year 2004.

Item	Loadings Within: Human environment	Loadings Within: Physical environment	Thresholds α_1	Thresholds α_2
Rel. prof. assistants	1.92		-2.27	2.02
Rel. professors	3.43		-4.49	3.58
Rel. techn. staff	0.94	1.54	-0.29	2.65
Lecture rooms		5.52	-0.24	3.49
Library		3.03	-1.94	1.96
Laborat.		5.43	0.08	3.92
Computers		2.87	-2.29	2.26
Ind. spaces		2.09	-1.93	1.19
Global satisfaction	1	1	-2.31	1.20
Factor variance	1.29	0.18		
Factor correlation	0.49			

Table 3.13: Two-level mixture factor model: latent classes probabilities. Students graduated (new system degree) at the University of Florence, year 2004.

k	1	2	3	4	5
$P(\eta_k^{(3)} = 1)$	0.414	0.053	0.244	0.109	0.179

to the third class and of 18% to belong to the fifth class. The probability to belong to the other classes is quite low.

The model results relative to the second level of analysis are reported in Table 3.14 and 3.15. The $\lambda_{hk}^{(3)}$ coefficients of equation (3.2) are represented with reversed sign for interpretational simplicity: the higher a coefficient is, the higher is the probability to be satisfied for each item. The null hypothesis of the Wald test states that all the effects associated with each indicator are zero. In this case study, all the indicators discriminate between the classes in a statistically significant way; only the indicator *Rel. prof. assistants* has a p-value slightly higher than 0.5.

Figure 3.4 shows the features of each class. Furthermore, for each item the constraint $\sum_{k=1}^{K} \lambda_{hk}^{(3)} = 0$ is used, so the class-specific effects should be interpreted in terms of deviation from the "average class" where the effects are equal to 0.

The third class of programs is the "best": all the coefficients are positive

Table 3.14: Two-level mixture factor model: parameter estimates, study program level. Students graduated (old system degree) at the University of Florence, summer session year 2004.

Item	Class	Coeff.	Wald test (=0)	df	p-value
	Class 1	-0.152			
	Class 2	0.012			
Rel. prof. assistants	Class 3	0.526	9.221	4	0.0560
	Class 4	0.002			
	Class 5	-0.387			
	Class 1	-0.592			
	Class 2	0.211			
Rel. professors	Class 3	1.221	15.306	4	0.0041
	Class 4	0.019			
	Class 5	-0.860			
	Class 1	0.430			
	Class 2	0.523			
Rel. technical staff	Class 3	0.804	81.560	4	0.0000
	Class 4	-1.025			
	Class 5	-0.732			
	Class 1	1.168			
	Class 2	-1.570			
Lecture rooms	Class 3	2.891	164.040	4	0.0000
	Class 4	-1.131			
	Class 5	-1.359			
	Class 1	0.434			
	Class 2	-1.454			
Library	Class 3	0.900	94.086	4	0.0000
	Class 4	0.442			
	Class 5	-0.322			
	Class 1	0.175			
	Class 2	-0.489			
Laboratories	Class 3	3.062	106.218	4	0.0000
	Class 4	-0.294			
	Class 5	-2.454			
	Class 1	0.752			
	Class 2	-1.838			
Computers	Class 3	2.060	165.908	4	0.0000
	Class 4	-0.663			
	Class 5	-0.312			

Table 3.15: Two-level mixture factor model: parameter estimates, study program level (*continued*). Students graduated (old system degree) at the University of Florence, summer session year 2004.

Item	Class	Coeff.	Wald test (=0)	df	p-value
	Class 1	0.399			
	Class 2	-0.946			
Ind. spaces	Class 3	0.712	62.330	4	0.0000
	Class 4	-0.496			
	Class 5	0.331			
	Class 1	-0.047			
	Class 2	0.120			
Global satisfaction	Class 3	0.472	20.983	4	0.0003
	Class 4	0.049			
	Class 5	-0.594			

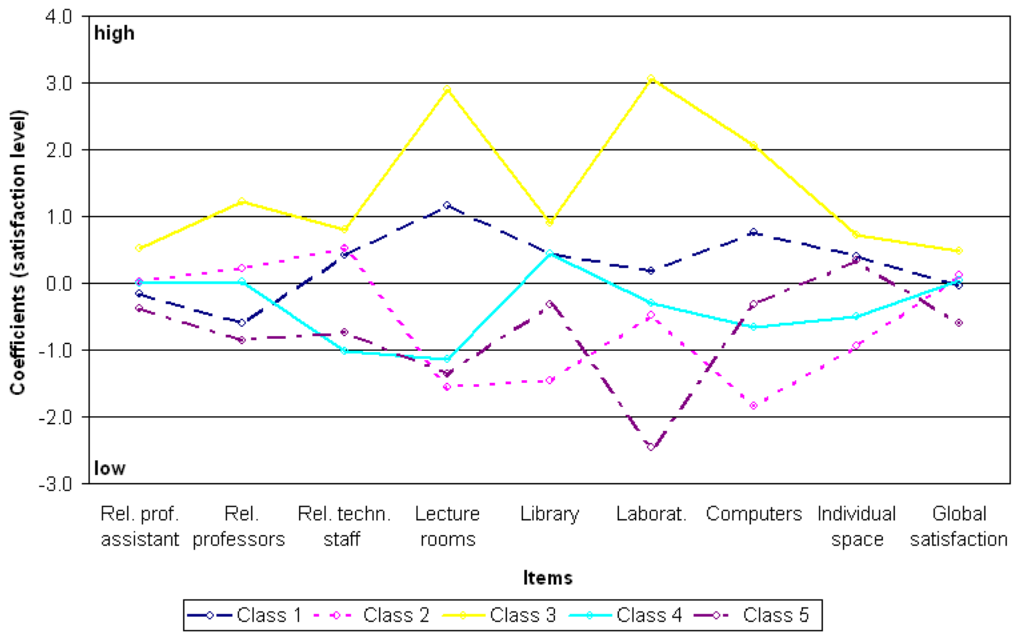

Figure 3.4: Two-level mixture factor model: latent classes features. Students graduated (new system degree) at the University of Florence, year 2004.

and almost all coefficients have the highest absolute value. A program belonging to the third class has higher satisfaction measured by each indicators than the programs belonging to the other classes. On the contrary, in the fifth class all coefficients are negative, except the parameters relative to the *Individual spaces*. Class 4 is a kind of "average class": the coefficients are quite near to 0 except for the items related to *Relationship with professors* and *Lecture rooms*. In class 2 there is a general satisfaction with the *Human environment* and dissatisfaction relating to the *Physical environment*. In all classes coefficients related to the two items *Relationship with professors* and *Relationship with professors assistant* have the same sign of the indicator *Global satisfaction*; this confirms the importance of these two aspects in determining students' satisfaction.

Finally, the coefficients range is higher for the indicators relating to the *Physical environment* than for the indicators *Relationship with professors* and *Relationship with professors assistant* and *Relationship with technical staff*, except for the indicator *Individual spaces*: the mean opinion of students is more heterogeneous with respect to the "physical" features of the programs, while it offsets relating to the personnel. This confirms the results of the multilevel factor model (section 3.1.1) and indicates the existence of an unique underlying dimension at the second level of the analysis relating to the *Physical environment*.

Giving the observed manifest variables at student level, the 38 programs are classified with the empirical Bayes modal prediction (section 1.5), as shown in Table 3.16.

Programs of the third class (the "best") belong mostly to the Faculties of Economics and Science. On the other hand, in the "worst" class (the fifth class) there are programs of the Faculty of Letters and Philosophy. In class 2, with only two elements, there are programs of the Faculty of Medicine. In this class there is a general satisfaction with the *Human environment* and dissatisfaction relating to the *Physical environment*, so the programs should improve the "physical" aspects. Programs belonging to class 4 present a quite high dissatisfaction relative only to the *Lecture rooms* and *Relationship with technical staff*. The first class is the biggest, with 16 programs belonging to different Faculties.

Some information on the goodness of the classification come from the so called "Classification Table" (section 1.5) that cross-tabulates modal and probabilistic class assignments. As shown in Table 3.17 the classification of second level units is quite good, giving very similar results with the two methods.

Table 3.16: Two-level mixture factor model: study program classification based on the empirical Bayesian posterior distribution. Students graduated (new system degree) at the University of Florence, year 2004.

Class	Program	Faculty	N resp.
1	sc. tecn. agrarie	AGRICULTURE	69
1	disegno industriale	ARCHITECTURE	9
1	sc. economia e gest. az.	ECONOMICS	223
1	ing. industriale	ENGINEERING	69
1	ing. civile e ambientale	ENGINEERING	34
1	ing. informazione	ENGINEERING	57
1	biotecnologie	INTER-FACULTY	13
1	sc. giuridiche	LAW	63
1	filosofia	LETTERS AND PHILOSOPHY	6
1	sc. e tecn. arti	LETTERS AND PHILOSOPHY	49
1	sc. storiche	LETTERS AND PHILOSOPHY	26
1	sc. e attività motorie	MEDICINE	24
1	sc. amministrazione	POLITICAL SCIENCE	10
1	sc. servizio sociale	POLITICAL SCIENCE	20
1	sc. sociologiche	POLITICAL SCIENCE	5
1	sc. biologiche	SCIENCE (MPNS)	6
2	prof. inferm. e ostetrica	MEDICINE	155
2	prof. sanitarie tecniche	MEDICINE	39
3	urbanistica	ARCHITECTURE	6
3	sc. statistiche	ECONOMICS	28
3	sc. economiche	ECONOMICS	29
3	sc. sociali per cooperaz.	ECONOMICS	12
3	sc. e tecn. farm.	PHARMACY	14
3	sc. pol. e rel. intern.	POLITICAL SCIENCE	21
3	sc. e tecn. chimiche	SCIENCE (MPNS)	18
3	sc. matematiche	SCIENCE (MPNS)	15
3	sc. e tecn. fisiche	SCIENCE (MPNS)	7
4	sc. architettura ing. edile	ARCHIT.-ENGINEERING	19
4	sc. educ. e a formaz.	EDUCATION SCIENCE	62
4	prof. sanitarie riab.	MEDICINE	39
4	tecn. conservaz. beni cult.	SCIENCE (MPNS)	9
5	lettere	LETTERS AND PHILOSOPHY	24
5	sc. geografiche	LETTERS AND PHILOSOPHY	7
5	lingue e culture moderne	LETTERS AND PHILOSOPHY	74
5	sc. beni culturali	LETTERS AND PHILOSOPHY	47
5	sc. comunicazione	LETT. AND PHIL. - POLIT. SCIENCE	53
5	sc. e tecn. psicologiche	PSYCHOLOGY	80
5	sc. e tecn. infor.	SCIENCE (MPNS)	32

Table 3.17: Two-level mixture factor model: classification table. Students graduated (new system degree) at the University of Florence, year 2004.

	Modal					
Probabilistic	Class 1	Class 2	Class 3	Class 4	Class 5	Total
Class 1	14.7	0.0	0.6	0.0	0.5	15.7
Class 2	0.0	2.0	0.0	0.0	0.0	2.0
Class 3	0.9	0.0	8.4	0.0	0.0	9.3
Class 4	0.1	0.0	0.0	4.0	0.1	4.1
Class 5	0.4	0.0	0.0	0.0	6.4	6.8
Total	16.0	2.0	9.0	4.0	7.0	38

3.1.3 Two-level factor model and two-level mixture factor model

In this section the results of the multilevel factor model and the multilevel mixture factor model are merged.

The focus of this section is on the between level of the analysis. While at student level the same underlying structure is used to explain the phenomenon of individual satisfaction, at the program level the multilevel factor model assumes continuous latent factors and the multilevel mixture factor model assumes a categorical latent variable.

The models give similar results relative to the differentiation of the programs.

In the multilevel factor model, at program level the only latent continuous factor underlying the phenomenon is measured by the items relating to the *Physical Environment* (items on lecture rooms, library, laboratories, computers, individual spaces and item on the relationship with technical staff); the programs differ only in the satisfaction with the *Physical Environment.*

Through the multilevel mixture factor model, the second level units are classified in 5 classes with homogeneous features. The classes have different size and are characterised mostly by the indicators relating to the *Physical Environment* and to the relationship with technical staff.

These results confirm that there is an unique underlying dimension at the second level of the analysis and that programs differentiate mostly on physical characteristics, probably because students opinion on human aspects are balanced. This is quite understandable, indeed university personnel have really different features that students, being different as well, may appreciate or not appreciate. Furthermore, in the university system there are many persons (for example students encounter many professors in their academic life), so it is probable that students have different opinions on different persons and these opinions are balanced at program level. In order to reach an higher level of mean satisfaction, programs should improve mostly the lecture rooms, the computers, the laboratories and the library and they should also pay attention on the recruitment of the technical staff. Of course, also the *Human Environment* is important to define the students' satisfaction, but these aspects influence more the differences among students (individual level) than the differences among programs (group level).

Useful information on the latent variables are obtained through the empirical Bayes prediction. In the multilevel factor model the 38 programs are ranked along a continuum (Tab. 3.9) and in the multilevel mixture factor model the programs are classified in five classes (Tab. 3.16).

In Figure 3.5 results of the multilevel factor model and the multilevel mixture factor analysis are merged[8].

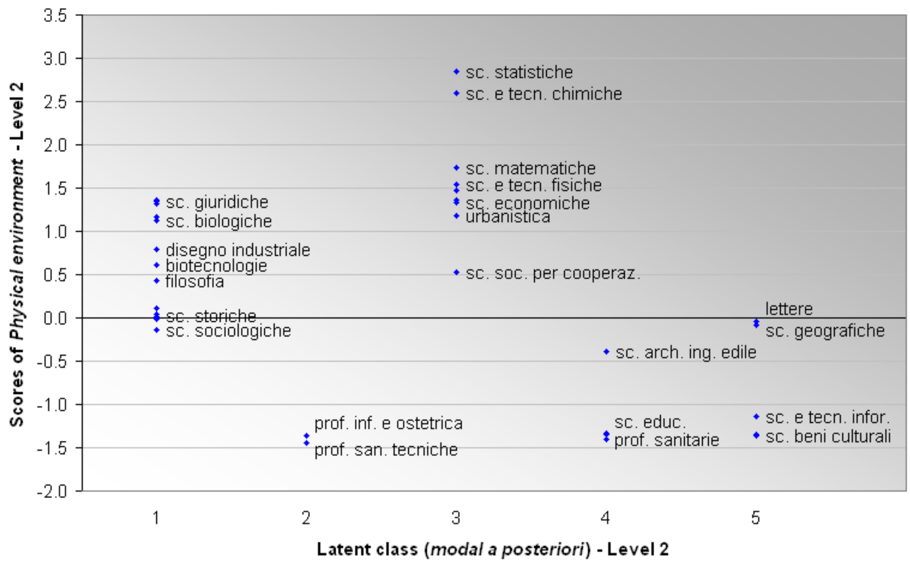

Figure 3.5: Two-level factor model and two-level mixture factor model: study programs factor scores and classes. Students graduated (new system degree) at the University of Florence, year 2004.

On the y-axis there is the global ranking of the programs. Looking only at this, some groups of programs cannot be distinguished. For example, the two programs *sc. storiche* and *lettere*, both of the Faculty of Letters and Philosophy have a rank of, respectively, -0.02 and -0.04; so they are quite similar. Looking at the results of the multilevel mixture factor model, the two

[8] In the figure, some program labels are not displayed due to lack of space.

programs are classified in different classes: *sc. storiche* is in class 1 and *lettere* is in class 5. These two classes have a similar global satisfaction level and a similar satisfaction with the relationship with professors and professors' assistant, but they are characterized by very different satisfaction level on lecture room and laboratories. Assuming that students of these programs probably do not use a lot the laboratories, *lettere* should improve especially the lecture rooms.

Another example is on the differences between *sc. giuridiche* (Faculty of Law) and *sc. economiche* (Faculty of Economics), ranked 1.35 and 1.34 on the continuous underlying factor at second level. The two programs belong respectively to class 1 and 3. Both classes have positive features, having almost all the parameters a plus sign; the main differences are relative to the items *Relationship with professors*, *Lecture rooms* and *Laboratories*. Looking at the results, two aspects need attention. The percentage of students that did not used or did not reply to the question on the laboratories is almost the same: about 40% on the 63 students who graduated in *sc. giuridiche* and 29 students who graduated in *sc. economiche*. Since 2004, the Faculty of Law and the Faculty of Economics are ubicated in the same area of Florence and their buildings have the same size but really different organization; the satisfaction level is quite different for the two programs probably because the internal spaces of the Faculty of Economics are better organized, for example, there are not so many students per each lesson. Concluding, *sc. giuridiche* should improve mostly the *Lecture rooms* and *Laboratories*.

The classification of programs obtained through the multilevel mixture factor model is represented on the x-axis. Looking only at this classification, programs inside each class cannot be distinguished; this is possibile using the information obtained through the multilevel factor model. For example, in class 3 the programs *scienze sociali per la cooperazione* and *scienze statistiche* have really different scores, respectively, of 0.53 and 2.85, and in class 1 *sc. sociologiche* and *sc. giuridiche* are scored -0.13 and 1.35. Programs with lower scoring have more critical aspects than the others; analysing the features of each class (section 3.1.2) the university government may solve these problems focusing its (economical and political) efforts.

Merging the results of the two analyses let to understand in a better way the phenomenon of satisfaction at program level. Of course, the substantive knowledge of the phenomenon is necessary to correctly interpret the statistical results, as shown with these simple examples.

3.2 University external effectiveness

In this section the results of the analysis of the university external effectiveness evaluated from the graduates' point of view are shown. The analysis is carried out through a multilevel mixture factor model. The job satisfaction (perceived quality) of students who graduated from the University of Florence is analysed as an indicator of the university external effectiveness. Data come from the AlmaLaurea survey on the "Employment opportunities, 2005". The focus is on students who graduated with the old university system during the year 2004 and are working at the moment of the interview.

The variables used in the analysis belong to the first section of the AlmaLaurea questionnaire described in section 2.3, with questions for graduates working at the moment of the interview. As mentioned previously, the item *Free time* is not used in the analysis, having low correlation with almost all other items. Therefore, the item used for the analysis are 14.

The 837 graduates working at the interview attended 42 different programs; because of the low number of students, some programs are grouped and some programs are excluded from the analysis: the final number of distinct programs is 23, and the final number of students is 826.

3.2.1 Two-level mixture factor model

As for the phenomenon of students' satisfaction about the university system, considering the nature of the phenomenon and the data coming from the AlmaLaurea survey, the most suitable statistical methodology for the the analysis of job satisfaction is the factor model. Data have a relevant hierarchical structure (first level units are the students and second level units are the study programs they attended), so multilevel tools of analysis are used.

The aims of the research are to study the phenomenon of job satisfaction at individual level (so to understand which are the important aspects of job that determine the individuals' satisfaction) and to classify the programs attended by the graduates into classes homogeneous respect to the job satisfaction. At the individual level of the analysis a traditional factor model is implemented, while at the second level a traditional approach to latent class analysis is used.

As the first step of the analysis, an exploratory factor analysis[9] is implemented. All indicators are expressed on an ordinal scale, with 10 categories. They are considered as continuous normal variables because of the number

[9] The EFA has been implemented with the software Mplus (Muthén and Muthén, 1998-2007).

of the categories. Of course, the normal distribution is only an approximation but, as suggested by Muthén and Kaplan (1985), the normal theory estimators perform quite well even with ordered categorical and moderately skewed-kurtotic variables, at least when the sample is not small, as in this case-study.

The fit statistics of the exploratory factor analysis are shown in Table 3.18. The two indexes Root Mean Square Error of Approximation (RMSEA) and Root Mean Square Residual (RMSR) are used to compare different models with different latent structures. The structure suggested by the fit statistics has 3 correlated factors; for substantive reasons also the structure with 4 factors is analysed with a confirmatory factor model.

Table 3.18: EFA: fit statistics. Students graduated (old system degree) at the University of Florence, summer session year 2004.

	3 Factors	4 Factors
CHI-SQUARE VALUE	179.191	89.698
DEGREES OF FREEDOM	42	32
PROBABILITY VALUE	0.000	0.000
RMSEA	0.063	0.047
RMSR	0.030	0.018

The exploratory factor analysis indicates the presence of three (four) underlying factors: one factor represents the satisfaction related to the earning and stability (*Earning*), one is related to the satisfaction with cultural aspects (*Cultural aspects*) and one is related to the satisfaction with the autonomy and the job environment (*Autonomy and environment*); the last factor can be subdivided into two correlated factors: *Autonomy* and *Environment*.

Confirmatory factor analyses are then implemented, the results are shown in Table 3.19 and in Table 3.20. The most proper solution, relatively both to statistical results and substantive reasons, has 4 latent continuous correlated factors referring, respectively, to the satisfaction with earning and stability (*Earning*), cultural aspects (*Cultural*), job environment (*Environment*) and autonomy (*Autonomy*).

The only not significant coefficient is the loading of the global job satisfaction on the latent factor *Autonomy*, while the most important aspect determining the job satisfaction is the factor *Cultural* aspects. The factor *Earning* and stability has a statistically significant, but not high effect on the global job satisfaction. Probably this is due to the fact that the analysis concerns only graduates working one year after the degree: the first job is considered as a first experience required to provide professionalism and experience, but it has usually a limited duration.

Table 3.19: Confirmatory factor analysis: parameter estimates. Students graduated (old system degree) at the University of Florence, summer session year 2004.

Item	Loadings			
	Earning	Cultural	Autonomy	Environment
Global satisfaction	-0.24	1.16	0.02[a]	-0.43
Steadiness	-1.00			
Coherence with studies		1.88		
Professionalism		1.45		
Prestige	-0.68	0.92		-0.41
Cultural interests		1.85		
Social utility		1.20		
Independence			-1.51	
Involvement		0.49	-1.22	
Schedule flexibility			-1.25	
Job place				-0.93
Relationship with colleagues				-0.78
Salary	-1.86			
Carrier	-2.12			

[a] Coefficient not statistically significant at $\alpha = 0.05$

Table 3.20: Confirmatory factor analysis: factor correlations. Students graduated (old system degree) at the University of Florence, summer session year 2004.

	Earning	Cultural	Autonomy	Environment
Earning	1			
Cultural	-0.62	1		
Autonomy	0.50	-0.60	1	
Environment	0.31	-0.37	0.54	1

Finally, a two-level confirmatory mixture factor analysis is implemented in order to confirm the hypothesized structure at the individual level and to classify programs in some homogeneous groups. The two-level mixture factor model for continuous outcomes is expressed by (section 1.3.2):

$$y_{hij} = \mu_h + \sum_{m=1}^{M_2} \lambda_{mh}^{(2)} \eta_{mij}^{(2)} + e_{hij}$$

and, for each latent factor $\eta^{(2)}_{mij}$, $m = 1, \ldots, M_2$:

$$E(\eta^{(2)}_{mij}) = \sum_{k=1}^{K} \lambda^{(3)}_{mk} \eta^{(3)}_{jk} \tag{3.3}$$

where $\eta^{(3)}_{jk}, k = 1, \ldots, K$ is an indicator variable taking the value 1 with probability π_k if unit i belongs to latent class k and 0 otherwise, and $\eta^{(3)}_j = (\eta^{(3)}_{j1}, \ldots, \eta^{(3)}_{jk})$ has a multinomial distribution; of course, $\sum_{k=1}^{K} \pi_k = 1$.

At the program level, it is assumed that programs differ in the mean level of latent factors at the individual level ($\eta^{(2)}_{mij}$). In other words, $\lambda^{(3)}_{mk}$ represents the mean of the m-th factor at individual level for the programs belonging to the k-th latent class. The model is represented in Figure 3.6.

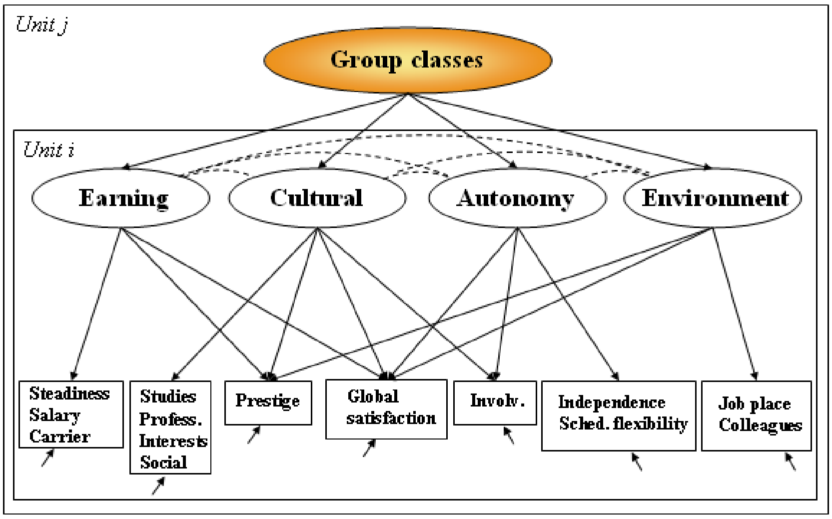

Figure 3.6: Two-level mixture factor model. Students graduated (old system degree) at the University of Florence, summer session year 2004.

A factor model for the between (programs) component could also be formulated in order to analyse the factor structure at between level. This analysis gives information about the differences of the underlying latent structure at both levels highlighting the differences between them. However, in this case study, it is not so important to understand which are the important aspects of job that determine the satisfaction of individuals at the between level. Indeed, the main aim of the thesis is to be a useful tool for the local university policy, and the university governments cannot act directly on specific aspects of the job. It is more important to classify the programs in classes similar in the mean level of individual job satisfaction in order to understand which are the programs that are able to "provide" a good perceived quality of the job.

Moreover the number of indicators of job satisfaction is quite high (14) and the analysis of the difference between the relative weight of each item at between level can be "distracting" for the university policy. After analysing which are the important aspects of job that determine the satisfaction of individuals, it is interesting to analyse the difference of programs in the mean level of each individual latent factor (less in the number respect to the items). A classification of programs gives information about similarities and peculiarities of programs relatively to the students job satisfaction.

In the questionnaire there is also a direct question on the job satisfaction. This is used in the model in order to estimate the effect of each underlying factor on global satisfaction (section 3.1.1). In this model, the variance of the factors at individual level is constrained to 1, and the effect of each factor on the job satisfaction is "measured" through its factor score[10]. At the second level of the analysis, the cluster-level item-specific errors $e^{(3)}_{hj}$, implicit in equation (3.3) are constrained to zero to reduce the computational burden, while $\sum_{k=1}^{K} \lambda^{(3)}_{mk} = 0$ for each m, $m = 1, \ldots, M_2$, to ensure the identification.

The Bayesian Information Criterion index calculated with N equal to the number of groups is used to choose between models with different number of classes at group level. Table 3.21 shows BIC values for models composed of 2 to 5 classes. The selected model has 3 classes.

Table 3.21: Two-level mixture factor model: loglikelihood and fit indexes. Students graduated (old system degree) at the University of Florence, summer session year 2004.

Classes N	Param. N	LogLikelihood	BIC (N obs.)	BIC (N groups)	AIC
Class1	54	-22252.83	44868.35	44674.97	44613.65
Class2	59	-22218.07	44832.41	44621.13	44554.13
Class3	64	-22209.27	44848.39	44619.20	44546.53
Class4	69	-22205.44	44874.33	44627.24	44548.89
Class5	74	-22200.02	44897.06	44632.06	44548.04

The bivariate residual statistics of the model are reported in Table 3.22.

[10]These constraints are different from the constraints used in section 3.1.1. In the study of satisfaction with the university system the responses are all ordinal and with Latent GOLD the factor variances cannot be constrained, but only the variance of the variables obtained through the Cholensky decomposition can be constrained.

Some values are higher than 3.84, so the local independence assumption does not hold for each pair of items. Relaxing the independence assumption for some items would increase the number of parameters to be estimated and complicate the interpretation of the model. Considering also that the classification statistics later presented indicate a quite good behaviour of the model, the model is retained.

At individual level, the factor structure (Table 3.23) is very similar to the structure found with the confirmatory one-level factor analysis (Table 3.20). All coefficients, except one, are statistically significant. The only coefficient not significant is the loading of the global job satisfaction on the latent factor *Autonomy*. The most important aspect determining the job satisfaction is the factor *Cultural* aspects. Of course, the latent factors at the student level are strongly correlated (Table 3.24).

The model classifies the programs in three classes with homogeneous individual factor means. The probabilities of the classes are quite different: a program k has a probability 0.68 to belong to the second class, 0.16 to the first class and 0.15 to the third class (Table 3.25).

Table 3.26 shows the parameter estimates of the model at program level, with the Wald tests for the equality of the parameters.

The three classes differ only for the mean level of the two latent factors *Earning* and *Cultural*, while the other two factors do not discriminate the classes. This can be easily seen also in Figure 3.7 and Figure 3.8. Due to the constraints $\sum_{k=1}^{K} \lambda_{mk}^{(3)} = 0$, the class-specific effects must be interpreted in terms of deviations from the "average class" where the effects are equal to 0.

Class 1 of programs is the "worst": the mean level of the latent factors *Earning* and *Cultural* is really low respect to the average class. Class 3 has a mean level of the satisfaction with *Earning* and stability, while the level of satisfaction with *Cultural* aspects is quite high. Class 2 has a behavior opposite to class 3: the satisfaction with *Earning* and stability is really high, while the satisfaction with *Cultural* is not so high.

The 23 programs are classified as shown in Table 3.27 using the empirical Bayes modal prediction (column 6).

The three programs assigned to the first class (the "worst") belong to the Faculties of Letters and Psychology. Usually, students with a degree in these programs need more than one year to find the job they studied for. In the third class there are two programs of the Faculty of Education and one program of the Faculty of Engineering. The other class is the biggest, with 17 programs; in the data there is not sufficient information to further divide this class.

Looking at the classification statistics (section 1.4.2) based on the poste-

Table 3.22: Two-level mixture factor model: bivariate residual statistics. Students graduated (old system degree) at the University of Florence, year 2004.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Global satisf.	.													
Steadiness	7.099	.												
Coher. studies	4.905	3.338	.											
Professionalism	0.058	0.528	0.685	.										
Prestige	1.467	0.524	9.023	19.056	.									
Cultural int.	0.796	25.463	26.566	23.058	3.936	.								
Social utility	0.255	0.045	0.030	5.726	4.227	20.701	.							
Independence	0.145	2.775	1.168	0.465	0.216	0.812	4.349	.						
Involvement	0.043	1.938	4.096	0.177	1.501	0.890	1.941	0.171	.					
Sched. flexib.	0.810	10.987	0.415	0.179	0.758	0.065	0.394	0.012	0.823	.				
Job place	0.506	5.437	0.078	0.028	1.312	0.115	0.037	0.002	0.093	3.928	.			
Relationship with colleagues	0.000	0.083	0.039	5.021	0.081	0.108	0.135	1.270	0.566	0.252	0.680	.		
Salary	5.819	9.131	2.120	2.680	0.133	10.835	9.390	0.405	0.425	0.058	0.723	0.010	.	
Carrier	8.905	7.776	7.712	6.358	0.621	9.905	3.169	2.516	0.508	0.780	0.427	0.089	0.016	.

Table 3.23: Two-level mixture factor model: parameter estimates. Students graduated (old system degree) at the University of Florence, summer session year 2004.

Item	Loadings			
	Earning	Cultural	Autonomy	Environment
Global satisfaction	0.22	1.12	0.02[a]	0.43
Steadiness	0.96			
Coherence with studies		1.81		
Professionalism		1.39		
Prestige	0.65	0.88		0.41
Cultural interests		1.77		
Social utility		1.14		
Independence			1.52	
Involvement		0.49	1.20	
Schedule flexibility			1.24	
Job place				0.93
Relationship with colleagues				0.77
Salary	1.80			
Carrier	2.03			

[a] Coefficient not statistically significant at $\alpha = 0.05$

Table 3.24: Two-level mixture factor model: factor correlations, individual level. Students graduated (old system degree) at the University of Florence, summer session year 2004.

	Earning	Cultural	Autonomy	Environment
Earning	1			
Cultural	0.60	1		
Autonomy	0.52	0.61	1	
Environment	0.32	0.37	0.54	1

Table 3.25: Two-level mixture factor model: latent classes probabilities. Students graduated (old system degree) at the University of Florence, summer session year 2004.

k	1	2	3
$P(\eta_k^{(3)} = 1)$	0.155	0.682	0.163

Table 3.26: Two-level mixture factor model: parameter estimates, study program level. Students graduated (old system degree) at the University of Florence, summer session year 2004.

Item	Class	Coeff.	Wald test (=0)	df	p-value
	Class 1	-0.296			
Earning	Class 2	0.342	54.952	2	0.0000
	Class 3	-0.046			
	Class 1	-0.476			
Cultural	Class 2	0.155	58.614	2	0.0000
	Class 3	0.321			
	Class 1	-0.087			
Autonomy	Class 2	-0.018	1.894	2	0.3900
	Class 3	0.105			
	Class 1	-0.094			
Environment	Class 2	0.006	1.243	2	0.5400
	Class 3	0.089			

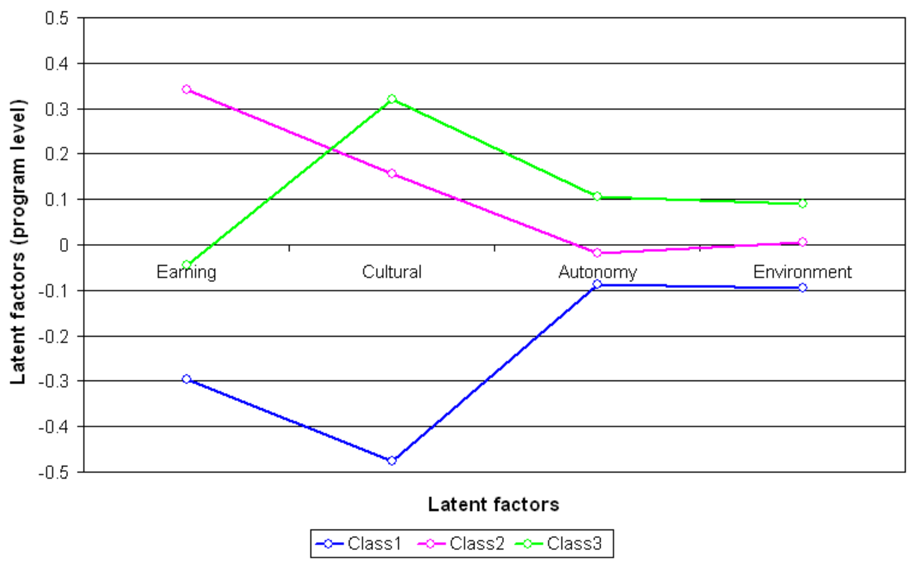

Figure 3.7: Two-level mixture factor model: latent classes features. Students graduated (old system degree) at the University of Florence, summer session year 2004.

Table 3.27: Two-level mixture factor model: study program classification based on the empirical Bayesian posterior distribution. Students graduated (old system degree) at the University of Florence, summer session year 2004.

Program	Faculty	Probability			Class	N
		Class 1	Class 2	Class 3	(Modal)	resp.
lettere	LETTERS	1	0	0	1	63
lingue e lett. straniere	LETTERS	0.99	0.01	0	1	38
psicologia	PSYCHOLOGY	1	0	0	1	131
gruppo scientifico	SCIENCE (MPNS)	0	0.92	0.08	2	13
altro sc. formazione	EDUCATION SC.	0.01	0.77	0.22	2	8
scienze statistiche	ECONOMICS	0.15	0.83	0.02	2	8
gruppo geo-biologico	SCIENCE (MPNS)	0.01	0.86	0.13	2	17
altro ingegneria	ENGENEERING	0	0.98	0.02	2	12
altro agraria	AGRICULTURE	0.01	0.97	0.03	2	9
architettura	ARCHITECTURE	0	1	0	2	155
economia aziendale	ECONOMICS	0	1	0	2	40
economia e commercio	ECONOMICS	0	1	0	2	34
filosofia	LETTERS	0.27	0.43	0.3	2	14
giurisprudenza	LAW	0.01	0.99	0	2	37
ingegneria civile	ENGENEERING	0	0.96	0.04	2	11
ingegneria elettronica	ENGENEERING	0	0.99	0.01	2	19
ingegneria meccanica	ENGENEERING	0	0.99	0.01	2	26
scienze politiche	POLITICAL SC.	0	1	0	2	66
ing. per amb. e territ.	ENGENEERING	0	0.92	0.08	2	11
sc. forestali ed amb.	AGRICULTURE	0.07	0.72	0.21	2	10
ingegneria informatica	ENGENEERING	0.04	0.35	0.62	3	10
scienze educazione	EDUCATION SC.	0	0	1	3	75
sc. formaz. primaria	EDUCATION SC.	0	0.01	0.99	3	19

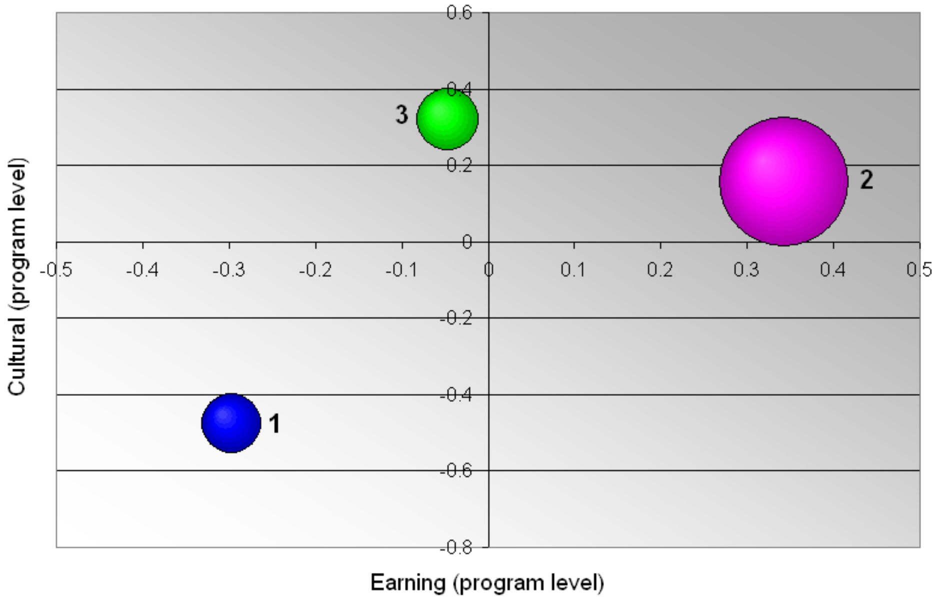

Figure 3.8: Two-level mixture factor model: latent classes features (latent variables *Earning* and *Cultural*). Students graduated (old system degree) at the University of Florence, summer session year 2004.

rior class membership probabilities, the proportion of classification errors is equal to 0.091, and the pseudo R-squared statistics $R^2_{\boldsymbol{\eta},errors}$, $R^2_{\boldsymbol{\eta},entropy}$ and $R^2_{\boldsymbol{\eta},variance}$ are respectively equal to 0.714, 0.717 and 0.724 indicating a global quite good classification[11] of the study programs.

Finally, the "Classification Table" cross-tabulates the two assignments based on the empirical Bayesian posterior distribution (Table 3.28).

Based on the empirical Bayes method (column 3, 4 and 5), summing the probabilities of belonging to the each class over all programs, 15.7 programs belong to the second class, 3.6 to the first and 3.8 to the third. The empirical Bayes modal method assigns 17 programs to the second class and 3 programs to the other classes. The most "uncertain" program is *filosofia* (Faculty of Letters) that has a probability of belonging to the second class equal to only 0.43 and a probability of 0.27 to belong to the "worst" class; *ingegneria informatica* (Faculty of Engineering) has a probability of belonging to the third class equal to 0.62 and a probability of 0.35 to belong to the second class, as the other programs of the Faculty Engineering. Probably these programs have particular features that should be analysed separately.

[11]The closer the values of the R-squared statistics are to 1 the better the predictions are.

Table 3.28: Two-level mixture factor model: classification table. Students graduated (old system degree) at the University of Florence, summer session year 2004.

	Modal			
Probabilistic	Class 1	Class 2	Class 3	Total
Class 1	3.0	0.5	0.0	3.6
Class 2	0.0	15.3	0.4	15.7
Class 3	0.0	1.1	2.6	3.8
Total	3.0	17.0	3.0	23

Analysing these results, the university can evaluate which are the “best” study programs relative to the job satisfaction one year after the degree and it can trigger a system of actions and counteractions aimed at improving the general quality of its activities relatively to the labour market requests.

Concluding remarks

During recent years there has been an increase in the size of the statistical literature on measuring the performance of public sector institutions (Bratti et al., 2004). This is linked to the need to efficiently allocate scarce public resources to public institutions, for example in the context of education and health, increasing the emphasis of public policy on institutional auditing and surveillance.

In this thesis we focus on the evaluation of the effectiveness of the university system, with the ultimate aim of providing policy advice to the university. In particular, we evaluate the university effectiveness from the users' point of view with two analyses: we study the perceived quality of students on the global university experience at the completion of the degree to evaluate the internal effectiveness of the university and we analyse the job satisfaction of individuals one year after graduation as an indicator of the university external effectiveness.

In the thesis, we implement multilevel mixture factor models that are included in the generalized latent variable modeling framework. This framework has been developed in recent years and tries to unify and extend latent variable models, integrating specific methodologies with different traditions and application fields in a global theoretical context. While most literature focus on models with a single type of latent variables (all continuous or all categorical), this thesis describes a very general framework proper for models with both continuous *and* categorical latent variables. In particular, the model used in the applications combine the features of factor models and latent class models in the multilevel framework. In the thesis, the term *factor analysis* is used to refer to models with all continuous latent variables and the term *mixture factor analysis* is used to refer to models with both continuous and categorical latent variables, regardless of the nature of the observed variables.

While Chiandotto et al. (2006) implement a structural equation model to analyse both internal and external effectiveness at the same time, we separate the analyses, focusing on a specific part of the measurement model.

We explain the correlation among observed random variables in terms of fewer unobserved random variables, called common *factors*. In our research they represent some hypothetical constructs, the global satisfaction and/or satisfaction on some specific features of the university. In particular, we analyse items on the individuals' perceived quality expressed on the same scale in order to stress on the interpretation of the model.

Actually, when we speak about university effectiveness we refer to the effectiveness of the study programs, relative to each other, since each program has its own features and organization. Students attending the same study program share common environments, experiences, and interactions that can influence their perceived quality (internal or external) of the university. Through the multilevel models we analyse at the same time the phenomenon both at the individual level and the group level. The focus of the thesis is especially on the group (program) level of the hierarchy and the aim is to show the differences and similarities among study programs.

Data used for the analysis come from two surveys of the consortium AlmaLaurea, which currently includes 51 Italian universities. Because of our knowledge of the context, we focus on data about the university of Florence.

In Chapter 1 we show the theoretical framework of our analysis: the generalized latent variable modeling framework. After describing the literature on the topic, the multilevel mixture factor model is introduced as a specification of the generalized latent variable model. In particular, we describe a very general framework proper for models with both continuous *and* categorical latent variables and we analyse the difference and similarities of models with different types of latent variables, taking into account the model formalisation, the likelihood expression, the estimation process and the posterior analysis.

In Chapter 2 and Chapter 3 the empirical part of the thesis is described.

In Chapter 2, after briefly describing the data, some traditional analyses are used. In both case studies, we first investigate the item distributions and the correlation between the items relating to specific aspects of the satisfaction.

For the evaluation of the internal effectiveness, we use the item on the global satisfaction and other 10 items concerning the students' opinion on the university "personnel" (such as professors, technical staff, students, etc.) and on specific services offered directly by the university (such as lecture rooms, library, laboratories, computers, etc.). Since all items are expressed on an ordinal scale, we analyse the polychoric correlations.

There are two groups of items where the correlations are quite high: the first group contains the 4 items about the satisfaction of students on their relationship with the university personnel, the second group is composed by other 4 items, with information about students' satisfaction on the university physical services. Global satisfaction is positively correlated with all other items; it has the highest correlation with the satisfaction on *Relationship with professors* and *Relationship with professors' assistants* and the lowest with *Computers* and *Individual spaces*.

The two items *Relationship with supervisor* and *Relationship with students* have an anomalous behaviour: they have a quite low correlation with all the other items. This peculiarity and the strongly asymmetric frequency distribution different from the other item distributions lead to not include these two variables in the subsequent analyses. On the contrary, the item *Relationship with technical staff* has a strong correlation both with the items relating to the personnel, and with the items *Lecture rooms*, *Laboratories* and *Library*: the perceived quality on the physical services depends also on the quality of some secondary services offered by the university, such as the competence and organization of the technical staff.

We next apply a multilevel regression model to global satisfaction using as covariates the students responses to the items on specifical aspects of the satisfaction. The first level units are the students, the second level units are the programs attended by the students. The aim is both to analyse the effect of each item on global satisfaction and to "quantify" the program effect and so the homogeneity between students attending the same program.

The satisfaction on *Computers*, *Library* and *Individual spaces* does not have a globally significant effect on global satisfaction at the student level, while the covariate with the highest effect on global satisfaction is *Relationship with professors*. All the estimates that are statistically significant have a positive effect on global satisfaction: the more satisfied a student is on specifical aspects of the university system, the more globally satisfied he is.

Both traditional analyses suggest that the university should stress more on the characteristics of the personnel in order to have more satisfied students, for example paying more attention on the recruitment of the professors or giving them the opportunity to participate to "refresher" courses, and training periodically the technical staff. Furthermore the university should improve also the quality of the lecture rooms and the laboratories. With the multilevel regression model it is also possible to quantify the "program effect" on the phenomenon of satisfaction; multilevel techniques are necessary to correctly interpret the phenomenon.

For the evaluation of the external effectiveness we analyse the correlations between 15 items, concerning the graduates' opinion on some job features, such as steadiness, stability, coherence with studies, prestige, relationship with colleagues, salary, etc.

All items are expressed on an ordinal scale with 10 categories, with 1 and 10 representing, respectively, the lowest and the highest satisfaction. The response frequency distributions are quite similar for all items; the most frequent scores are usually 7 and 8 and the score 1 has an higher percentage than the score 2.

The graduates' responses to the items on *Coherence with studies*, job *Steadiness* and *Carrier* reveal some critical features of the university system. Indeed, one year after the degree a high percentage of working graduates consider their job not stable and with low carrier opportunities; furthermore, the job is often not coherent with the studies of graduates. This may be naturally linked to the necessity of some graduates to continue their educational and training programs in order to get the job they studied for and the connected necessity to do occasional jobs. This indicates that the university is not able to meet the market requests, at least in the short period.

The highest correlation is between the items *Salary* and *Carrier*, followed by the correlation between *Carrier* and *Prestige*, *Prestige* and *Professionalism*. The item *Prestige* is correlated with almost all other items. The highest correlations with *Global satisfaction* are for the items *Coherence with studies*, *Professionalism*, *Prestige*, *Cultural Interests*, *Involvement* in working activity and in decisional processes, *Salary* and *Carrier*. On the contrary, the item *Free time* is not correlated with any other item except with *Schedule flexibility*: the graduates' opinion on the free time is not associated with the other aspects of job satisfaction. For this reason, the item *Free time* is not included in the subsequent analysis.

The correlations show a complex structure of the observed association among the items due to the complexity of the job satisfaction phenomenon.

In Chapter 3 we use more proper tools of analysis for the study of both the internal effectiveness of the university and its external effectiveness.

The satisfaction phenomenon is a complex theoretical construct that depends on several latent aspects measured with multiple indicators, so we use tools of analysis relating to the study of latent variables. Furthermore, available data have a hierarchical structure: first level units are the students and second level units are the study programs they attended. As a result, we propose different specifications of the *multilevel mixture factor model*.

For the implementation of the models, we use the syntax module (*Beta version* at 1st of April 2007) of Latent GOLD software, version 4.5 (Vermunt and Magidson, 2007) that allows defining models with any combination of categorical and continuous latent variables at each level of the hierarchy.

In the analysis of the university internal effectiveness, we first use a *multilevel factor model*. With continuous latent variables at both levels of the analysis we study the latent constructs underlying the phenomenon of students' satisfaction both at the student and program level. The final model has different structures at within and between level: at the individual level, there are two latent dimensions underlying global satisfaction (*Human environment* and *Physical environment*), while at the program level the latent dimension is only one (*Physical environment*). The mean opinion of students differ respect to the "physical" features of the programs, while it offset relating to the personnel. At students level, the only item measuring both the satisfaction with *Human environment* and *Physical environment* is relative to the students' relationship with technical staff. In order to improve the mean level of students' satisfaction, the programs should improve the quality of the lecture rooms, the computers, the laboratories and the library. They should also pay attention on the recruitment of the technical staff, that is the only important aspect in defining the satisfaction at program level among the items relative to the university personnel.

Finally, the analysis of the communalities suggests that more latent dimensions that are not included in the model (such as the satisfaction with the program contents) are probably necessary to better explain the variability of the global satisfaction: the questionnaire should include other items measuring these latent dimensions.

Compared to the regression model, the multilevel mixture factor model correctly interpret the items on the students' satisfaction as measures of some latent factors and gives us more detailed information on the phenomenon both at the student and program level; that is, we are able to determine which aspects students are more satisfied with and, connected to this which aspects, determine the mean level of satisfaction with a program.

With the empirical Bayes prediction, we rank the 38 programs on the continuous factor *Physical environment*. The programs with the highest students' satisfaction belong to the scientific area: the first is *statistica*, followed by 3 programs of the Faculty of Science. On the contrary, at the end of the ranking there are programs of the Faculties of Medicine, of Science and of Letters and Philosophy.

Next, we apply to the same data the *multilevel mixture factor model*. At the individual level we retain the latent structure of the multilevel factor model, while at the program level we use a categorical latent variable instead of a continuous variable following the traditional approach to latent class analysis. The aim is to classify the second level units (programs) into a small number of classes, which differ with respect to the item intercepts of the specified factor analysis model. The use of the multilevel mixture factor model to obtain clusters of second level unit is quite innovative. Indeed, the latent class approach is well known in the one-level framework and is usually used to obtain clusters of individual units with the same profile (Hagenaars and McCutcheon, 2002). As results, we find 5 classes of programs, differing in their features and their size.

The first class of programs is the biggest, with 16 programs belonging to different Faculties. The third class is the second biggest and has the "best" features: a program belonging to the third class (mostly of the Faculties of Economics and of Science) has a satisfaction level measured by each indicator higher than the satisfaction level of programs belonging to the other classes. On the contrary, the fifth class (programs of the Faculty of Letters and Philosophy) is the "worst": all coefficients indicate a low satisfaction level, except the parameter relative to the *Individual spaces*. Class 4 is a kind of "average class": the coefficients are quite near to 0 except for the items related to *Relationship with professors* and *Lecture rooms*. In class 2 there is a general satisfaction about the *Human environment* and dissatisfaction relating to the *Physical environment*; in this class there are only two programs of the Faculty of Medicine.

At the end of the two analyses, we merge the results relative to the program level to better understand the phenomenon. The use of different types of latent variables in an unique model is an interesting approach, already applied only in one-level context (Muthén, 2001); it allows obtaining useful information, not available with the separate use of categorical or continuous latent variables.

With the multilevel factor model we rank the 38 programs along a continuum (*Physical environment* satisfaction) and with the multilevel mixture factor model the programs are classified in five classes. Merging the results we distinguish programs that have a similar ranking but belong to different classes and we distinguish programs that are in the same class but have different rankings. For example, the two programs *sc. storiche* and *lettere*, both of the Faculty of Letters and Philosophy, have a similar ranking but they belong, respectively, to class 1 and class 5. These two classes have a similar global satisfaction level and a similar satisfaction with the relationship with professors and professors' assistant, but they are characterized by very different satisfaction level on lecture room and laboratories. Assuming that students of these programs probably do not use a lot the laboratories, *lettere* should improve especially the lecture rooms. At the same time, it is possible to see which are the best programs inside each class; for example in class 3 *scienze statistiche* is much better than *sc. sociali per la cooperazione*.

With this analysis the university obtains a huge quantity of information; obviously, only the substantive knowledge of the phenomenon and of the context let to properly use these information in order to act corrective interventions.

In the analysis of the university external effectiveness from the graduates' point of view, we use a different specification of the *multilevel mixture factor* *model* respect to the first case-study. The aim is to see if the programs (or groups of programs) differ in the values of the latent variables at the individual level representing the job satisfaction. In particular, with continuous latent variables at the individual level we reduce the dimensionality of the phenomenon and with one categorical variable at program level we classify programs relatively to the obtained latent dimensions.

At the individual level, there are 4 latent continuous factors referring, respectively, to the satisfaction with *Earning* and stability, *Cultural* aspects, job *Environment* and *Autonomy*; the factors are strongly correlated. The only not significant coefficient is the loading of the global job satisfaction on the latent factor *Autonomy*, while the most important aspect determining the job satisfaction is the factor relating to the satisfaction on *Cultural* aspects. The factor *Earning* and stability has a significant but not high effect on the job satisfaction; probably this is due to the fact that the analysis concerns only graduates working one year after the degree. Usually, the first job is considered as a first experience that has to provide professionalism and experience, but it has a limited duration over the time; so, the autonomy and the earning are not so important in defining the global job satisfaction.

The model classifies second level units (programs) into three classes with homogeneous factor structures: 68% of the programs belongs to class 2, 16% to class 1, and 15% to class 3.

The three classes differ only for the mean level of the two latent factors *Earning* and *Cultural*. Class 1 of programs is the "worst": the mean level of the latent factors *Earning* and *Cultural* is really low respect to the average class. Class 3 has a mean level of the satisfaction on *Earning* and stability, while the level of satisfaction on *Cultural* aspects is quite high. Class 2 has a contrary behavior of class 3: the satisfaction on *Earning* and stability is really high, while the satisfaction on *Cultural* aspects is not so high.

With the empirical Bayes modal prediction, programs belonging to the first and third class are 3. Programs of the first class (the "worst") belong to the Faculties of Letters and Psychology: usually, the students who graduated in these programs need more than one year to find the job they studied for. In the third class there are two programs of the Faculty of Education and one program of Engineering. The other class is the biggest, with 17 programs.

Analysing these results, the university can evaluate which are the "best" study programs relative to the job satisfaction one year after the degree and it can trigger a system of actions and counteractions aimed at improving the general quality of its activities relatively to the labour market requests. For the university it is really interesting to note that the satisfaction with *Cultural* aspects of the job (measured also by the *Coherence with studies*) is an important factor in determining the global job satisfaction.

The two applications have different features. From a theoretical point of view, different specifications of the multilevel mixture factor models have been used. To analyse the university internal effectiveness we assume that groups of programs differ relatively to the item intercepts of the factor model at the individual level, while to analyse the university internal effectiveness we assume that groups of programs differ in the mean level of each latent factor at individual level. Obviously, these two specifications derives from the data characteristics and from the different aims of the research. From an applied point of view, the first application provides the univesity with information on the aspects that students consider more important and, as a result, with information on the services that must be improved, the second application provides some information that university can use indirectly to increase the programs quality in order to encounter the labour market requests.

In this dissertation, highlighting the interpretational features of the models, we show the extreme flexibility of the generalized latent variable modeling framework. In particular, we describe differences and similarities between the use of continuous and categorical latent variables in factor models with respect to the likelihood formalisation, the estimation process and the posterior analysis, highlighting the information that can be obtained only with the combined use of different typologies of latent variables. In this context, we show for the first time how to combine the results of analyses with continuous and categorical latent variables at the highest level of the hierarchy in order to better explain a phenomenon. At the same time, we obtain information to evaluate the internal and external effectiveness of the University of Florence from the users' point of view.

The multilevel mixture factor models are extremely flexible and supply at the same time a huge quantity of information on different levels of analysis. Furthermore, the estimation process, even if it depends on many factors such as the number of latent variables, is quite fast and this technique can be used in the analysis of many social phenomena. On the other hand, the correct interpretation of these models requires some care, and the literature has tended to focus mostly on theoretical aspects.

In the thesis, we do not use covariates in our models and we measure the students' satisfaction as it is experienced in the real world. Indeed, our main aim is to provide police advice for university and it is difficult for university to act in different ways depending on students' characteristics (covariates at first level) or study programs' characteristics (covariates at second level). In the future, it would be interesting to add covariates in the models in order to evaluate the "net" effectiveness of the study programs, controlling for their composition and their features. From an applied point of view, the use of covariates will lead to a better knowledge of the phenomenon of satisfaction and, as a result, will lead the university to focus its economical and political resources on particular aims. For instance if, controlling for the other features, females are more satisfied than males on the computers, the programs with a majority of males should pay more efforts in obtaining the same global results than the programs with a majority of females. From a theoretical point of view the general modeling framework presented in the thesis, together with likelihood formalisation, estimation process and posterior analysis, should be extended in order to include the covariates in the models.

In the analysis of the external effectiveness, we focus on data relative to students who graduated with the old university system and we study only data on students who graduated one year before the interview. A possible field of research is the analysis of data of students enrolled and graduated with the new university system in order to evaluate the success of the Italian university reform. Furthermore, it will be interesting to analyse the time effect on the job satisfaction using all information collected by AlmaLaurea on students who graduated 1, 3 and 5 years previously. The aim will be to understand if the time has a significant effect on the latent structure at the individual level, so if the relative importance of the specific aspects and the programs effectiveness change over the time. The analysis could be carried on with a MIMIC (Multiple Indicator and Multiple Cause) model (Kline, 2005).

In order to compare models with different number of latent classes we use the index BIC, suggested by several textbooks and articles. In particular, since we implement models with a categorical latent variable only at the program level, the number of observations in the BIC formula is the number of programs. In the literature there is no common acceptance of the best criteria for determining the number of classes of categorical latent variables: this represents another possible field of research.

This thesis describes a very general framework proper for models with both continuous and categorical latent variables. Because of the aims of the applications, we use models with different types of latent variables at each level of the analysis. Another possible field of research is the modeling using a combination of continuous and categorical latent variables at the same level of the hierarchy; this extends the latent class approach by allowing for variability of the phenomenon under study within classes.

In this context, an open research problem is also the issue of identification and outliers. These topics have been not handled in the thesis. In particular, the topic of identification did not receive the proper attention in the literature, especially in the multilevel framework.

Finally, the multilevel mixture models have been implemented using of the

Latent GOLD software, but other packages can do the work: a comparison of competing software would be an important value for applied research.

Bibliography

Agresti, A. (2002), *Categorical Data Analysis, 2nd Edition*, Wiley and Sons, Inc., Hoboken, New Jersey.

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

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

Bartholomew, J. D. and Tzamurani, P. (1999), 'The goodness of fit of latent trait models in attitude measurement', *Sociological Methods Research* **27**, 525–546.

Bassi, F. (2007), 'Latent class factor models for market segmentation: an application to pharmaceuticals', *Statistical Methods and Applications* **16**, 279–287.

Bock, R. D., Gibbons, R. D. and Muraki, E. (1988), 'Full-information item factor analysis', *Applied Psychological Measurement* **12**, 261–280.

Bollen, K. A. (1989), *Structural equations with latent variables*, J. Wiley, New York.

Bratti, M., McKnight, A., Naylor, R. and Smith, J. (2004), 'Higher education outcomes, graduate employment and university performance indicators', *Journal of the Royal Statistical Society Serie A* **167(3)**, 475–496.

Cagnone, S. (2003), Latent variable models for ordinal variables, Phd thesis, University of Bologna.

Chiandotto, B. (2004), 'Sulla misura della qualità della formazione universitaria', *Studi e note di Economia* **3**, 27–61.

Chiandotto, B. and Bacci, S. (2004), Un modello multilivello per l'analisi della condizione occupazionale dei laureati dell'ateneo fiorentino, *in* C. Crocetta, ed., 'Modelli di statistici per l'analisi della transizione Universita-lavoro', Cleup, Padova, pp. 211–234.

Chiandotto, B. and Bacci, S. (2006*a*), Measurement of university external effectiveness based on the use of the acquired skills, *in* L. Fabbris, ed., 'Effectiveness of University Education in Italy. Employability, Competences, Human Capital', Physica-Verlag HD, pp. 89–104.

Chiandotto, B. and Bacci, S. (2006*b*), Un indicatore sintetico per la misura dell'efficacia interna della formazione universitaria, *in* C. Crocetta, ed., 'Metodi e modelli per la valutazione del sistema universitario. OUTCOMES', Cleup, Padova, pp. 85–105.

Chiandotto, B., Bini, M. and Bertaccini, B. (2006), Quality assessment of the university educational process: an application of the ECSI model, *in* L. Fabbris, ed., 'Effectiveness of the University Education in Italy. Employability, Competences, Human Capital', Springer-Verlag HD, pp. 43–54.

Dobson, A. J. (2002), *An introduction to generalized linear models*, Chapman & Hall/CRC, Boca Raton, Florida.

Eskildsen, J. K., Martensen, A., Grønholdt, L. and Kristensen, K. (2000), 'Benchmarking student satisfaction in higher education based on the ecsi methodology', *Sinergie - Rapporti di Ricerca* **9 (18)**, 385–402.

Everitt, B. S. (1988), 'A monte carlo investigation of the likelihood ratio test for number of classes in latent class analysis', *Multivariate Behavioral Research* **23**, 531–538.

Fielding, A. (2004), 'Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models', *Quality & Quantity* **38**, 425–433.

Galindo-Garre, F. and Vermunt, J. K. (2005), 'Testing log-linear models with inequality constraints: a comparison of asymptotic, bootstrap, and posterior predictive p-values', *Metodología de las Ciencias del Comportamiento* **4**, 325–337.

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

Goodman, L. A. (1974), 'Exploratory latent structure analysis using both identifiable and unidentifiable models', *Biometrika* **61**, 215–231.

Grilli, L. and Rampichini, C. (2007*a*), 'Multilevel factor models for ordinal variables', *Structural Equation Modeling* **14(1)**, 1–25.

Grilli, L. and Rampichini, C. (2007*b*), 'A multilevel multinomial logit model for the analysis of graduates' skills', *Statistical Methods and Applications* **16(3)**, 381–393.

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

Harvey, L. and Knight, P. T. (1996), *Transforming Higher Education*, Open University Press, Taylor and Francis, Bristol, UK.

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

Hox, J. J. (1995), *Applied multilevel analysis*, TT-Publikaties, Amsterdam.

Jöreskog, K. and Moustaki, I. (2001), 'Factor analysis of ordinal variables: A comparison of three approaches', *Multivariate Behavioral Research* **36**, 347–387.

Kaplan, D. (2000), *Structural equation modeling: foundations and extensions*, Sage Publications, Thousand Oaks, London, New Delhi.

Kline, R. B. (2005), *Principles and practice of structural equation modeling*, Guilford Press, New York.

Langeheine, R., Pannekoek, J. and Van de Pol, F. (1996), 'Bootstrapping goodness-of-fit measures in categorical data analysis', *Sociological Methods and Research* **24**, 492–516.

Lo, Y., Mendell, N. R. and Rubin, D. B. (2001), 'Testing the number of components in a normal mixture', *Biometrika* **88**, 767–778.

Lockheed, M. E. and Hanushek, E. R. (1994), 'Concepts of educational efficiency and effectiveness', *International Encyclopedia of Education* pp. 1779–1784.

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

Magidson, J. and Vermunt, J. K. (2001), 'Latent class factor and cluster models, bi-plots and related graphical displays', *Sociological Methodology* **31**, 223–264.

Magidson, J. and Vermunt, J. K. (2002), 'Latent class models for clustering: A comparison with k-means', *Canadian Journal of Marketing Research* **20**, 36–43.

Magidson, J. and Vermunt, J. K. (2004), Latent class models, *in* D. Kaplan, ed., 'The Sage Handbook of Quantitative Methodology for the Social Sciences', Sage Publications, Thousand Oakes, pp. 175–198.

Martensen, A., Grønholdt, L., Eskildsen, J. K. and K., E. J. (2000), 'Measuring student oriented quality in higher education: Application of the ecsi methodology', *Sinergie - Rapporti di Ricerca* **9 (18)**, 371–383.

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

Millsap, R. E. (2001), 'When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis', *STRUCTURAL EQUATION MODELING* **8(1)**, 1–17.

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

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

Muthén, B. O. (2004), 'Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data', *Handbook of quantitative methodology for the social science* pp. 345–368.

Muthén, B. O. and Kaplan, D. (1985), 'A comparison of some methodologies for the factor analysis of non-normal likert variables', *British Journal of Mathematical and Statistical Psychology* **38**, 171–189.

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

Nylund, K. L., Muthén, B. O. and Asparouhov, T. (2007), 'Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study', *Structural Equation Modeling: A Multidisciplinary Journal* **14**, 535–569.

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

Pawitan, Y. (2001), *In all likelihood: statistical modelling and inference using likelihood*, Clarendon, Oxford.

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

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

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

Self, S. G. and Liang, K. Y. (1987), 'Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions', *Journal of the American Statistical Association* **82**(398), 605–610.

Skrondal, A. and Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman and Hall/CRC, Boca Raton, FL.

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

Spearman, C. (1904), 'General intelligence, objectively determined and measured', *The American Journal of Psychology* **15**, 201–292.

Stroud, A. H. and Sechrest, D. (1966), *Gaussian Quadrature Formulas*, Prentice Hall, Englewood Cliffs, New Jersey.

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

Vermunt, J. K. (2005), 'Mixed effects logistic regression models fo indirectly observed discrete outcome variables', *Multivariate behavioral research* **40(3)**, 281–301.

Vermunt, J. K. (2007), 'Multilevel mixture item response theory models: an application in education testing', *ISI 2007 Proceedings* .

Vermunt, J. K. and Magidson, J. (2003), 'Latent class models for classification', *Computational Statistics and Data Analysis* **41**, 531–537.

Vermunt, J. K. and Magidson, J. (2005*a*), Factor analysis with categorical indicators: A comparison between traditional and latent class approaches, *in* A. Van der Ark, M. Croon and K. Sijtsma, eds, 'New Developments in Categorical Data Analysis for the Social and Behavioral Sciences', Erlbaum, Mahwah, pp. 41–62.

Vermunt, J. K. and Magidson, J. (2005*b*), *LatentGOLD 4.0 user's guide*, Statistical Innovations Inc., Belmont, Massachussetts.

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

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.

Vermunt, J. K. and Van Dijk, L. (2001), 'A nonparametric random-coefficients approach: the latent class regression model', *Multilevel Modelling Newsletter* **13**, 6–13.