Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Latent growth modelling: Analysis of the change in students' satisfaction for the University of Padua

Relatore: Prof. Francesca Bassi
Dipartimento di Scienze Statistiche

Correlatore: Prof. José G. Dias
Departamento de Métodos Quantitativos, ISCTE Business School

Laureando: Marco Guerra
Matricola N. 1084521

Anno Accademico 2015/2016

Contents

List of Figures

List of Tables

Introduction

This thesis work finds its main aim in analysing the change in students' satisfaction with the courses within the University of Padua. We know from the literature on this topic that students' satisfaction has gained more and more importance over the last years. In fact, students' opinions, gathered through a questionnaire, are fundamental to detect any bad aspect of the teaching process. They allow then to improve the same teaching process and also the learning one, as a consequence. This should lead even to an increase of satisfaction with the course. Most of studies involving questionnaires of didactics are focused on the study of one single year. That allows a deep understanding of the "static" level of students' satisfaction, missing an overall view of the "dynamic" process in which satisfaction is involved, though. This work is meant to cross that line. It will be interesting to analyse a usual set of data from a relatively new and more complete point of view.

To achieve our objective, data was collected from students' responses to the questionnaire of didactics. Students are asked to answer such a questionnaire at the end of every course they attended, before taking the exam. It is composed of many different items, each catching a particular aspect of the didactic activity plus a direct question on the overall satisfaction with the course. Our analysis will be focused on the last three academic years available, from 2012/13 to 2014/15. To conduct a longitudinal analysis like this, only courses available throughout the considered period will be taken into account. Having experienced a change in the items since academic year 2013/14, the questionnaire is not entirely comparable over time. For this reason, only a subset of the items will be considered, namely the ones available in all years analysed.

A first approach to the data will be made in order to assess the latent structure underlying the measurement scale, i.e. the set of items at our disposal. It will be accomplished by means of a Factor Analysis. Results will be rather different from the ones obtained in previous researches on the same kind of data, as we will show later on. This analysis will be made cross-sectionally, considering only one academic

year at a time. It will allow a preliminary knowledge of the data collected and will provide an initial idea of the satisfaction level within our university institution. This is thus a good starting point for more complex and meaningful analyses on the data, always keeping in mind our primary aim.
The final analysis we will show represents a longitudinal study on the data. Hence, we will simultaneously analyse the items value for each course in each year. This will be made using different kinds of latent growth models. The results will provide the growth path(s) of satisfaction and will allow us to make many final considerations.

The present work is structured in five chapters.
Chapter 1 introduces the topic of the thesis and it is mostly a work of literature review. It explains fundamental concepts such as quality of didactics and students' satisfaction. Moreover, it stresses the importance of evaluating teaching activity through questionnaires given to students. An excursus is finally made on the context in which the University of Padua is set.
Chapter 2 reports the theory of our preliminary analysis on the data. In particular, Factor Analysis is explained, together with many indices to assess the reliability of a measurement scale and the goodness of the factorial model estimated.
Chapter 3 provides the theoretical concepts of latent growth modelling. It describes three different growth models, the ones estimated in the analytical part of the work.
Chapter 4 shows at first the results of factor analysis using the full measurement scale available on the last academic year. Subsequently, two reduced scales are presented and are used to perform the same analysis, always on the last year. After having chosen the model with the best fit, Factor Analysis is conducted on the three years separately.
Chapter 5 displays the last analysis on the data, the one of more interest. Different models are presented, some of them including course- and teacher-related covariates. These variables are likely to affect the growth trajectory of satisfaction. The results of the preferred model and the related conclusions are reported at the end.

Chapter 1

Quality in didactics and students' satisfaction

The concepts of quality in didactics and of students' satisfaction, which are in fact really close ideas, have both acquired more and more importance over time. In particular in these last years, it has become clearer how much an improvement in the services provided by the university can affect in a good way the same image of the educational institution. It would bring more prestige and for this reason it is likely that more people would choose university to continue their course of study on the basis of it.

To reach the achievement of improvement, at the end of every course and before the final exams, all the students are asked to fill in a questionnaire regarding some aspects of the classes they just attended. Through these questions the students can say, anonymously, the characteristics of the course they found appreciable and the ones they did not. Their answers allow the management of university to know the degree of students' satisfaction and how the quality of didactics could be improved. It is important to understand which aspects of teaching are actually not working and are probably affected by "bad quality", because it is reasonable to think that satisfaction is strictly related to the teaching method. Thus, satisfaction could be raised if professors are made aware of what they could improve in their lessons. For these and other reasons that will be explained better later, the importance of students' opinion for an institution such as a university is clear. But before dealing with this fundamental aspect, it is better to figure out the various meanings that the concept of quality could assume.

1.1 The concept of quality

According to Harvey (2006), the term *quality* embodies “the essential nature of a person, object, action, process or organization” and it is referring often to concepts such as distinctiveness, exclusivity, goodness or reliability. There are at least five ways to define the concept of quality:

- **Excellence**: quality is seen as something special that has overtaken high standards. If these standards are set at an even higher level, the quality will be improved. This acceptation of quality implies exclusivity. The excellence in higher education can affect both input (teaching process) and output (the students’ learning capability).

- **Perfection or Consistency**: in general terms, a quality product is a consistent one and with no defects. This notion deals with the idea of reliability and, unlike the first one, it allows all products to be potentially quality products. University’s task is to ensure a reliable and “zero defects” information system, in addition to a consistent service provided.

- **Fitness for purpose**: in this case, a quality product is one fitting perfectly the purpose it was meant for. As the concept of perfection, it is inclusive and so all products and services could be quality ones. But unlike the previous one, it emphasises the need to reach some generally accepted standards. According to this concept, a product or service can be considered a quality one if the requirements and needs of consumers are met. But, otherwise, it can consider even the purpose of the institution which is providing the product/service: for instance, a university could consider a quality service the one that meets its own objectives.

- **Value for money**: quality is in this case seen as a “return on investment”. A quality product or service is the one which can be obtained at the lowest cost or at a predetermined cost that the costumer considers acceptable. The value for money quality can be related to the terms of efficiency and effectiveness, even in the field of education. There is efficiency if some educational activities are provided using a lower amount of money, if a course is replaced by one considered more useful or if useless activities are just eliminated. An institution could be efficient, but not effective if it does not achieve to meet its purposes.

- **Transformation**: according to this acceptation, there is quality when a "qualitative change" occurs from one state to another. In higher education, transformation is referring both to the improvement and change a student can experience through his learning process and to institution internal changes that allow it to provide improved learning processes to students.

So it is clear how all these ways to see the concept of quality can be applied to the higher education field. In fact, according to Fabbris (2002), the supply of a didactic service is a multidimensional process, which involves a lot of resources and actors from different levels of the education hierarchical scale (professor, student, department, school, athenaeum). The process implies lasting interactions among these levels and each of them has a different perception of what quality means. So the concept of quality is quite relative because it depends on the personal perspective of the people linked to the university, the ones that are interested in having quality educational services.

It is necessary to introduce the term of *customer* of an university institution. It can be said that university customers are all the subjects obtaining some benefits through the services the university offers to them. So students from different studies (bachelor, master or even PhD in all the fields), private or public authorities that commission some surveys and even the social and economic systems are all customers of the university, of course in really different ways. This fact makes it easier to understand how several criteria are used to define quality, in relation to the stakeholder taken into consideration. For instance, students see university quality in terms of excellence: choosing a well-known high education institution (an excellent one) means to have a good educational and personal formation, besides a higher probability of finding a job once taken the degree. Instead, external institutions supporting university research could see quality mainly in a financial way, looking at the investments they made on the surveys. Moreover, professors see quality as perfection, having complete working satisfaction (Srikanthan and Dalrymple, 2003).

As it has been previously said, one of the acceptations of quality concerns transformation and so it is strictly related to the concept of change: that is what the evaluation process should do, because it should bring to a transformation and possibly an improvement of the evaluated person. This particular concept brings to the definition of *dynamic quality*, as it was originally thought by Pirsig in his work of 1991. The dynamic quality, seen as a "sparkle" generating a change, is not an easy subject to be dealt with because it goes beyond a precise definition. It is in contrast with the *static quality*, which can be exactly defined. In the educational context, it is determined by elements such as a clear and complete teaching program or the definition of consistent aims and contents: these are the principal characteristics defining static quality in university didactics. In the same field, dynamic quality could be expressed through the involvement and love for learning (Dalla Zuanna et al., 2015). Although these two kinds of quality are totally juxtaposed, they have to coexist in order to allow a good learning process: as Pirsig (1991) stated, "without dynamic quality the organism cannot grow. Without static quality the organism cannot last. Both are needed".

It is clear that the concept of quality has its own complexity, incorporating very different meanings in itself. It strongly depends on the context and on the actors taken into account: as it has been shown, even within the field of high education didactics, there are multiple dimensions of quality with dynamic nature.

1.2 The satisfaction of university students

The concept of quality is strictly linked to the one of satisfaction. University is an institution providing, in fact, some educational services to many people, the students: certainly it can be said they really are the customers of the university. For this reason, if students feel satisfied or not it depends on the consumption process they experience, like for any other kind of product or service. Many references on this topic are present in the literature and five phases are usually outlined in the consumption decision process (Figure 1.1). According to Guido et al. (2010), these phases are:

1. **Needs awareness**: it represents the beginning of a decision process, growing from the existence of a need that has to be satisfied. The client perceives this need comparing his/her actual situation to the desired one. In the academic context, it could be referring to the desire of a person to start studying at the university or, once started this formative experience, the desire of attending a particular class. This need can come, for example, from the will to have more probabilities of finding a successful job in the future.

2. **Information search**: in this phase, useful information is collected and through this, the consumer should be able to make a decision that fulfils the need. In the educational field, this phase represents the searches the students, potential or actual ones, make in order to choose one degree course or a single class

within their degree courses. Obviously the information could come from students previously attending that particular course. In that case, the role of the so-called *word of mouth* becomes really important. The fame of the university plays a great role too.

3. **Evaluation of alternatives**: after the collection of different kinds of information, consumers have to choose among some alternatives. Students are asked to judge which one of the possible choices they prefer.

4. **Purchase decision**: the moment of the purchase decision concerns the real choice of one of the possible options. It implies the purchase of the product and its use. From the students' perspective, this phase deals with the final decision of the course to attend and the act of actual attending.

5. **Post-purchase dispositions**: in this final phase, consumers evaluate the choice made earlier and decide how to behave in the future. It is clear that students cannot decide to attend an identical class or degree twice. Instead, after the attendance, a student can decide to study other subjects strictly connected to the one learnt in that class because it was interesting and useful or, on the contrary, can decide to leave the university because it was not what the student really wanted.

Although the satisfaction of a consumer is often attributed to the sole last phase of this process, it is not completely correct. In every moment of the purchase process, the satisfaction is creating and assuming different shapes depending on the context. Every service provider has to define and measure its costumers' satisfaction and that is what the same university has to do, belonging to that category: an institution of this kind has the goal of making people grow through the learning process. But it must assure these people to be fully satisfied with their course of study, so they will still be enrolled as students and will continue to guarantee an income to the institution.

There are many ways to interpret satisfaction. A general definition could be found in the perception of customers to have obtained the best, in proportion to what they were expecting (Cherubini S., citation contained in Iasevoli, 2007). However, the most common definition of satisfaction is derived from the so-called discrepancy paradigm: satisfaction is given by the comparison between customers' expectations and the perception of the service they actually received (Iasevoli, 2007).

In high education, a student enrolling in a class and having some expectations can

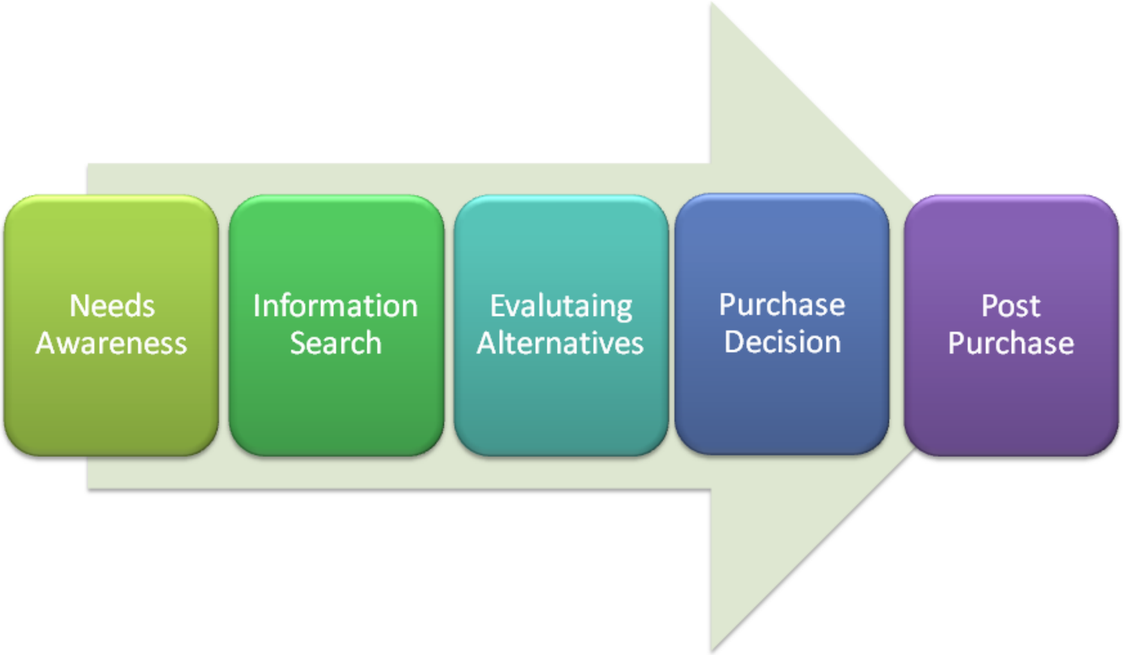

Figure 1.1: The five phases of the consumption process.

find them confirmed or disconfirmed. In this last case, it could be both positively or negatively. If the student thinks the performance received is worse than expectations, it is a negative disconfirmation; on the opposite side, if the performance goes beyond any expectation, there will be a positive disconfirmation; if performance meets perfectly expectations, there is confirmation. If there is negative disconfirmation, feelings such as disappointment can grow among the students. In the case of a positive one, emotions as excitement and happiness are predominant. The confirmation of expectation brings to gratification (Athiyaman, 1997).
The overall satisfaction of a student towards the university is directly linked to his/her opinion about the quality of the course, but even on other kind of factors linked to that university. Students are more likely to recommend their university to their acquaintances if they found a good campus environment, otherwise they will probably not (Elliott and Shin, 2002). However, the concept of satisfaction is different from one person to another. For example, while a student could appreciate a professor interacting more with the students attending the course, another one could feel uncomfortable with it: the level of satisfaction they declare about the same course would be really different.

In the past, students' satisfaction has not been the principal method to evaluate and understand the level of quality within the university courses. Focusing the attention

on the teachers rather than on the entire university organization, it should be easier to achieve an improvement in high education quality (Chen et al., 2014).
In the last years, it has become very common to evaluate the quality of didactics through a questionnaire filled in by the students, in which their opinion about the attended course are collected. The questionnaire could be done using one single question (single-item), but it would be possible to capture only the opinion on the overall satisfaction, or using different questions (multi-item). In the last instance, it is possible to measure both the different dimensions of the service provided and the overall impression about the performance (Elliott and Shin, 2002).
In the next paragraph, some literature about the topic of students evaluating the didactics will be summarised.

1.3 Students' Evaluation of Teaching

The Students' Evaluation of Teaching (SET) is a widespread method to measure teaching performances in high education, on the basis of a questionnaire proposed to students regarding an individual university course. SETs are collected towards the end of the semester, when the course is nearly over, and it seems to be generally accepted that students' opinion is really helpful to enhance the quality of education. But, even if this method is almost used all over the world, it is still the centre of controversy and strong disagreement (Zabaleta, 2007).
So why is it so important to collect students' feedbacks?
It is sure that they are playing an always increasing role through the years, principally for three reasons (Bassi et al., 2016):

1. they show the students' point of view and their level of satisfaction about professors and didactic activities in general;

2. they allow the same professors and the management of the university to be conscious and to reflect about their effective work, so they can be helpful to decision-making about promotion and tenure;

3. they should bring to an increase of the quality of the services offered by the University and to a general improvement of the didactics.

It is not only important to involve students in the evaluation process by means of a questionnaire, but it is also relevant to acquire some information capable to assure an improvement in teaching. As Zabaleta (2007) stated, SETs have the aim to improve both the students' learning process and the professors' teaching one. If some teachers use students' opinion as a stimulus to improve their teaching performance, many others still do not recognise the importance of SETs and their usefulness. Thus, they tend to ignore advice and comments supplied by the students (Spooren et al., 2013).

Through an accurate literature review, Spooren et al. (2013) confirmed that thousands of research studies have been conducted on SETs since 1927, year of publication of the first report dealing with this topic. These studies have been focusing on the validity of students' opinion and on the possibility they could be subject to some bias factors, not necessarily related to quality of teaching. These types of factors can be divided into three groups:

- student-related factors;
- teacher-related factors;
- course-related factors.

Many researchers have been conducting research on this topic: in the next subparagraphs some evidence from the literature will be revised (specifically from Beran and Violato, 2005; Spooren, 2010; Spooren et al., 2013; Dalla Zuanna et al., 2015).

1.3.1 Student-related factors

One of the first variables influencing the SETs is the attitude of the students towards the same SETs: it does influence the SETs, because it would make the difference if the student is doing an effort in answering accurately. If students cannot see an immediate connection between their effort in fulfilling the questionnaire and the effects of their evaluations (about some organizational or teaching aspects), the questionnaire tends to become only something to do: students would lose the will to answer correctly. For this reason, the importance of this kind of survey should always been explicated to students, as well as the impact and the effects that accurate answers could have (Dalla Zuanna et al., 2015).

The attendance to the classes seems to have a sort of effect on students' rating. In particular, the more students attend the course, the more is likely that they will give a high rating to the course. That could be explained by the fact that attending students are usually more motivated and interested in the course than students who

are not attendee, so they tend to give higher ratings (Beran and Violato, 2005).
The dedication of students could be related to SETs and professors encouraging them to study and apply themselves are normally the ones who receive the best feedbacks.
Both expected and actual grades seem to influence the rating on the questionnaire: the higher the grade, the higher the evaluation of the course.
The gender of the student is likely to affect the ratings of the teaching too: as explained by Spooren (2010), male students do not choose often as expected a female professor as their favourite teacher. In addition, female students are giving generally higher grades than male ones.
About student's age, the situation is still not clear. Students of master's degree usually give higher ratings than students attending bachelor's courses. However, it could be both because respondents attending an upper-level course are more mature or because specialised courses are considered more interesting by the students.

1.3.2 Teacher-related factors

The gender of the student is not the only characteristic which could affect evaluation of teaching. It seems that women receive higher ratings, but researches do not always confirm such a result.
Experience as an instructor, charisma, personality, availability and respect of the students are all variables positively associated to evaluations of teaching.
The age of the professor is still an uncertain factor. Most of the studies do not show a significant association between this variable and the SETs, while a few others found a significant negative correlation. It seems that lower evaluations are given to old professors, while young ones are preferred.

1.3.3 Course-related factors

There are also some factors related to the characteristics of the course.
The complexity of the topics is plausible to play an important role, because usually the more difficult the course is, the lower the students' grade and satisfaction.
Students coming from different schools and studies have completely different approaches to the courses and to the university in general. Researches found out that frequently humanities students are more satisfied and tend to be more generous in their evaluations than students studying scientific subjects.
According to previous researches (see Spooren, 2010), the size of the class could

have a negative relation with the SETs, but also a non-linear relation with them. It could be possible for small-sized and big-sized classes to be more appreciated than the medium-sized ones.

The relation between the workload and SETs is still a little mysterious. Some studies reported that the rating increases with the increase of the workload. Students can feel to be challenged by a teacher who is requiring them a huge workload. For this reason they can become more committed to study, learn more and then give a higher evaluation. Other studies concluded exactly the opposite. For instance, feeling the weight and the pressure of an excessive workload, students could be angry and feel resentment towards the teacher, which brings to a low rating.

The type of the course may be a significant variable too. Laboratories and practical lessons are usually more satisfying than theoretical classes, because they allow the students to apply what they have just learnt.

It is clear how many factors not related to didactics quality could affect the opinions of university students. It becomes compulsory to take into account the context in which the students are, because they may be influenced by some variables external to university. But researches by Beran and Violato (2005), Spooren et al. (2013), Dalla Zuanna et al. (2015) found out that the effects on the SETs of all these and other possibly biasing factors are in fact really small. These factors explain only a minimal percentage of the total variance of the SETs rating.

1.4 The case of the University of Padua

The University of Padua is one of the biggest universities in Italy and one of the oldest educational institutions in the world. Formally founded in 1222, it counts more than sixty thousand students.

The following lines of the present section are going to refer to the work of Martinoia and Stocco, contained in the technical report by Dalla Zuanna et al. (2015).

The opinion of the students about the courses they attend is very important for the University of Padua, because through a specific questionnaire it allows to understand if students are satisfied with the didactic activities carried out.

Collecting students' feedback is useful to provide to professors information about the way they are teaching, in order to start a continuous improvement process in the quality of the didactics and in the services connected to the same didactics.

The Athenaeum of Padua has paid attention to students' opinions in particular since the first semester of the academic year 1999/2000, before it became compulsory by law (Law 370/1999). By means of a common instrument, since that moment all the attending students have been required to express their point of view about didactic activities they were enrolled in. Previously there had been other experimentations, which de facto did not involve all the University Faculties. In the year 1999/2000, for the first time, all the thirteen Faculties of the University of Padua, even if with different levels of acceptance, decided to participate in this initiative.

In the first years the questionnaires were done making the students fill in a paper, in which there were both multiple-choice and open questions. The first type allowed the University to have information easy to collect and compare. The second one permitted the students to have the freedom of saying exactly whatever they thought about the courses attended, including negative and positive aspects of the lessons and suggestions to the teacher. The questionnaires were meant to be anonymous and used only at aggregate level, in order to guarantee students' freedom of expression.

The open-question part of the questionnaire was given to the same professor immediately after it was filled in, so he/she could have an instantaneous perception of what students thought about his/her teaching methods and the didactic activity itself.

The survey submitted to the students has been constantly modified over the years, in order to have a continuous improvement of the questionnaire and of the quality of services supplied by the educational institution.

Starting from the academic year 2010/2011, students were asked to answer an on-line questionnaire about didactic activities they attended (on-line survey model CAWI, *Computer Assisted Web Interview*). The questionnaires were proposed through *Uniweb*, the information system of University of Padua allowing all the regular students to access, via Internet, the information about their course of studies and to manage directly their own academic career. Through these computerised questionnaires, students' opinions about single courses and professors of the current academic year are collected, asking them some specific characteristics they have to evaluate and rank. The paper questionnaire is still used and proposed to students during a lesson, usually when the course is almost finished. It is composed of the usual open questions, in which students are free to say what they liked or disliked most about the course and some advice to give directly to the professor to improve the quality perceived of the didactic activity.

1.4.1 Data collection

This research uses a sample of 61,488 questionnaires answered by students. They were filled in during the last three academic years, from 2012/13 to 2014/15. The total number of the university courses evaluated and considered, which are in fact the units of the following analysis, is equal to 1,854. To collect this data, only the answers of students that effectively attended more than the 50% of the course were used. The aim of this thesis is to analyse if students of the University of Padua are generally satisfied with the courses provided by this institution and if their satisfaction has increased or not in the last three years. In the first case it would mean that the quality of didactic activities, or at least the quality perceived by the students, got better over time; otherwise it would mean that something about didactics provided should still be changed and University should find a way to improve teaching and learning processes.

1.4.2 The items

The items composing the evaluation scale which have been considered in this analysis are twelve. The first eleven deal with different aspects of the course or of the teacher, while the last one expresses a judgement on students' overall satisfaction about the classes they attended. In each question, respondents are asked to rate from 1 to 10 their level of satisfaction with a certain aspect of the course, being 1 the lowest level and 10 the highest one.

In the last three academic years some questions of the computerised questionnaire have been changed. These ones have been excluded from the data set to permit the comparison among the years and a longitudinal analysis of the data.

Before starting the questionnaire, three questions are posed to the students. The first one is if they are available to participate to the survey (whether the students are not, the questionnaire has to be considered concluded). The second one asks them the percentage of the classes of the course under examination the students have attended. The third question is about the period in which students have attended the course. If students attended less than a half of the total classes, they have to be considered not attending and a reduced questionnaire is posed to them. This is not our case, because in collecting this data only the questionnaires of actually attending students were taken into account. Even if students have attended the course in a previous academic year to the one in which they are filling in the questionnaire, the number of the questions to be answered is reduced. In our data only the students

who have just attended the courses are considered.
The twelve items proposed to the students are the following ones:

- **Item 1**: At the beginning of the course, were aim and contents clearly shown?
- **Item 2**: Were the examination arrangements clearly explained?
- **Item 3**: Were timetables of the didactic activity observed?
- **Item 4**: Was your preliminary knowledge sufficient to understand all the topics?
- **Item 5**: Independently on how the course was taught, were the contents interesting?
- **Item 6**: Did the teacher stimulate some interest towards the topic?
- **Item 7**: Did the teacher explain the contents clearly?
- **Item 8**: Was the suggested material adequate for study?
- **Item 9**: Was the teacher available during the office hours for further explanations?
- **Item 10**: Were laboratories/practical activities/workshops, if included, adequate?
- **Item 11**: Was the requested workload proportionate to the credits of the course?
- **Item 12**: How much are you satisfied with this course?

Every year, in its website, the University of Padua publishes some information collected with the questionnaire about didactics, in particular three indicators. The first one is represented by the last item described (the overall satisfaction), which is a *gold standard* (DeVellis, 1991). Thus, it can be used to validate the measurement scale of students' satisfaction, because the scale validity is assured when it has a strong association with this gold standard. Item 12 could be seen as an alternative method to measure directly satisfaction, rather than using the scale composed by the other eleven items: this is the reason why it will be not included in the analysis. The other two indicators are related to the organizational aspects of the course and to the efficacy of didactics: they are respectively obtained as the arithmetic mean of items 1 (clearness of the aims), 2 (examination arrangements), 3 (observance of the timetables), 8 (material suggested) and of items 6 (stimulus of interest), 7 (clearness of explanation).

Previous studies confirm that the last two indicators and the measurement scale can be considered valid and reliable to measure the satisfaction of the students of University of Padua (Bassi et al., 2016; Dalla Zuanna et al., 2015). The concept of validity refers to the degree with which the measured variables correspond to the underlying construct. A measure can be considered reliable if it produces stable and consistent results, so if independent but comparable variables explaining the same construct match each other. Reliability is necessary, but not sufficient, to assess validity (Bassi et al., 2016).

In the work by Bassi et al. (2016), the measurement scale is wider than the one here presented, counting eighteen items in total. Twelve of them are exactly the ones used in this thesis. We remind that one of them (the gold standard) is excluded in advance from the study. Using Factor Analysis, the authors found that the seventeen-item scale is multidimensional and that there are four latent factors underlying it. As it will be shown in the following chapters, using only the eleven items specified before, results change significantly. Through a Factor Analysis, there is no way the scale adopted could be more than just unidimensional.

1.4.3 The context variables

The data set has some other variables referring to peculiar traits of the courses or of the professors teaching a particular course. It is likely to expect they can have a real influence on students' satisfaction. The University of Padua does not collect information on the sex of respondents (students) and the professor being assessed, despite the evidence that this variable may explain the appreciation of the course and should be controlled for this reason. The same happens to the age of students and/or professors, that could have been important to control in the study and measures the impact on satisfaction.

The variables which have been object of analysis are:

- **Academic Year**: the year in which the questionnaire was filled in and in which the student attended the course. As said before, it could be 2012/13, 2013/14 or 2014/15.

- **Number of questionnaires filled in**: the number of completed questionnaires per course every year.

- **School**: numerical code (from 1 to 8) associated to the eight schools existing within the educational institution. For privacy reasons, they cannot be labelled.

- **Kind of degree**: variable indicating if the course is taught in a bachelor's degree, master's degree or 5-year-long degree.

- **Borrowed**: dichotomous variable explaining if the course is provided within the same degree course of the student attending or if it is not.

- **ECTS**: number of credits assigned to the course.

- **Hours of didactic activity**: the total number of hours of the didactic activity. It corresponds to the sum of the hours of all the teachers who have taught in that course.

- **Role of the teacher**: it expresses the role the professor has and it could be full professor, associate professor, assistant professor, external partner (a person who normally does not work within the university, but who is asked to collaborate) and others (very uncommon or not available teaching roles).

In the next two chapters of this work, we are going to introduce the main theory concepts about methods and models that we used to study the data set and hence the change in students' satisfaction.

Chapter 2

An introduction to validity and reliability of the measurement scale

Aim of this chapter is to briefly explain the concepts of validity and reliability of a measurement scale, showing the main methods to assess them. It must be said that the measurement scale under study has already been proven to be valid and reliable (see, for instance, Dalla Zuanna et al., 2015; Bassi et al., 2016). Thus, the validation of the questionnaire is not the principal intention of this work, anyway it could be confirmatory of the goodness of the scale adopted by the institution.
In our case, the methods to assess validity and reliability will be of fundamental importance to understand the latent structure underlying the questionnaire of didactics of the University of Padua. Moreover, they can be considered the basis for further and more explanatory analyses on the data.

2.1 The latent variable: validity and reliability of its measures

One of the most important concept in this work is the idea of *latent variable*, namely a variable which is not observable or observed. This could be possibly a manifest variable but hidden or referring to an abstract concept and not effectively representable. Even though it is actually not directly observable, it can be linked to other manifest variables (i.e. directly observable ones) by means of a mathematical model.

Let us consider an example which is meant to be explicative of this situation. Suppose there is a construct C, which can be, e.g., students' satisfaction. Paraphrasing Churchill (1979), every student within the university has a personal "true" level of satisfaction X_T at any given time point. The questionnaire includes items which are supposed to measure the non-observable construct. The perfect result is obtained when the measurement of each item produces an observed score X_0 which is exactly equal to the true level X_T. In this optimal situation, a difference in the levels X_0 measured by two items would be due to effective differences between the latent characteristics the items are trying to measure. Thus, it is attributable to true differences in X_T. But it is clear that this fortunate case rarely happens. In most of the cases, the X_0 level differences reflect other factors, such as (Selltiz et al., 1976):

- stable factors affecting the score, for example the individual's will to express his/her personal true feelings;
- temporary personal factors, like individual mood;
- situational factors, e.g. if the questionnaire is filled in at home or not;
- sampling factors, in particular for the sampling of items, because the exclusion of specific items could lead to some differences in the observed scores; even the change in the items wording could affect the scores;
- lack-of-clarity factors, due to some ambiguous questions posed to respondents and which can be misinterpreted by them;
- mechanical factors, e.g. answers coded in a wrong way;

These are all factors that bring to differences between the observed scores and the true one. Not all of these factors are present in every measurement, but they can generally affect every kind of questionnaire. The impact of these factors on the X_0 level varies from one case to another, but it is predictable. The aforementioned factors bring the observed scores to be distorted from the effective ones.
Assuming the relation between the scores and the factors to be linear, Churchill (1979) expressed it as:

$$X_0 = X_T + X_S + X_R. \tag{2.1}$$

In Equation (2.1) X_S represents the set of systematic sources of error, e.g. stable characteristics of the interviewed and of the measure affecting item score, while

X_R is the set of random sources of error affecting the observable score, such as non-permanent factors (e.g. personal reasons, like the mood or fatigue of the respondent, etc.).

It is necessary at this point to give some definitions about the observable measures, which are the items of the questionnaire in this specific context.

An item is *valid* when it measures well what it is supposed to measure. That is, a measure will be valid if it coincides exactly with the phenomenon of interest. Using Churchill formulation, the validity occurs when $X_T = X_0$. It means that both systematic and random errors have to be (approximately) equal to zero.

An item is instead *reliable* when its measure leads to consistent and stable results (Peter, 1979). Thus, reliability is verified if independent but comparable measures of the same latent construct agree. It is clear that the concept of reliability strictly depends on the random error. Using the same formulation above-mentioned, it means that $X_R = 0$. For this reason, an item has to be considered reliable when its score is not due to any kind of random error.

The mathematical formulations of these two concepts make it obvious the relation between the two. Validity implies reliability, because to have validity both the sources of error have to be zero. But, on the contrary, a reliable measure could be not valid: under the assumption of reliability, $X_R = 0$, the observed score X_0 could still be equal to $X_T + X_S$. Therefore, reliability is a necessary but not a sufficient condition to assess validity (Churchill, 1979).

2.2 The different kinds of validity

This short section is meant to be explanatory of the three main kinds of validity. At first, let us remind that validity is the degree to which an item is able to measure exactly what it is supposed to.

Construct validity

Construct validity is verified if the items measuring a latent construct are effectively related to it. For example, a tool that is supposed to measure satisfaction is "constructually" valid if all the measures contained in the tool are capable to measure exclusively aspects that are theoretically related to satisfaction. On the other hand, if the items used are also capable to measure other concepts strictly related to satisfaction (it could be, e.g., customer's loyalty), they might not have enough construct

validity as measures of satisfaction (DeVon et al., 2007).
It is possible to evaluate the construct validity of an item or a scale in many different ways. One of these is Factor Analysis, which is going to be adequately explained later on in the next section.

Translational validity

The translational validity can be divided in two other kinds of validity.
Face validity establishes that the item seems to be measuring the latent construct under study. It is the easiest way to assess validity, but at the same time it is the weakest form of validity. Actually, it does not provide an indication of how well the instrument measures the latent construct of interest. However, it could give an idea of how potential respondents might interpret and answer to the items of the questionnaire (DeVon et al., 2007). Face validity can be confirmed checking the grammar and syntax of the items, in order to be sure that they look appropriate and have a logical flow.
Content validity is assessed if the items of a measurement scale cover most of the domain of the concept under study. Being this concept a latent one, it is obviously impossible to manage to cover all the aspects of the construct. Referring to the previous literature and research on the topic might be a solution to have content validity of a measurement scale. Then, a group of experts in the subject could confirm the accuracy and correctness of the set of items included in the scale.

Criterion validity

Criterion validity concerns the demonstration of the relation existing between the score given to an item and another variable, which is called *criterion variable.* This variable is usually included in the measurement scale and it consists in a question which is supposed to evaluate directly the latent construct. One also refers to it as *gold standard* (DeVellis, 1991). As it was said in Chapter 1, the measurement scale under study finds its own criterion variable in item 12, the one asking directly to students how much they feel satisfied with the course they attended. Thus, it is clear that the measures contained in the scale have to be related to this standard as much as possible. That is because a strong association between an item and the criterion variable would suggest an equally strong relation between the same item and the latent construct it has to measure. In this case the criterion-related validity is fulfilled.

2.3 Methods to assess reliability

Reminding what was said above, reliability refers to the capability of an instrument to measure consistently a construct. It is a necessary component but, as we already explained, not sufficient to assess validity.

In this section we want to focus on several approaches useful to know if the instruments used to express a latent variable are reliable and effective. These measures we obtained in the empirical part of the study demonstrated to be really helpful in the comprehension of the latent structure of the data.

2.3.1 Correlation indices

A measurement scale is considered reliable when all the items which it consists of are strictly related to the latent construct underlying the scale. As a consequence, the items are correlated to each other and this *inter-item correlation* (for each pair of items) is thus an indicator of reliability. An inter-item correlation above the threshold 0.3 (Hair et al., 2010) is considered good enough to state the existence of a relationship between the items.

Another way to assess reliability is calculating the *item-total correlation*, which expresses the strength of the relationship between an item and the overall scale. This kind of correlation is strictly related to the concept of reliability: the more every item correlates to the whole scale, the more likely they are correlated to each other.

The item-total correlation has inflated values, due to the presence of the same item in the measurement scale. There is a corrected version of this index, usually called *item-rest correlation* and which has to be preferred to the first one. As its name suggests, it shows the strength of the relationship between an item and the rest of the measurement scale (the same scale without that specific item). As Hair et al. (2010) proposed, when the item-rest correlation is above the threshold of 0.5, it is sign of high coherence of the item with respect to the rest of the items. Otherwise, if the value is below that threshold, it does mean that the item is not so consistent with the measurement scale and, thus, with the latent construct underlying the measures. In this situation, it is preferable to study deeper the non-coherent items and, where appropriate, decide to exclude them from the measurement scale.

2.3.2 Cronbach's alpha

Another measure to express a scale reliability is obtained calculating its degree of internal consistency, by means of the Cronbach's alpha. This coefficient, introduced by Cronbach in 1951, shows how much the items of a scale are related as a group. It can be written as a function of the number of the items composing the scale and of their mean correlations. Hence, it follows that using a large number of items leads to think there is strong internal consistency, while only few of them could be actually related to each other.

According to Cronbach (1951), the coefficient alpha can therefore be indicated as a variances ratio. More specifically, it is the proportion of the shared variance among the items to the total variance, rescaled to a function of the number of items. The proportion of shared variance is just the ones' complement of the proportion of items unique variance. To have it clearly expressed, for $i, j = 1, ..., k$ and $i \neq j$

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma_{ij}}{\sigma_t^2} = 1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_t^2}, \tag{2.2}$$

where k is the number of items used in the scale; σ_{ij} is the covariance between the items i and j; σ_i^2 is the variance of the single item i (unique variance of the item); σ_t^2 is the total variance and consists of the sum of all the k^2 elements of the items variance-covariance matrix.

Cronbach's alpha is therefore obtained multiplying this ratio by a correction factor:

$$\alpha = \left(\frac{k}{k-1}\right)\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma_{ij}}{\sigma_t^2}. \tag{2.3}$$

Being a ratio, the coefficient in Equation (2.3) varies from 0 to 1. If it is measured for the overall scale, a low value of the index is sign of poor internal consistency and the scale should therefore be reviewed as a whole; on the contrary, a value near to the maximum expresses the goodness of the set of items in measuring the latent construct.

This coefficient can be computed for each item singularly, removing the item itself from the measurement scale. In this case, if the value of Cronbach's alpha has reduced compared to the overall value, it suggests that the item excluded needs to be reintroduced in the scale. If the elimination of the item brings instead to an increase in the value of the alpha, it is not so consistent with the rest of the scale and it should be removed.

Acceptable values to assess a good overall internal consistency are over 0.7, but excellent values are the ones near 0.9 (Hair et al., 2010). Obviously, there are not predefined values for the coefficients without a single item, which must be compared, from time to time, to the overall alpha coefficient.

2.3.3 KMO coefficient

Kaiser-Meyer-Olkin (KMO) coefficient is a measure of sampling adequacy to perform a Principal Component Analysis or a Factor Analysis. It provides the proportion of variance that is in common among the items and thus indicates the presence of a common factor underlying the observable variables.

Assume that we dispose of the correlation matrix R and its generic element r_{ij} for $i, j = 1, \cdots, k$ and being k the total number of items composing the measurement scale. We can define then the partial correlation matrix Q, with its generic element q_{ij}, where i and j are defined as for R. With the term partial correlation it is meant the correlation between two items, keeping constant all the other items.

This coefficient can be calculated for the overall scale or just to assess the adequacy of each single item. As Dziuban and Shirkey (1980) reported, the KMO coefficient for the single item j (thus, keeping j constant in the formula) is

$$KMO_j = \frac{\sum_{k \neq j} r_{jk}^2}{\sum_{k \neq j} r_{jk}^2 + \sum_{k \neq j} q_{jk}^2}. \tag{2.4}$$

The individual coefficient given in Equation (2.4) assesses if that specific item represents adequately the domain of the latent construct. For this reason, it expresses the worthiness to include the same item in the scale.

The global KMO coefficient is obtained from the previous formula (2.4), taking all the possible values for j and maintaining $j \neq k$.

The KMO meaasure varies from 0 to 1. A value near to the minimum one indicates that the partial correlations are much larger than the correlations. This is a problem for Factor Analysis, because it indicates that the correlation are widespread and not clustering among the items. An index value near to the maximum indicates that Principal Component and Factor Analyses are good methods to proceed with the study of the data.

Kaiser in 1974 expressed some thresholds for his KMO index, as follows:

”in the .90s, marvelous
in the .80s, meritorious
in the .70s, middling
in the .60s, mediocre
in the .50s, miserable
below .50, unacceptable”.

2.3.4 Principal Component Analysis

Principal Component Analysis (PCA) is a mathematical technique able to reduce a set of correlated variables in a smaller set of uncorrelated components, representing the latent constructs underlying the items. However, these components retain most of the information contained in the original variables.
Thus, this method usually allows a great simplification in the complexity of the data, but it requires a set of mutually correlated items to be effective.
PCA indicates every principal component as a linear combination of the manifest variables (see, for instance, Jolliffe, 2002). The number of components calculated by this method is equal to the number of original variables:

$$\begin{gathered} c_1 = \lambda_{11}Y_1 + \lambda_{12}Y_2 + \cdots + \lambda_{1k}Y_k + \varepsilon_1 \\ \cdots \\ c_k = \lambda_{k1}Y_1 + \lambda_{k2}Y_2 + \cdots + \lambda_{kk}Y_k + \varepsilon_k, \end{gathered} \tag{2.5}$$

where the Y_i is one of the k items; λ_{ij} is the loading linking the observable variable Y_j to the component c_i; ε_i is the error term in the i^{th} equation. The loading is thus expressible as the correlation between the item and the latent variable which it is meant to measure.
The first step in the principal component procedure provides for a first component, which extracts from the items the highest possible of the total variance. After having removed the amount of variance explained by the first component, the procedure continues with the further steps. A second principal component is extracted and it explains the maximum of the remaining unexplained total variance. The procedure goes on, as we said before, until the number of components equals the number of items and all the variance is thus explained.
The main aim of PCA is therefore to choose a reduced set of (uncorrelated) variables to explain data variability. Hence, selection criteria are used to decide on the right

number of components to be retained in the solution. Kaiser Criterion (Kaiser, 1960) is a simple method to take this decision and it states that all the components having an eigenvalue above 1 have to be considered in the new set of variables. The eigenvalues associated to each component indicate the share of total variance which the components are explaining. Thus, just components explaining more than a fixed proportion of variance are chosen. Another criterion often used to assess the number of principal components is the *scree test*, introduced by Cattell in 1966. It is a graphical method and provides a plot of the eigenvalues against the number of components in order of extraction. Cattell suggests that the point at which the scree begins indicates the true number of components to be kept.

It is not a foregone conclusion that these two criteria will provide the same results. It is the researcher's task to understand which solution makes more sense and it is therefore preferable.

2.3.5 Factor Analysis

Factor Analysis (FA) is one of the most important methods to assess the reliability and the real dimensionality of a measurement scale. It is obviously related to the above-mentioned PCA, but they are often erroneously confused. While PCA considers the whole variance of the items, again let us call it total variance, FA uses only a portion of this variance. In this procedure, each item variance, derived from the correlation matrix, is divided into two parts: a *common variance*, namely the variance which is shared among the different indicators of the scale; a *unique variance*, which is composed of a variance specific to the indicator (i.e., variance which is not shared with the other items) and of random error variance (due to measurement error or unreliability of the item). For references see, for example, Bryman and Cramer (2005) and Brown (2006).

PCA ignores the difference between unique and common variances, because this procedure analyses all the total variance. FA considers only common variance among the manifest variables and hence it tries to exclude the unique variance from the analysis (Bryman and Cramer, 2005).

While the first procedure is used mainly to obtain simplicity in a complex data set, the second one has also another important aim. In addition to complexity reduction, FA allows a deep understanding of the structure underlying the measurement variables. It is in fact oriented to the relationship linking the items and the latent constructs, called in this case factors, and the items among each other. The factors are therefore unobservable variables explaining the correlations among the observed indicators. More explicitly, we can say that the items are correlated because they have a common cause and so they are influenced by the same latent factor (Brown, 2006).

Differently from what happened for PCA, Factor Analysis sees each observed variable Y_i as a linear combination of the latent factors F_j (plus an error term):

$$\begin{gathered} Y_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1k}F_k + \varepsilon_1 \\ \cdots \\ Y_k = \lambda_{k1}F_1 + \lambda_{k2}F_2 + \cdots + \lambda_{kk}F_k + \varepsilon_k, \end{gathered} \tag{2.6}$$

where λ_{ij} indicates the factor loading linking the item Y_i to the factor F_j. Just as in PCA, the number of factors is equal to the number of items, but only a reduced number of them is retained.

This method is supposed to extract the proportion of variance that is due to common factors and that is shared among the observed measures. This proportion of variance is usually called *communality*. This concept is in contrast to the one of *uniqueness*, which namely expresses the proportion of variance unique to each item. Being factors the cause of shared variance among the items, it would be desirable to have a high value of communality to perform an effective Factor Analysis.

Factor Analysis can be divided in two distinct phases: a preliminary phase, called *Exploratory Factor Analysis* (EFA), and an advanced phase, the *Confirmatory Factor Analysis* (CFA). These phases are extremely different, but complementary in their uses and aims.

The first objective of EFA is the evaluation of the dimensionality of a questionnaire, achieved by discovering the minimum number of latent factors able to explain the correlations among the items. EFA does not require the knowledge of the specific structure of the data, i.e. the exact number of latent factors. Thus, there are not a priori formulations on the relationship between the indicators and the latent constructs (Brown, 2006).

Many methods can be used to extract the factors. One of them is the *Principal Component Method* and it is the one used in the analyses of the present study. Through this method, the factors extracted coincide exactly with the principal components and for this reason there is not need for further explanation on this procedure.

There are still debates on the suitability of using PCA as an estimation method of Factor Analysis. Fabrigar et al. (1999) attested that in some circumstances PCA and EFA lead to different results, in particular when communalities among the items are low or when there are few indicators for each factor. As we will show later on, these circumstances are not verified in our study and thus we can assume that PCA and EFA provide in this case similar results.

Just as EFA, CFA finds its main objective in identifying the latent factors responsible for the variance and covariance among a set of manifest indicators. However, while the first procedure has just descriptive and exploratory intention, the second one requires a preliminary knowledge of the latent structure of the data. Thus, the number of factors and the item-factor loadings pattern have to be decided in advance. This is the reason why CFA is used in later phases of analysis, after the underlying constructs have been purportedly decided by means of EFA and theoretical studies on the topic (Brown, 2006).

CFA is a particular type of Structural Equation Modelling (SEM) and, in our case, it relies on the maximum likelihood (ML) estimation method. When such a method is used, it is worthy to evaluate the CFA model through different goodness-of-fit indices. These indices state how well the solution proposed replicates the observed variances and covariances among the items.

The next brief section is going to introduce the main goodness-of-fit indices, which were used to verify the goodness of our CFA models.

Goodness-of-fit indices

Likelihood Ratio test (LR test) is a test proposed for the first time by Wilks (1938) and it was used in the present study to verify if the proposed model differs substantially from the saturated model (i.e., the model reproducing exactly all the variances, covariances and means of the observable variables). If it does not, the likelihood ratio will be near to 1, being the two likelihoods statistically equivalent. In the same way, if log-likelihoods are considered, their difference will be close to 0, leading to the same conclusion. It was demonstrated by the same Wilks that this test is distributed as a χ^2 with degrees of freedom equal to the difference of the number of parameters used in the two tested models.

The *Root Mean Square Error of Approximation* (RMSEA) is one of the most important and most reported goodness-of-fit measures in the Structural Equation Modelling context. As expressed by Browne and Cudeck (1992), a RMSEA value below 0.05 is sign of very good fit of the model. A value between 0.05 and 0.08 of the same index denotes mediocre fit, while a value above 0.08 indicates poor fit. Some

researchers set the last acceptable threshold to 0.10 (see, for instance, Baker, 2008).
The *Akaike Information Criterion* was formulated by Akaike in 1974. Given a model with p parameters to be estimated and its maximum likelihood estimate L, $AIC = 2p - 2ln(L)$.

Thus, AIC expresses the information a given model is able to explain, weighted by the number of parameters present in the same model. As it is notable in its formulation, AIC penalises complex models, i.e. models with many parameters. Given a set of estimated models, the model to be chosen is the one with the lowest AIC value. Therefore, this index allows comparisons even among models with different numbers of parameters. However, it does not provide a measure of model fit quality in absolute terms.

The *Comparative Fit Index* (CFI), formulated by Bentler (1990), establishes a model fit by analysing the ratio between the specific model and the baseline model discrepancies. For baseline model, it is meant the model in which all the variables are considered uncorrelated, i.e. there are not latent variables; in this case, discrepancy of a model is considered as the difference between the observed and predicted variance-covariance matrices. CFI shows the worthiness in using the model of interest rather than the baseline model. Varying from 0 to 1, this measure indicates optimal fit when it is near to the maximum. Hu and Bentler (1999) considered value 0.95 as a threshold beyond which the model shows very good fit.

The *Tucker-Lewis Index* (TLI) by Tucker and Lewis (1973), also known as non-normed fit index, is similar to the previous measure. As CFI, it is based on the discrepancy between the model selected and the baseline one. TLI has a range from 0 to 1, being 1 sign of excellent fit and, just as for CFI, Hu and Bentler (1999) set TLI threshold at 0.95.

The *Standardized Root Mean Square Residual* (SRMR) represents a standardised solution of the square root of the difference between the sample correlation matrix and the hypothesised model one. The standardisation allows this coefficient to vary in a range that goes from 0 to 1, where lower values mean better fit. A SRMR value below 0.05 is index of high fit quality, even if Hu and Bentler (1999) suggested that values up to 0.08 should be considered acceptable for a good fit. It must be highlighted that SRMR will be lower in presence of a model with a high number of parameters and based on large sample size (Hooper et al., 2008).

The *Coefficient of Determination* (CD) provides information that is similar to the well-known R^2 computed in OLS regressions.

Average Variance Extracted (AVE) and *Composite Reliability* (CR) are two addi-

tional methods to assess reliability in a measurement scale.
AVE was developed by Fornell and Larcker (1981) and it measures the level of (common) variance explained by the latent construct, in opposition to the level of variance due to random measurement error. Being a proportion, AVE varies from 0 to 1 and usually values above 0.7 are considered good, whereas between 0.5 and 0.7 are however acceptable (Hair et al., 2010).
CR index shows how much the items share in measuring the underlying construct and therefore the extent to which the latent factor can be considered reliable. According to Hair et al. (2010), CR values exceeding 0.7 support the hypothesis of internal consistency and reliability of the construct.

2.4 Unidimensionality of a measurement scale

The particular structure of the data at our disposal, further outlined later on, requires the explanation of the concept of *unidimensionality* in respect of a scale.
Due to the limits concerning single-item measures of a latent construct (Churchill, 1979), respondents are often asked to give more measures (i.e. a scale), which are supposed to be alternative indicators of the same non-observable construct (Segars, 1997). Unidimensionality refers to the presence of one single latent construct underlying an entire items scale (Gerbing and Anderson, 1988). In few simple words, Hattie managed to state clearly the meaning and the relevance of this concept: "One of the most critical and basic assumption of measurement theory is that a set of items forming an instrument all measure just one thing in common" (Hattie, 1985). The mathematical definition of unidimensionality can be traced back to the CFA model specified above in Section 2.3.5. In CFA unidimensional model, a set of items share one single common underlying factor Z. An example of such a model is represented in Figure 2.1. To simplify the idea, in the example shown there is only one factor. There could be more factors in the model, each with different and not shared items, but unidimensionality would be established anyway. Each indicator y_i is linked to the latent variable Z through the factor loading λ_i. To express it formally,

$$y_i = \lambda_i Z + \varepsilon_i, \tag{2.7}$$

where $i = 1, ..., k$ being k the number of the items and ε_i is the generic residual, uncorrelated with other the factors (if present in the model) and with the other items residuals (Segars, 1997).

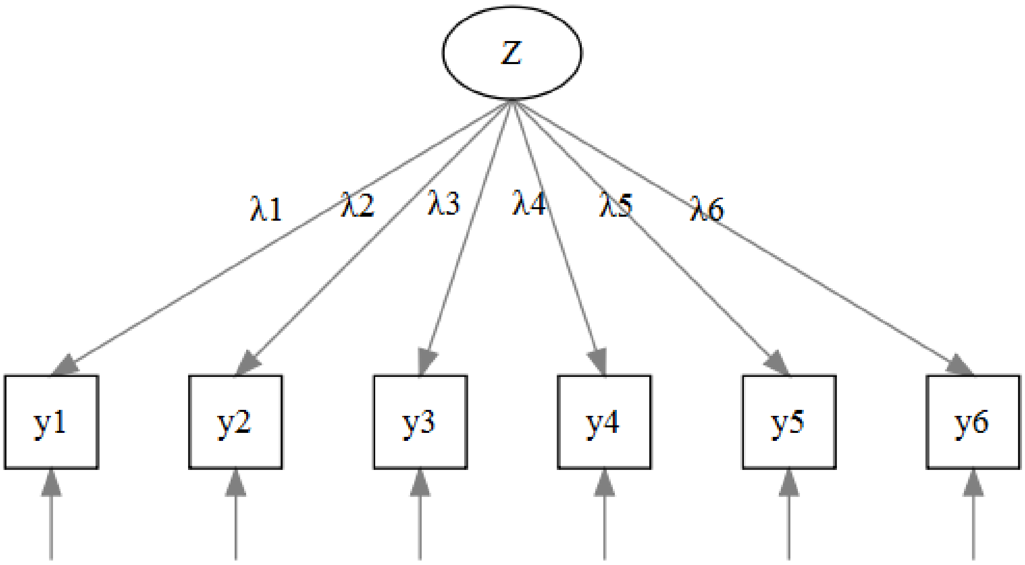

Figure 2.1: A representation of a unidimensional CFA model with six indicators.

The development and evaluation of a measurement scale is usually based on the methods expressed in the previous Section 2.3, for instance Cronbach's alpha, item-total correlations, EFA or CFA. These are all instruments to evaluate reliability. However, they could lead to draw different conclusions because they are different choice criteria. Among these, only Confirmatory Factor Analysis is able to test effectively unidimensionality as it was defined in Equation (2.7) (Gerbing and Anderson, 1988).

After having determined the latent structure underlying the data by means of the methods analysed in this chapter, it is then possible to use more complex models. In particular, the present work aims to investigate the change in students' satisfaction by using latent growth models. The explanation of this kind of models is the main objective of the following chapter.

Chapter 3

Latent Growth (LG) analysis: definition of LG models

Approaches such as Principal Component Analysis, Factor Analysis and Structural Equation Modelling, previously described in Chapter 2, are variable-centred ones. Thus, they describe the connection among the variables and aim to identify and explain the way these observable variables are related to each other.

There is another kind of approach, including methods such as Cluster Analysis and Latent Curve Analysis. These ones are usually defined as individual-centred approaches. It means that the focus is moved to the relationships between the individuals. This kind of approaches aim to classify the individuals into different sub-populations based on their characteristics and their responses. This is made in order to have similar observations within the same group, which are different from the observations in the other groups (Jung and Wickrama, 2008).

Latent curve analysis, or latent growth modelling (LGM), is a particular parametrization of SEM which has proven to be a good method for analysing change of individuals in time. Contrary to what usual SEM method does, LGM considers longitudinal data and not cross-sectional one. It allows to study the growth of a latent construct using, typically, the same observable variables as indicators over time. LGM can provide the estimates of relevant features of change, like the individuals' status at the measurement starting point, individuals' growth trajectories across time and the variability among the individuals in their starting points and in their growth rates (Hancock and Buehl, 2008).

This chapter has the purpose of explaining the main features of latent growth models, to make it clear their usefulness in the specific context of students' satisfaction.

Such models were estimated by means of the EM algorithm, an iterative approach computing maximum-likelihood estimations. Each iteration of this algorithm consists of two different steps, first the expectation step (E-step) and then the maximisation one (M-step). For a better explanation about the algorithm, one can refer to Dempster et al. (1977), who explained the use of this method in presence of "incomplete data" (i.e. data including both observable variables and other unobservable variables but expressible through the observable ones).

3.1 Unconditional latent growth model

The first step in latent growth analysis is to consider an "unconditional model", i.e. a model in which there are not covariates affecting the latent structure and predicting the growth over time. This is the simplest model of its kind, but even if it will not be shown in our study, it clarifies the structure of its extensions we actually used in the analyses.

The conventional growth modelling here presented assumes the observations in the dataset to come from a single population and therefore, one single growth curve is considered sufficient to approximate adequately the entire population.

A latent growth model can be described as a multilevel, random-effects model: there is some variability among the individuals and it can be explained by latent random effects (continuous latent variables, i.e. intercept and slope). It is usual to refer to latent intercept and slope, called respectively α and β, as growth factors (Muthén, 2004). The intercept α represents the value of the observable variable when the growth curve begins, i.e. its initial level. Hence, it is the average score of the variable in the first time point (the first year, in our example). The slope β indicates how much the curve grows in time, because it represents the rate of change. Figure 3.1 shows an example of the model just described. All the observations were considered deriving from the same population (class) and the manifest variables y_t for $t = 1, 2, 3$ represent the same measure (item) collected in three different time points. The latent growth factors underlying the observable variables, i.e. α and β, explain the change over time through a linear growth trajectory. Different constraints may be imposed on the factor loadings connecting the observable variables to the growth factors. The shape of the growth path can be estimated using the mean and covariance matrices of the observed measures. For sake of simplicity, the models here presented assume a linear growth, a good approximation of the real trajectory, as it will be shown later on. To specify that linear growth, the loadings of the intercept α are constrained to 1, while the ones of the slope β are increasing from 0 to 2, as the measures are collected in three consecutive years (constraints are considering equally spaced time points).

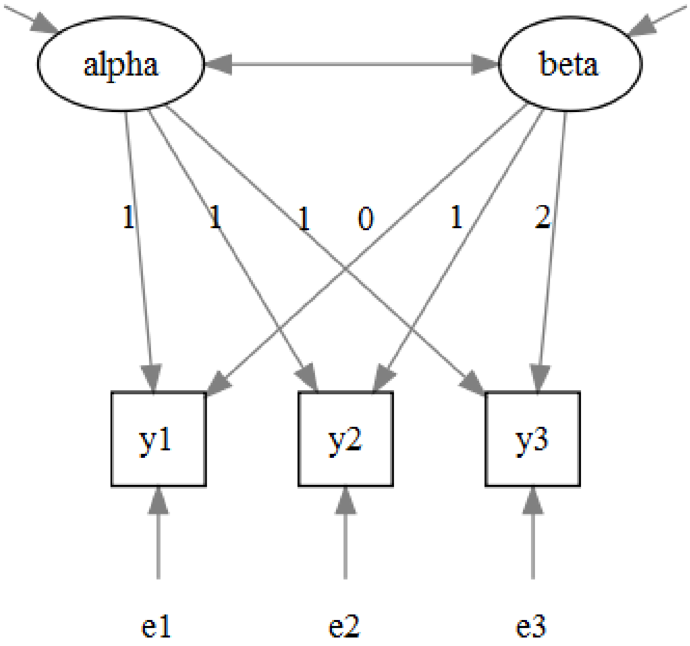

Figure 3.1: A single-class latent growth model.

In mathematical terms, let us consider

$$y_{it} = \alpha_i + \lambda_t \beta_i + \varepsilon_{it}, \tag{3.1}$$

where the variable y_{it} is referring to the value of the item for individual i ($i = 1, ..., N$, being N the total number of observations in the population) at time t ($t = 1, ..., T$, being T the number of occasions in which the variable is measured); the vector ε_i includes all the error terms for the i^{th} individual at each time occasion and these errors are distributed normally with an average of zero. The random intercept and slope can be expressed as it follows:

$$\alpha_i = \mu_\alpha + \zeta_{\alpha_i} \tag{3.2}$$

$$\beta_i = \mu_\beta + \gamma\alpha_i + \zeta_{\beta_i}, \tag{3.3}$$

where $\zeta_{\alpha_i} \sim N(0, \psi_\alpha)$, $\zeta_{\beta_i} \sim N(0, \psi_\beta)$; ζ_{α_i}, ζ_{β_i} and ε_{it} are mutually independent, for each individual and in each time point. The term γ in Equation 3.3 represents the regression coefficient of the slope on the intercept. The parameters of interest in this model are in particular the means and variances of the latent growth factors, as well as the regression coefficient γ linking these two random effects (Salgueiro et al., 2013).

We recall that this model assumes all the individuals to have the same change across time. This is a very restrictive assumption, as it oversimplifies the real changes occurring in the different observations. There may be subsets of individuals whose change trajectory is completely different from the estimated average trajectory. Moreover, there might be more latent levels and several covariates affecting the growth factors and/or the latent construct. For all these reason, the aforementioned model can only be considered as a good starting point for the application of more complex models, which will be explicated in the next sections.

3.2 Unconditional second-order LGM

The second-order latent growth model is a further extension of the previous model. It is called so because there are two latent construct levels in it, while there is still only one level of observable variables. This kind of models is useful when the variable of interest is unobservable and it is important to study the change in the same latent variable, rather than the change in the observable ones. Figure 3.2 shows an example of such a model.

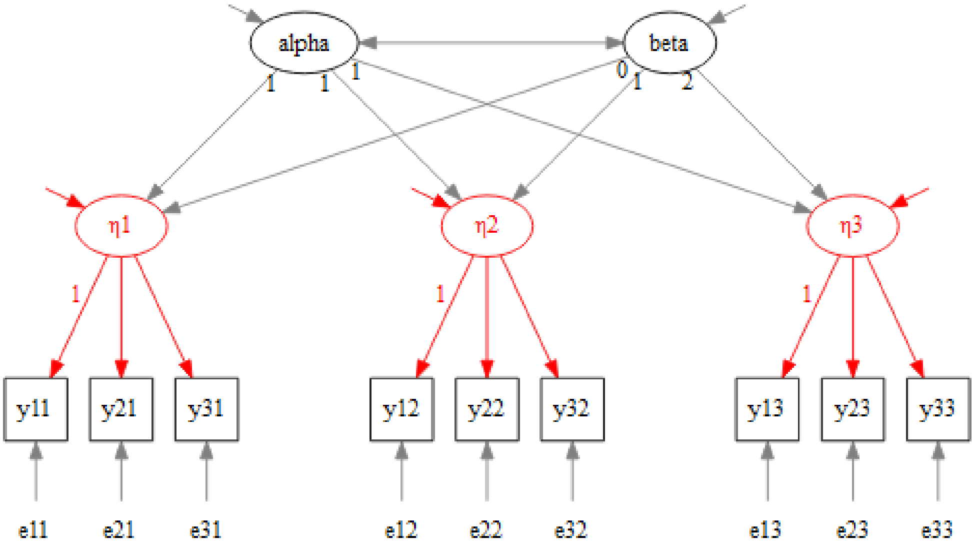

Figure 3.2: An unconditional second-order latent growth model. Highlighted in red, the additional latent level.

The one proposed is an unconditional latent growth model, in which a latent variable η is measured by three observable indicators (y_1, y_2, y_3) in three different time points. As it will be shown in the analyses, the factor loadings of the model can be seen as invariant in time. This restriction is obtained imposing equality constraints on the factor loadings linking the items to the factors. The error terms of the items could be correlated over time, but in the case presented later on these correlations have to be considered negligible.

In order to give a general description of the model, let us recall that the aim is studying the change occurred in the latent construct through T different time points. Let η_t be the latent construct at time t, which underlies J observable variables y_{jt}, for $j = 1, ..., J$ and $t = 1, ..., T$. Each variable is measured for every one of the N individuals. The equation linking the observable variables to first level latent construct is

$$y_{jit} = \tau_{jt} + \lambda_{jt}\eta_{it} + \nu_{jit}, \tag{3.4}$$

where y_{jit} is the j^{th} measured variable in time t for the individual i; τ_{jt} is the intercept for the variable j at time t; Λ_{jt} is the factor loading connecting observed indicator j at time t to latent construct η_t; ν_{jit} is a normal random error referred to variable j for individual i at time t. In the same way, the growth occurring in the η_t is described by the following formula:

$$\eta_{it} = \alpha_i + \lambda_t\beta_i + \zeta_{it}, \tag{3.5}$$

with latent growth factors α and β expressed as in Equations 3.2 and 3.3. The error terms ζ_{it} have to be considered mutually independent for $i = 1, ..., N$ and $t = 1, ..., T$. They are normally distributed with zero mean and a variance dependent on t, i.e. $\zeta_{it} \sim N(0, \theta_t)$. The first latent factor, intercept α, is the initial amount of η_t. The second one, slope β, is the rate of change for the same construct over time. The key parameters are the ones specified in the previous model, namely means and variances of the random effects and regression coefficient of random slope on random intercept, as well as the residual variances over time and the factor loadings of the measurement part of the model (Salgueiro et al., 2013; Hancock and Buehl, 2008).

3.3 Conditional second-order LGM

Latent growth models shown in the previous Sections 3.1 and 3.2, i.e. the unconditional ones, try to describe a growth over time. Their further extensions, conditional latent growth models, try to explain the growth using predictors that could affect the individual change in time. Figure 3.3 shows the conditional part of the model, in which the covariate x is a time-invariant variable (or a set of variables), both continuous or dummy, influencing the growth factors. Time-variant explanatory variables x_t are also introduced in the model and they affect the first latent level of the model, the construct η_t.

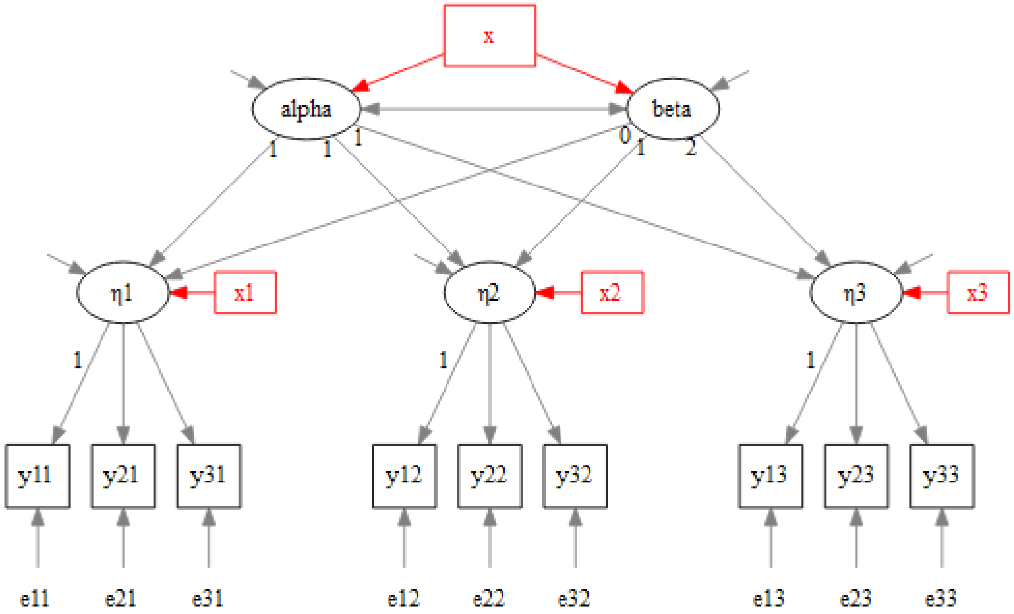

Figure 3.3: A conditional second-order latent growth model. Highlighted in red, the conditional part of the model (covariates).

Unconditional models can be easily extended in a conditional shape by correcting the Equations 3.2 and 3.3, the ones relating to random intercept and slope, as it follows:

$$\alpha_i = \mu_\alpha + \gamma_\alpha x_i + \zeta_{\alpha_i} \tag{3.6}$$

$$\beta_i = \mu_\beta + \gamma \alpha_i + \gamma_\beta x_i + \zeta_{\beta_i}. \tag{3.7}$$

The additional key parameters of this model include the regression coefficients of the covariates on the random effects. The parameter γ_α, related to the random intercept, is the expected change in the mean of the latent growth factor α for an unit increase in the covariate x. The parameter γ_β, associated to the random slope, is the expected change occurring in the growth rate for a unit change in the covariate x (Salgueiro et al., 2013).

3.4 Second-order latent growth mixture model

The growth mixture model (GMM) is an extension of the model described in Section 3.3. It relaxes the assumption that considers all the individuals coming from the same population. Actually, it takes into account different groups within the whole population, in order to explain longitudinal data. That is because assuming homogeneity in the growth factors (i.e. same latent intercept and slope for all observations) is often not realistic (Bassi and Dias, 2013).
Sub-grouping of all observations is achieved through a categorical latent variable. This variable allows for each group to have its own change trajectory, different in mean and form from the other groups ones. It results in different growth models, one for each latent class, each with its own estimates. GMM can take into account also covariates influencing the latent variables (both categorical and continuous predictors). Figure 3.4 presents an example of this kind of model. Covariates x, x_1, x_2, x_3 are affecting the growth factors α and β (conditional part) and the categorical latent variable C splits the population in different subgroups.
The decision on the number of latent classes to use in the analysis could be suggested by many procedures, such as BIC or LMR likelihood ratio.
The Bayesian Information Criterion was defined by Schwarz in 1978 as

$$BIC = -2\left[ln(L)\right] + p\left[ln(N)\right]. \tag{3.8}$$

In Equation (3.8) p is referring to the number of parameters in the model, while N to the sample size; $ln(L)$ is the log-likelihood. A small BIC value corresponds to a good model with large log-likelihood and a small number of parameters. It has to be reminded that this criterion does not indicate the goodness of a model in absolute terms, but only the goodness in comparison with another model.
Lo et al. (2001) proposed a likelihood ratio-based method, in order to test a number of $k-1$ different classes versus k. In the GMM context, LR test involves nested models in which the constrained model is obtained from the other one when a parameter assumes a value on the border of its parametric space (in this specific

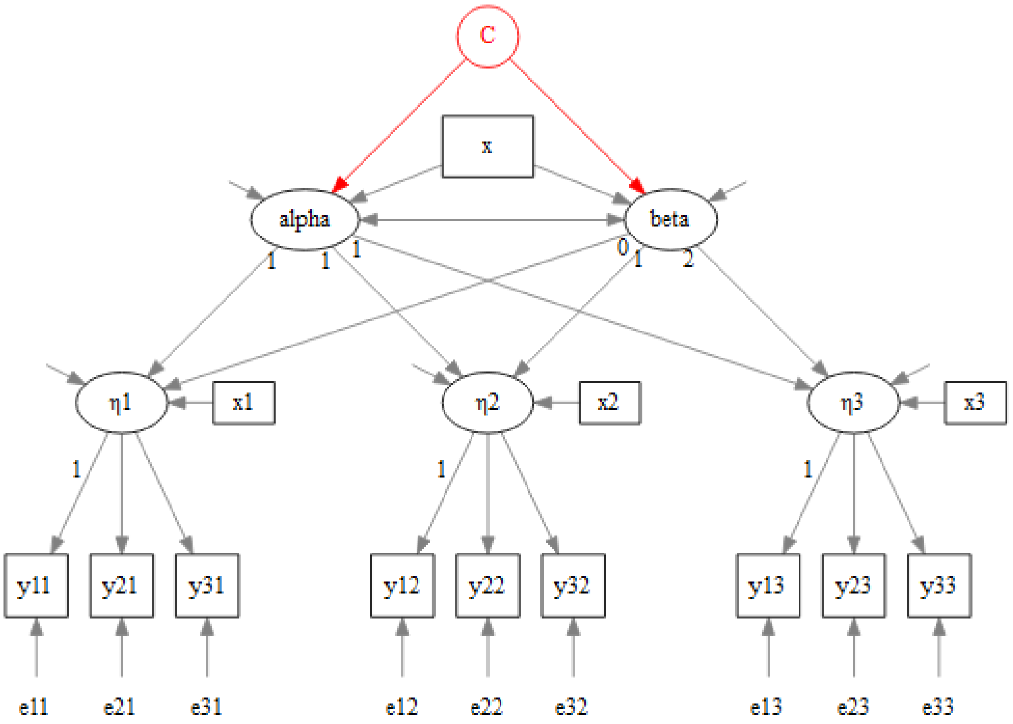

Figure 3.4: A conditional second-order latent growth mixture model. Highlighted in red, the mixture part of the model (categorical latent variable).

case, it would be a latent class probability equal to zero). This likelihood ratio does not follow a chi-square distribution, as instead the classical LR test does. Lo et al. use the same ratio, having managed to derive its exact distribution. A low p-value leads to the rejection of the constrained model (with less classes) in favour of a model with a higher number of classes (at least one more). Determining the appropriate number of categories depends not only on the fit indices, but even on elements such as parsimony, research question and interpretability of results. The modality c_i of the categorical latent variable C indicates the unobserved membership for the individual i to a specific class. The variable C assumes K different values, being K the number of latent subgroups. Considering for simplicity a single covariate x, it is possible to have a model in which the covariate affects only the latent intercept and slope of each class. In this case, the covariate does not influence class membership in any way. Thus, the probability to belong to class c_i for the individual i is calculated only on the basis of the item scores, the observable values. The effects of the covariate on the growth factors of each class are estimated only subsequently. This is the situation we will show later on. However, for sake of completeness, it is possible to estimate class probabilities considering x since the beginning. The latent variable

C could be related to the covariate by means of a multinomial logistic regression. It means that x could have a direct effect on both class membership and the growth factors of each class. Let us consider C as a K-categories variable,

$$P(c_i = k|x_i) = \frac{e^{\gamma_{0k}+\gamma_{1k}x_i}}{\sum_{s=1}^{K} e^{\gamma_{0s}+\gamma_{1s}x_i}} \tag{3.9}$$

with the standardizations $\gamma_{0k} = 0$ and $\gamma_{1k} = 0$. For clarity, let us assume that C is a binary variable, assuming values 1 and 2. Therefore only two latent subgroups are considered within the population. According to the logistic model,

$$P(c_i = 1|x_i) = \frac{1}{1+e^{-l_i}}, \tag{3.10}$$

where l_i is the following logit (i.e. log-odds):

$$log\left[\frac{P(c_i = 1|x_i)}{P(c_i = 2|x_i)}\right] = \gamma_{01} + \gamma_{11}x_i. \tag{3.11}$$

Thus, as it is notable from Equation (3.11), γ_{11} is the increase in the log-odds for being in the first group, due to a unit increase in the covariate x. Assuming, for sake of simplicity, the covariate x to be a dichotomous variable, $e^{\gamma_{11}}$ represents the odds ratio for being in the first class rather than in the second one. It means that the odds of being in the first group is $e^{\gamma_{11}}$ times higher for the ones presenting the characteristic of $x = 1$.

In case the covariate has significant effects on the latent variables (categorical C, growth factors α and β), all the other models shown in this chapter would lead to biased results. In fact, not considering the influence of covariates on class membership, the observable variables could be not appropriately associated to the latent classes. To understand it clearly, consider the analogous case of a misspecified regression model, in which the estimates would be biased not using an important predictor in the analysis. The bias of the effect of the categorical latent class variable on latent intercept and slope would cause wrong class probability estimates. Hence, the individuals could be associated to the wrong class. It is important to estimate both models with and without covariates affecting class membership and then compare the results (Muthén, 2004). As we will explain later on, in our case covariates did not demonstrate to have a particular influence on class membership.

At this point, all the useful theory concepts have been given to the reader. Therefore, it is possible to proceed with the second (and most important) part of the work. The application of these concepts to real data, concerning students' satisfaction with the course, will be shown in the next chapters.

Chapter 4

Understanding the latent structure of the data

The present chapter explores the battery of items available, in order to reveal the structure underlying the data in hand. This objective was achieved by means of Principal Component Analysis (PCA) and Factor Analysis (FA). Both these methods have been explained adequately in Chapter 2.

At first, a PCA was conducted to the data of most recent year using the eleven items available; being the twelfth item a general question on the overall satisfaction, it was not included in the analysis. Second, Exploratory and Confirmatory Factor Analyses (EFA and CFA) were conducted to explain explicitly the relationship between the observable variables and the underlying construct. For reasons given later on, two alternative scales were then obtained using a reduced set of items or particular combinations of them. The aforementioned analyses were conducted using both these reduced scales. The CFA model with improved fit to the data and explaining more variability was chosen to analyse the evolution in students' satisfaction over the last three academic years.

4.1 The measurement scale

The data is composed of 1,854 courses and it includes only the courses existing in all the considered years. Observations with missing data or evident errors were excluded in advance. Moreover, only the questionnaires filled in by regular students were taken into account. For this reason, the responses of Erasmus students and external students who are not regularly enrolled in a degree (but, for instance, are attending a single course) were excluded from the data set. In order to have a good overview of the satisfaction with teaching, only data from regular students attending at least half of the whole course was included in the analyses; otherwise results would turn to be biased. In fact, these students may not be aware of the structure of the course and the teaching, given they missed too many classes for personal reasons. As a result, their assessment of the course can be affected by other factors not so consistent with the regular teaching. For this reason, their responses may be lower than the ones by regular students, which may lead to a negative bias in the results (see, for instance, Massingham and Herrington, 2006). The further exclusion of the courses with few questionnaires filled in (i.e. three or less) would not affect the results in any way. These courses were therefore kept in the data.

Following the description given in Chapter 1, questionnaires about evaluation of didactic activities in the University of Padua are measured by a battery of twelve items. It is the subset of the questionnaire that is shared across the three years. As we said previously, the questionnaire has been changed between Academic Years 2012/2013 and 2013/2014, thus not all the items were comparable. Selecting only these twelve items, a direct comparison of the responses over time was then possible. The first eleven items deal with specific characteristics of the course, while the last one measures the overall impression of the students about the course they were attending.

A first approach to the data involves the last year available, Academic Year 2014/2015, as it contains the most recent data on students' opinion. For this reason, it is a good starting point for an improvement in didactics.

Table 4.1 shows a strong positive correlation between all the items in the battery. All the inter-item correlation coefficients are high, never below 0.67, and they are all statistically different from zero. These values suggest that items are measuring the same latent construct (students' satisfaction) and the measurement scale has internal consistency. For instance, there is a strong correlation between items 6 and 7, respectively stimulus and clearness of explanation. These two items, together, define the "Efficacy of didactics" indicator and thus they refer to teacher's skills. Their similarities are proven by the comparison of their means and standard errors, very close to each other. Item 7 presents a high correlation with item 8 too, which indicates materials suggested to students by the professor.

The scores assigned to the items by the students are quite different and that can be seen comparing the means of the ratings shown in Table 4.1. Students give lower scores to preliminary knowledge and workload, respectively items 4 and 11, compared to the other items, while they assign higher ratings to timetables (item 3). The twelfth item, students' overall satisfaction with the course, is positively associated with the other eleven items and has to be considered a "gold standard", as it assures the validity of the entire scale (DeVellis, 1991).

Only the items referring to particular aspects of the course will be used in the following analyses, while the overall-satisfaction item will be left out.

Table 4.1: Inter-item correlations, means and standard errors of the items (2014/2015).

	I01	**I02**	**I03**	**I04**	**I05**	**I06**	**I07**	**I08**	**I09**	**I10**	**I11**	**I12**
I01	1.00											
I02	0.88	1.00										
I03	0.80	0.81	1.00									
I04	0.77	0.73	0.67	1.00								
I05	0.81	0.76	0.72	0.75	1.00							
I06	0.88	0.82	0.75	0.76	0.86	1.00						
I07	0.89	0.81	0.74	0.76	0.82	0.94	1.00					
I08	0.88	0.82	0.76	0.78	0.79	0.88	0.91	1.00				
I09	0.80	0.80	0.83	0.68	0.74	0.80	0.78	0.81	1.00			
I10	0.81	0.76	0.75	0.72	0.74	0.83	0.82	0.84	0.84	1.00		
I11	0.74	0.72	0.69	0.72	0.68	0.75	0.75	0.76	0.70	0.74	1.00	
I12	0.88	0.82	0.77	0.77	0.87	0.94	0.93	0.90	0.82	0.86	0.80	1.00
Mean	7.95	8.03	8.21	7.68	8.09	7.83	7.81	7.75	8.09	7.83	7.61	7.80
Std Error	1.22	1.18	1.10	1.17	1.18	1.34	1.34	1.27	1.22	1.29	1.29	1.29

4.2 PCA and FA with eleven items on Academic Year 2014/2015

The analysis starts with a PCA and the measurement model for the most recent year. It is clear through PCA that one single component should be enough to explain a great proportion of the total variability.

In Figure 4.1 the scree plot shows the eigenvalues associated to every component. The first principal component has a very high associated eigenvalue, of almost 9. It is the only one having an eigenvalue above 1 (threshold based on the Kaiser criterion), and it retains 80.58% of the total variance in the data. All the other principal components have eigenvalues quite below the threshold: they do not add a considerable percentage to the amount of explained variance, such as to justify the choice of more than just one component.

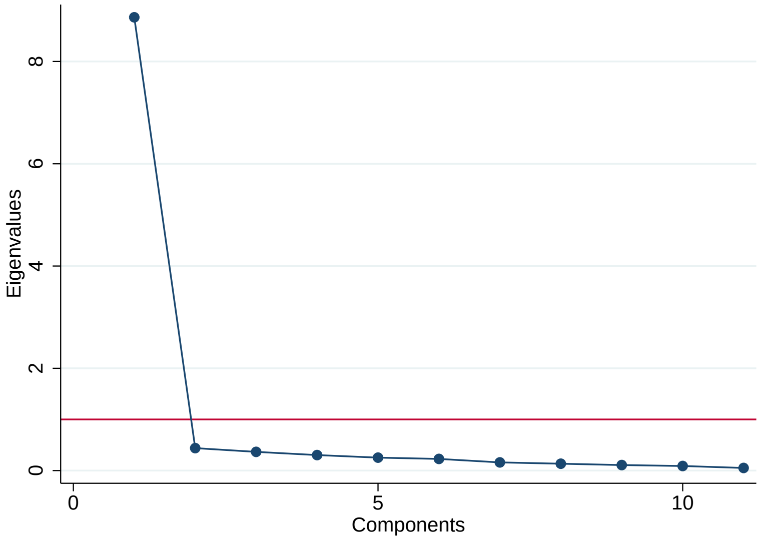

Figure 4.1: Scree plot of the eigenvalues after PCA on the total set of items (2014/2015).

In order to understand the dimensionality of the construct underlying the items, an Exploratory Factor Analysis was conducted. From the previous analysis, it was already expected to find one single factor underlying the observable indicators. Extracting the factors with methods such as the iterative maximum likelihood did not provide useful results, because a Heywood case was encountered[1]. A solution was found in extracting the factors through the principal-components factoring, which gave for obvious reasons the same results of the previous PCA. Thus, it resulted in one single latent dimension found underneath the eleven original variables, which could be simply named as “students’ satisfaction”. This solution does not confirm the results obtained in other works concerning this kind of data (see, e.g., Dalla Zuanna et al., 2015; Bassi et al., 2016). These papers deal with questionnaires collected in 2012/2013 in the same University and, partially, with the same measurement scale. It is only a partial match, as the questions posed to students in the questionnaire are more than the ones at our disposal. The previous research found out that the items in academic year 2012/13 were measuring different latent constructs. Those items were actually composing a multidimensional scale with three or

[1] A Heywood case is encountered whenever the communality between the original variable and all the other ones is equal or superior to 1. It means that the variance explained by that variable is totally shared with the other variables composing the scale. Thus its uniqueness is equal to 0. Clearly, this situation could lead to some estimation problems (Fabrigar et al., 1999).

four underlying dimensions. On the contrary, the present results show that a single dimension underlies the observed 11-items responses. Thus, the items comparable over time are composing a unidimensional measurement scale.

Table 4.2 shows the results obtained after EFA on the data. The second column provides standardized factor loadings, i.e., the correlations between the single latent factor and each of the eleven items. All of the indicators prove to have a high positive association with satisfaction, the underlying construct being measured. That is confirmed by noting that loadings range from a minimum of 0.836 (workload) to a maximum of 0.942 (aims). The adjacent column in the same Table contains the estimates of each item uniqueness. Let us remind that the level of uniqueness (i.e. the proportion of variance not shared among the items) must be low to perform an effective Factor Analysis. We conclude that the degree of uniqueness is quite low for every item and they are therefore adequate to express the latent factor all together. Additionally, all item-rest correlations are above 0.80, which confirms the internal consistency and reliability of the instrument used to measure satisfaction.

The KMO coefficients of the items are all "in the 0.90s". For this reason, using Kaiser's words (Kaiser, 1974), they can be considered "marvelous". Thus, the items are really adequate to describe the latent factor.

Table 4.2: Factor loadings and indices using the eleven-items solution (2014/2015).

Item	**Loading**	**Uniqueness**	**Item-rest Corr.**	**KMO**	**Alpha**
I01 Aims	0.942	0.114	0.927	0.961	0.972
I02 Exam	0.904	0.184	0.882	0.960	0.973
I03 Timetable	0.863	0.255	0.835	0.959	0.974
I04 Prel. knowl.	0.843	0.289	0.812	0.970	0.975
I05 Interest	0.880	0.226	0.854	0.961	0.974
I06 Stimulus	0.940	0.117	0.924	0.928	0.972
I07 Clearness	0.935	0.125	0.919	0.929	0.972
I08 Material	0.936	0.124	0.920	0.965	0.972
I09 Office hours	0.890	0.207	0.867	0.948	0.974
I10 Laboratories	0.897	0.195	0.874	0.967	0.973
I11 Workload	0.836	0.301	0.804	0.977	0.975
Total				0.956	0.976

The overall value of Cronbach's alpha index shown in Table 4.2 equals 0.976, indicating a desirable level of internal consistency of the items composing our measurement scale. The other alpha coefficients, each of them associated to a single item, are respectively the index calculated if the item is removed from the scale. In this case, each elimination would lead to a decrease in the alpha value. This fact attests that each item contributes to raise the scale consistency. However, alpha value does not experience a considerable reduction, in particular dropping one between item 4 (preliminary knowledge) or item 11 (workload).

After EFA, it is confirmed the presence of only one latent factor underlying the items. It is reasonable to believe this latent factor to be students' overall satisfaction with the course they attended. Assuming the structure of the model, i.e. eleven items measuring a single latent trait, it is possible to start with a first Confirmatory Factor Analysis.

Table 4.3 presents the results of the factorial model. Having set the variance of the latent factor to 1, the standardized coefficients vary between -1 and 1. All the coefficients are highly statistically significant, proving once again their strong relation with satisfaction. They are all positive and close to 1, which means that the score given to the items increases with the increase in the overall satisfaction with the course. A higher intercept, called "constant" in this case, means a higher score when the course is considered averagely satisfactory. That is because the latent variable is assumed to be distributed as a normal with mean zero. A higher slope coefficient (i.e. the factor loading) indicates a faster increase in the score assigned to the item, as satisfaction rises. Hence, higher loadings are signs of stronger relations with the latent construct.

One of the most remarkable figure in Table 4.3 is the constant of the third equation, the one concerning timetables. It has a value considerably above all the other intercepts. This value is consistent with the mean of the same item, shown previously in Table 4.1, which was higher than all the other items mean. However, the slope coefficient in the same part (the one related to timetables) is one of the lowest. Workload is the item with the lower coefficients: when students are averagely satisfied (factor equals zero), they assign to the twelfth item a bad score; the same score does not grow so much, even when in fact students' satisfaction does.

The coefficients of stimulus and clearness of explanation, respectively items 6 and 7, are very similar to each other (it was visible looking at the items mean and standard error in Table 4.1). It suggests that students may consider these items meaning really similar to each other. In fact, teachers who are able to explain clearly their subjects are usually the ones stimulating more interest towards the topic in the students. Though these items are characterized by low intercepts (and, therefore, means), they show a considerable slope, similar to the one of aims (item 1) and materials (item 8).

Table 4.3: CFA coefficients using the eleven-items solution (2014/2015).
Levels of significance: p-value<0.001 ∗∗∗; 0.001<p-value<0.01 ∗∗; 0.01<p-value<0.05 ∗

	Coefficient	**Standard Error**
Aims	0.941 ∗∗∗	0.003
Constant	6.535 ∗∗∗	0.110
Exam Arrangement	0.888 ∗∗∗	0.005
Constant	6.787 ∗∗∗	0.114
Timetables	0.832 ∗∗∗	0.007
Constant	7.435 ∗∗∗	0.124
Preliminary knowledge	0.818 ∗∗∗	0.008
Constant	6.560 ∗∗∗	0.110
Interest	0.868 ∗∗∗	0.006
Constant	6.851 ∗∗∗	0.115
Stimulus	0.947 ∗∗∗	0.003
Constant	5.850 ∗∗∗	0.099
Clearness	0.947 ∗∗∗	0.003
Constant	5.842 ∗∗∗	0.099
Material	0.937 ∗∗∗	0.003
Constant	6.102 ∗∗∗	0.103
Office hours	0.865 ∗∗∗	0.006
Constant	6.638 ∗∗∗	0.111
Laboratories	0.880 ∗∗∗	0.005
Constant	6.060 ∗∗∗	0.102
Workload	0.806 ∗∗∗	0.008
Constant	5.911 ∗∗∗	0.100

The previous model attested the effective validity of the items to measure the underlying factor, but actually it shows some fitting problems. For example, the Likelihood Ratio statistics is equal to 2,072.77 and, under a chi-square distribution with 44 degrees of freedom, the null hypothesis is clearly rejected. Thus, the model with eleven items is not fitting the data very well and for this reason it should be modified. This lack of fit is due to the fact that the model does not consider co-varying error terms, while the residual correlations among the errors are still quite strong and not explained by the model. Too many highly correlated items were used in the previous analyses. In fact, it might be sufficient to explain the single latent factor through a subset of the eleven items, as they are clearly overlapping and strongly correlated.

4.3 Two solutions with reduced sets of items

Two solutions are here proposed, in order to avoid the redundant usage of eleven items for just a unidimensional scale.
The first one was obtained considering only five items out of the original eleven. One by one, all the items with the highest residual correlations were removed from the model (backward elimination) by means of their Modification Indices. The selected items are:

1. **Item 3**: Timetables;
2. **Item 4**: Preliminary Knowledge;
3. **Item 6**: Stimulus;
4. **Item 8**: Material;
5. **Item 10**: Laboratories.

As it was said before, the original items are strongly related among each other. For example, items 6 and 7, regarding respectively stimulus and clearness of explanation, are very similar and it was confirmed looking at the output of the CFA model (Table 4.3). To stress the concept, teachers using a clear teaching method are likely to be the ones who stimulate the students more. Thus, a good teacher could even increase the interest of students towards the subject taught, relating these two items to the fifth one (interest of the student). Furthermore, a teacher could stimulate the students providing good material for the lessons or spending some hours in useful laboratories. Moreover, even clearness of explanation could be influenced by efficient study material provided.
The way students feel about the workload requested (item 11) might be influenced by their preliminary knowledge of the subject (item 4). If students have a good background preparation before attending the course, they will consider the workload adequate to the number of credits assigned to that specific course. This connection could explain why the scores given by students to these two items are strictly close.

The aims of the course (item 1) are strongly linked to item 8 (material) too. In fact, many teachers often use the first lesson and the first slides to clarify the aims of the topics they are about to lecture throughout the course.

The opinion on office hours (item 9) is linked to timetables (item 3). If teachers do not respect the time they had planned to spend in the office to handle students' doubts, their image could be compromised. Therefore, students will tend to give a bad rating even on the question about lessons timetable, as both of these items deal with teacher's schedule reliability.

The exam arrangement (item 2) is related to laboratories (item 10), because often labs are used by teachers to solve with the students many exercises from the past exam sessions. Furthermore, in specific degrees many classes are taught in computer or science labs, in which students can apply what they learned in theoretical classes. In most of the cases, the exam includes both theoretical and practical parts. Hence, it is even through laboratories that students are capable to understand how their exam will be structured.

All these logical connections make it clear the redundancy in the use of eleven items to explain a single latent dimension. Therefore, they justify the use of reduced sets of items.

A second solution to reduce the number of the items used in the analyses was found in literature. Several items were combined to form two indicators published yearly by the University of Padua in order to indicate the level of its students' satisfaction. These two indicators deal with "Organizational Aspects" (OA) and "Efficacy of Didactics" (ED). They are obtained averaging respectively item 1 (aims), item 2 (exam arrangement), item 3 (timetable), item 8 (material) and item 6 (stimulus), item 7 (clearness). The validity and reliability of these instruments were already established (see, for instance, Bassi et al., 2016). Using these indicators instead of the items they are composed of, the number of measures used decreased from eleven to seven. However, the residual correlations among the remaining items were still quite high. A further reduction was made, eliminating once again the items showing the greatest residual correlations. The final choice is the following:

1. **OA**: Mean of items 1, 2, 3 and 8;

2. **ED**: Mean of items 6 and 7;

3. **Item 4**: Preliminary Knowledge;

4. **Item 10**: Laboratories;

5. **Item 11**: Workload.

Thus, only items 5 and 9 (interest and office hours) were excluded with this solution. OA and ED indicators include the information of many items, but it has to be considered that an items average leads inevitably to the loss of part of the information. Clearly, it would be desirable to lose the smallest information possible.
In the following Section 4.4, the analyses done so far will be reproduced according to these new sets of items. The aim is to find the model that fits the data best.

4.4 Factor Analysis with reduced sets on Academic Year 2014/2015

The two reduced sets of items both consist of five items. The first set, let us name it "Model 1", contains items 3, 4, 6, 8 and 10, whereas the second set, named "Model 2", covers the indicators OA and ED and items 4, 10 and 11. As expected, the EFA using these two solutions did not produce different results from the ones with the original set of items. The factors extracted were obtained, as before, through principal-components factoring. The scree plots shown in Figure 4.2 confirm that there is only one latent dimension, since in both cases one single eigenvalue is higher than 1. The eigenvalue of the first factor is significantly lower than the full model one. That is an obvious consequence of the reduction of complexity, because the sum of the eigenvalues is equal to the number of items.

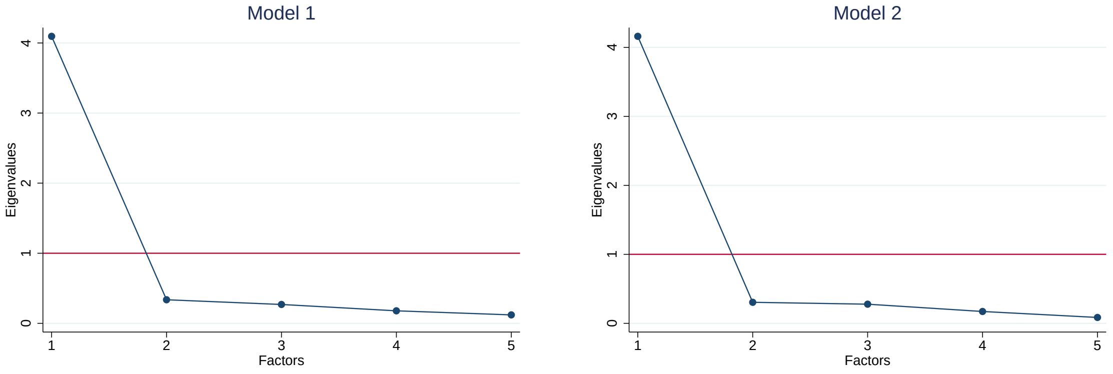

Figure 4.2: Scree plots of eigenvalues after EFA on Model 1 (left) and Model 2 (right) (2014/2015).

The first factor of Model 1 catches 81.90% of the total variance explained by the five items of the set. It represents a little increase compared to the initial model, whose first factor could explain more than the 80% of the total variance. The first factor of Model 2 is able instead to explain 83.21% of the total variability, showing a good improvement.

Table 4.4 shows the factor loadings for both models, together with extra measures of reliability. The factor loadings associated to the items are still very high. In most cases they are even higher than the same coefficients obtained using the whole scale. The indicators OA and ED show remarkable connections with the latent factor. The loading linked to the first one is the highest of all, with a value above 0.95. It confirms that the correlation existing between the organizational aspects of the course and the latent trait is strong. The loading associated to ED is very similar to the ones of items 6 and 7, the measures composing the indicator, expressed in Table 4.2. They previously showed to have similar loadings between each other.

Table 4.4: Factor loadings and indices using the five-items solutions (2014/2015).

Item	**Loading**	**Uniqueness**	**Item-rest Corr.**	**KMO**	**Alpha**
			Model 1		
I03 Timetable	0.868	0.247	0.797	0.945	0.941
I04 Prel. Knowl.	0.864	0.254	0.791	0.946	0.942
I06 Stimulus	0.933	0.130	0.889	0.882	0.924
I08 Material	0.942	0.112	0.904	0.865	0.921
I10 Laboratories	0.915	0.163	0.863	0.914	0.929
Total				0.906	0.944
			Model 2		
OA	0.954	0.090	0.923	0.843	0.926
ED	0.941	0.115	0.902	0.865	0.930
I04 Prel. Knowl.	0.876	0.233	0.809	0.946	0.946
I10 Laboratories	0.910	0.172	0.856	0.932	0.938
I11 Workload	0.878	0.230	0.812	0.947	0.946
Total				0.900	0.949

The uniqueness is very low for all the items and we remind this is desirable in a Factor Analysis context.

The item-rest correlation clarifies the strict relation among every measure and the rest of the items composing the measurement scale.

The KMO index is still high, even with the reduction made to the original set of items. The items are still adequate, but now some of them present an index “in the

0.80s" (Kaiser, 1974) in both sets.

The Cronbach's alpha indices show high values for the two reduced models. In both sets, the further elimination of the remaining items would compromise the value of the overall alpha. OA and ED, which were associated to the lowest KMO indices in Model 2, have the lowest Cronbach's alphas too. It should be reminded that it is not a good sign for a KMO value to be low, as it is not good for a specific alpha to be greater than the overall alpha. In this case, Cronbach's alpha implies it could be worth to use OA and ED indicators to describe students' satisfaction.

The predicted scores of these factor analyses are very similar to the ones with the full set of items. The correlation between 11-items and Model 1 predictions is 0.991, while the one between the full set and Model 2 predictions is 0.990. It is clear that both these choices lead to the same conclusions in terms of measuring satisfaction. The loss of information resulting from the simplification of the model is thus negligible.

After running EFA, results confirmed once again that there is only one dimension underlying the observable variables. A CFA was run in order to understand how the items composing the two selected sets are affected by satisfaction. Results for Models 1 and 2 are shown in Table 4.5. The items reduction does not influence the estimates of the constants or their standard errors, but only the coefficients related to factor loadings. All the measures used before in the eleven-items model confirmed to have similar loadings estimations in the reduced models. The "Organizational aspects" indicator has a high intercept (and mean) and even a high slope, proving a strong correlation with the latent factor. The other indicator, associated to "Efficacy of Didactics", presents a high loading estimation, but a low intercept. However, these values are both consistent with the ones of items 6 and 7 in the full model.

Table 4.6 presents several measures indicating the goodness of fit relatively to the three proposed models. The Likelihood Ratio (LR) test proved how poorly the first model (the complete one) could fit the data. This is due in particular to the significant correlations that are still present among the items residuals. Reducing the number of the items had advantages in both the examined cases. The null hypothesis is still rejected, but the LR statistics experienced a notable decrease reducing the complexity of the measurement scale. This scale was clearly containing overlapping variables. The complete model has an unacceptable Root Mean Squared Error of Approximation (RMSEA), equal to 0.158. The reduced models present lower levels of error, in particular Model 1. In addition, the same Table shows for every model the probability of having a RMSEA below 0.05. Model 1 is the only one having a significant possibility to have an error inferior to that threshold.

Table 4.5: CFA coefficients using the five-items solutions (2014/2015).
Levels of significance: p-value<0.001 ∗∗∗; 0.001<p-value<0.01 ∗∗; 0.01<p-value<0.05 ∗

	Coefficient	**Standard Error**
	Model 1	
I03 Timetables	0.817 ∗∗∗	0.008
Constant	7.435 ∗∗∗	0.124
I04 Prel. Knowledge	0.817 ∗∗∗	0.008
Constant	6.560 ∗∗∗	0.110
I06 Stimulus	0.927 ∗∗∗	0.004
Constant	5.850 ∗∗∗	0.100
I08 Material	0.943 ∗∗∗	0.004
Constant	6.102 ∗∗∗	0.103
I10 Laboratories	0.891 ∗∗∗	0.005
Constant	6.060 ∗∗∗	0.102
	Model 2	
OA	0.964 ∗∗∗	0.003
Constant	7.174 ∗∗∗	0.120
ED	0.943 ∗∗∗	0.003
Constant	5.939 ∗∗∗	0.100
I4 Prel. Knowledge	0.823 ∗∗∗	0.008
Constant	6.560 ∗∗∗	0.110
I10 Laboratories	0.884 ∗∗∗	0.006
Constant	6.060 ∗∗∗	0.102
I11 Workload	0.820 ∗∗∗	0.008
Constant	5.911 ∗∗∗	0.100

According to Akaike Information Criterion, shown again in Table 4.6, the natural choice is Model 2.

Comparing the Comparative Fix Index, both the reduced measurement scales are quite good in fitting the data and show an improvement compared to the original scale. Even in the case of Tucker-Lewis Index, Models 1 and 2 have a better fit than the complete model.

The Standardized Root Mean Squared Residual (SRMR) is clearly better in the

reduced models than in the complete one and it is the lowest in Model 1.
Looking at the Coefficient of Determination (CD), Model 2 is preferable to Model 1, but the models show all a good fit.
According to the Cronbach's alpha of reduced models, the items of Model 2 benefit from a higher internal consistency. The Average Variance Extracted (AVE) values indicate for all the models the validity of the latent and observable variables, just as the Composite Reliability (CR) indices demonstrate the reliability of all the adopted scales.

Table 4.6: Goodness-of-fit indices for the three models (2014/2015).

Index	**Full Model**	**Model 1**	**Model 2**
LR test (df)	2072.765(44)	33.490(5)	69.823(5)
RMSEA	0.158	0.055	0.084
Prob RMSEA$<$ 0.05	0.000	0.277	0.001
AIC	40560.152	21284.995	20526.883
CFI	0.927	0.997	0.993
TLI	0.909	0.994	0.987
SRMR	0.026	0.007	0.012
CD	0.981	0.957	0.967
Cronbach's alpha	0.976	0.944	0.949
AVE	0.785	0.776	0.790
CR	0.976	0.945	0.949

Thus, Table 4.6 provides evidence to consider the reduced models as an improvement of the initial full model. The small information loss experienced from a 11-items model to a 5-items one is negligible, compared to the huge reduction of the scale complexity. The goodness-of-fit indices here presented do not lead to a unique choice of the best model, but they are almost evenly split between Models 1 and 2. Therefore, it has been demonstrated that both the reduced solutions could be adequate to study students' satisfaction. We decided to keep the solution containing "Organizational aspects" and "Efficacy of didactics" indicators, Model 2. It would be useful to take the more information possible and Model 2 seems to be the right choice to achieve this aim. The second reduced set of items will be used in the next Section 4.5 in order to reproduce Factor Analysis in the three Academic Years under study. This will confirm similarities or differences among the years and will allow to proceed then with further statistical models.

4.5 CFA on the three Academic Years available

The replication of PCA and FA on the first year available, the academic year studied in previous papers, brought to the same conclusion of the most recent year. It is the same even for the second year. Thus, just one latent dimension underlies the measurable indicators.

Nevertheless, it is useful to replicate the previous studies. We remind again that the items used here to analyse the change in students' satisfaction represent a subgroup of the questions that students are actually asked to answer to. The questions are yearly revised and can be eliminated or changed from one year to another. Therefore, dropping the differing items was a necessary step to be done, in order to ensure a comparison between the answers of different years. The substantial difference between our study and the previous ones is that the eliminated items were the ones making the scale multidimensional. Even the choice to drop all the observations with at least one missing value may have had unidimensionality as a consequence.

The analyses done so far were referring exclusively to the last academic year available. They were made with the purpose of choosing an adequate measurement scale that could allow to find the model fitting the data best. It could be better to recall the items that are composing this scale: they are the indicator "Organizational aspects" (average of items 1, 2, 3 and 8), indicator "Efficacy of didactics" (average of the items 6 and 7) and items 4, 10 and 11 (respectively preliminary knowledge, laboratories and workload).

The confirmatory factorial model used before was applied, this time in order to discover similarities and especially differences between the three academic years. The coefficients of the three models can be seen in Table 4.7.

As it was expected from the previous analyses, all the loadings associated to the five items are strongly significant and close to 1 for each academic year. It is notable that the factor loadings have generally increased over the last three years. In particular, OA and ED demonstrate to have a strict relation with the latent dimension, since each of their loadings is considerably above 0.90. The only exception is represented by laboratories loading, which is lower in the second year than in the first one. However, this loading have increased in the last year, with a value quite close to the first-year one. In any case, it is confirmed the strong relation connecting each item to the underlying construct. Thus, an increase of the loading coefficient can be considered as an improvement from the measurement point of view, since the item is more linked to the latent factor than before.

Table 4.7: CFA coefficients using Model 2, comparison of the three years.
Levels of significance: p-value<0.001 ∗∗∗; 0.001<p-value<0.01 ∗∗; 0.01<p-value<0.05 ∗

	2012/2013		**2013/2014**		**2014/2015**	
	Coefficient	**Std Error**	**Coefficient**	**Std Error**	**Coefficient**	**Std Error**
OA	0.941 ∗∗∗	0.004	0.955 ∗∗∗	0.003	0.964 ∗∗∗	0.003
Constant	8.247 ∗∗∗	0.137	7.442 ∗∗∗	0.124	7.174 ∗∗∗	0.120
ED	0.930 ∗∗∗	0.004	0.938 ∗∗∗	0.004	0.943 ∗∗∗	0.003
Constant	6.490 ∗∗∗	0.109	6.104 ∗∗∗	0.103	5.939 ∗∗∗	0.100
Prel. Knowl.	0.745 ∗∗∗	0.011	0.799 ∗∗∗	0.009	0.823 ∗∗∗	0.008
Constant	7.126 ∗∗∗	0.119	6.568 ∗∗∗	0.110	6.560 ∗∗∗	0.110
Laboratories	0.891 ∗∗∗	0.006	0.861 ∗∗∗	0.007	0.884 ∗∗∗	0.006
Constant	7.120 ∗∗∗	0.119	6.200 ∗∗∗	0.104	6.060 ∗∗∗	0.102
Workload	0.768 ∗∗∗	0.102	0.807 ∗∗∗	0.009	0.820 ∗∗∗	0.008
Constant	6.493 ∗∗∗	0.109	6.037 ∗∗∗	0.102	5.911 ∗∗∗	0.100

The increase of factor loadings over time does not ensure that students' satisfaction is increasing. In fact, the intercept values of all the items, without exception, are constantly decreasing. This fact denotes that students are averagely giving lower and lower ratings to the items. However, their evaluation grows faster than before as their satisfaction rises, considering the increasing slopes (loadings) over time.

In Table 4.8 the usual goodness-of-fit indices are presented. They were calculated after the application of Model 2 to each of the three academic years. The results were rather satisfactory, as they were just for the last academic year available.

The LR test value in the second year is noteworthy, since it is quite below the ones of the other years. The Root Mean Square Error of Approximation is good for each time, but again especially in the second period. With a value of 0.068, the second-year RMSEA is the only one having a concrete possibility to be below 0.05 (even if this chance is still under the 5%).

All the other indices are additional signs of the model good fit, as they are far beyond their threshold, which were adequately expressed in the previous Chapter 2.

To conclude, the analyses presented thus far attested that only one single latent dimension is underlying the measurement scale under study. However, the original measurement scale, composed of eleven items, was redundant for analysing just one latent construct. Due to many items overlapping, the model resulting from the scale did not fit the data well. Thus, two reduced scales, of five items each, were later proposed. The models estimated using these scales proved to fit the data better than the full model with the whole set of items.

Table 4.8: Goodness-of-fit indices for Model 2 on the three years.

Index	**2012/2013**	**2013/2014**	**2014/2015**
LR test (df)	68.222(5)	49.592(5)	69.823(5)
RMSEA	0.083	0.069	0.084
Prob RMSEA< 0.05	0.001	0.030	0.001
AIC	20113.519	21069.553	20526.883
CFI	0.992	0.995	0.993
TLI	0.985	0.990	0.987
SRMR	0.012	0.011	0.012
CD	0.954	0.960	0.967
Cronbach's alpha	0.932	0.941	0.949
AVE	0.738	0.765	0.790
CR	0.933	0.942	0.949

Different goodness-of-fit indices were then calculated, in order to decide which model would be preferable. Finally, it was possible to analyse the differences between the last three academic years by means of a factorial model with the preferred items set. The intercepts of the items have experienced a constant decrease in the last years. It means that students are averagely giving lower scores to the items than in the past. At the same time, the items loadings have slightly increased over time. Hence, if satisfaction gets higher, students' ratings grow faster than in the previous years. However, this model allowed just a cross-sectional study of the data, as the three years were analysed separately.

Once the latent structure of the data has been verified, the further step is a longitudinal analysis. It will provide more detailed and accurate information on how students feel about university courses. Moreover, it will be clearer the change of students' satisfaction over time. The results of the longitudinal analysis, accomplished through latent growth modelling, will be subject of the next and final chapter of this work.

Chapter 5

Evolution of students' satisfaction in the last three academic years

The aim of this work was to establish the changing in students' satisfaction with the academic courses in the context of the University of Padua. The results shown in the previous Chapter 4 allowed us to be conscious of the latent structure of the measurement scale. It is a crucial step to proceed with further analyses on the same data. The results attested that only one latent dimension, i.e. satisfaction, underlies our scale. Moreover, a preliminary knowledge of students' satisfaction level was provided, separately for each year at our disposal. Let us remind that only a reduced scale of measures was used, instead of the set of eleven items originally available. This solution was taken in order to avoid the unavoidable overlap a eleven-item set was causing. The scale is composed of "Organizational aspects" and "Efficacy of didactics" indicators (respectively obtained from the average of items 1, 2, 3, 8 and 6, 7), items 4 (preliminary knowledge), 10 (laboratories) and 11 (workload).

The present and final chapter will be focused on a longitudinal analysis, studying simultaneously the three years. This analysis will be provided by means of latent growth models, which were introduced in Chapter 3 and are particularly fit for our purpose. At first, an unconditional second-order latent growth model was run considering just the reduced scale of indicators. Lacking of any covariate, this model is the simplest one and it has to be considered just as a starting point for other more complex models. Before exploring more deeply students' satisfaction, a descriptive analysis of the available covariates will be given. It is important to know what kind of course- and teacher-related variables could affect satisfaction. These covariates were used in the application of the following models. The conditional model is the natural extension of the previous unconditional one, with the introduction of covariates. In this case, all the observations are still considered coming from the same population, with just one latent trajectory. The latent growth mixture model is a further extension of the conditional one. Taking into account covariates, it divides the observations into different groups, each with its own latent trajectory. These two last models will be shown together, in order to catch immediately eventual similarities and differences. Thus, it will be easier to understand if there are actual different growth patterns among the courses or if they follow the same change trajectory.

5.1 The unconditional second-order LG model

The unconditional model here presented is perceived as the beginning of our longitudinal analysis. It is in fact a model with one measurable level, constituted by the items, and two different latent levels. The first level is composed of three latent constructs, one for each year, which are linked to the measures collected in the corresponding time. The second level is represented by the so-called growth factors (i.e. latent intercept and slope), which are related to the first latent level. Being the simplest model, it does not provide for covariates affecting the growth factors.
This growth model, just as the following ones, requires measurement invariance over time. Thus, factor loadings are equal in the different years. Furthermore, the first loading (the one associated to "Organizational aspects") is fixed to one and it is the same for the items residual variances, while all residual covariances are set to zero. This constraints choice was used also in other works (see, for instance, Bassi and Dias, 2013).
Our sample size was reduced for the following analyses from the initial 1,854 to 1,843 observations. Such a reduction is due to few rare roles of professors which were not included. However, it will be motivated better later on, when we will deal with the several covariates available in the data set. We can affirm that a reduction of 11 observations in the whole population (0.59%) would not affect the previous results in any way. For this reason, we decided to keep the complete sample to perform Factor Analysis.
The estimates of the unconditional model are presented in Table 5.1. In the previous analysis we found out that factor loadings have decreased in the last years. Through this model we assume instead to have loadings time-invariance. We suppose therefore that the same item will not change its relation to the underlying dimension over time. This is plausible and logical. After CFA, we were inclined

Table 5.1: Estimation of the unconditional second-order latent growth model.
Levels of significance: p-value<0.001 $***$; 0.001<p-value<0.01 $**$; 0.01<p-value<0.05 $*$

	Estimate		**S. E.**
Loadings			
OA	1.000		$---$
ED	1.188	$***$	0.008
I04	0.896	$***$	0.010
I10	1.074	$***$	0.010
I11	0.989		0.011
Residual Variances (1$^{\text{st}}$ level)			
θ_1	0.472	$***$	0.044
θ_2	0.599	$***$	0.028
θ_3	0.578	$***$	0.050
Covariance (2$^{\text{nd}}$ level)			
ψ_α	0.388	$***$	0.045
ψ_β	0.037		0.022
$\psi_{\alpha\beta}$	-0.001		0.025

to think that OA indicator was the measure with the highest loading of all. The current model shows it is not as it seemed. Taking OA as the baseline, Table 5.1 shows loading significances in comparison to the reference measure. ED indicator and item 10 (laboratories) proved to have higher loading values than the baseline and these differences are both statistically significant. As to item 4 (preliminary knowledge), it has a statistically significant loading, which is lower than the OA one. Item 11 (workload) did not prove to have a statistically-different loading than the reference. The residual variances θ_t, related to the first- level latent constructs η_t (for $t = 1, 2, 3$), are statistically different from zero. This is sign of a lack of fit and the model could be improved. Nevertheless, the second-level variances lead already to interesting considerations. The variance of the latent intercept α, namely ψ_α, is statistically significant. It means that not all the courses have the same initial level. We remind that α represents the point where the latent curve begins. Thus, some courses have higher scores than others since the beginning of the period we are analysing. Both growth factors covariance and latent slope variance, respectively $\psi_{\alpha\beta}$ and ψ_β, are not statistically significant. This fact means that the growth factors are not related, since the intercept varies across individuals (courses) while the slope is not. In fact, the courses share the same growth rate which is null, because the

means of all the latent variables are set to zero in this model.
The unconditional model provided a first impression on satisfaction evolution in time. However, this model is rudimentary and might actually be improved introducing some covariates, which could influence the satisfaction of students.

5.2 Descriptive statistics on the covariates

Together with student's answers to the questionnaire, other information is always collected. The present section is focused on a preliminary study of this kind of information, necessary to proceed with our analysis. It will allow a better comprehension of the university context in which courses are set. University of Padua gathers in particular information about teachers and courses, while information about single respondents is not considered.

5.2.1 Number of questionnaires filled in

The number of questionnaires filled in per course every year do not represent actually the effective class size, because the first quantity is always smaller than the second one. This is due to many attending students who decide not to compile the questionnaire or to students having not attended enough lectures. We remind that our sample was made collecting only the answers of effective attending students, who took part in more than 50% of the classes. Nevertheless, we can consider the number of questionnaire collected per course as a proxy of the real class size, which is much more difficult (if not impossible) to collect. It is likely to collect more questionnaires concerning big courses and, on the contrary, to receive less answers for small-sized ones. This covariate represents the only time-variant variable at our disposal. In conditional models, it will affect therefore the first-level latent constructs, i.e. satisfaction in the three years. In our case, a total of 61,488 questionnaires were collected. Removing the eleven courses which have a rare professor role, this number decreases to 61,252, still quite high. The answers available are split evenly among the academic years they refer to.
The box plots shown in Figure 5.1 demonstrate that the average number of completed questionnaires per course are approximately the same, regardless of the year, and it is slightly above ten. Many outliers are present for each time, sign of several big-sized (thus extremely attended) courses.
The years under study show similar distributions for the number of compiled ques-

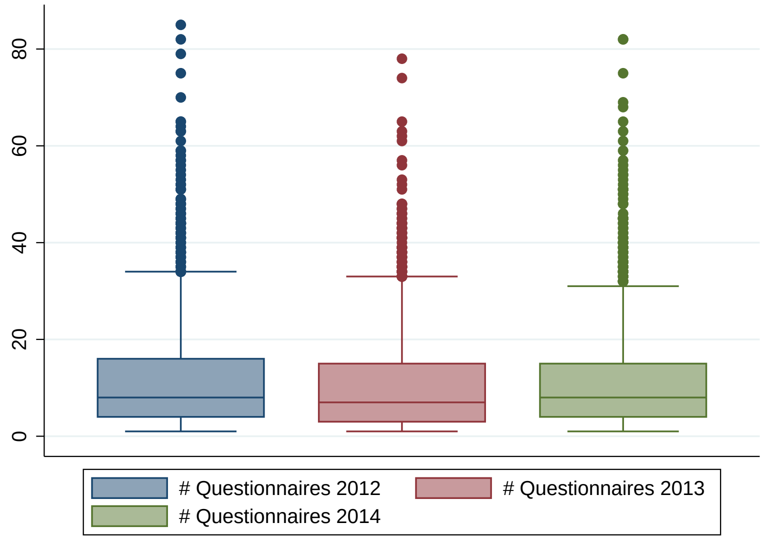

Figure 5.1: Box plots of the number of questionnaires filled in for each year.

tionnaires and this is shown in Table 5.2. The majority of the observations have five or less students answering (between 35 and 40% every year) and the percentage of respondents decreases drastically after that threshold. Suffice it to say that only about 16% of the courses have more than 20 respondents at least in one year. Despite of few answers in most cases, we will take into account the entire sample of 1,843 observations. We noticed that keeping courses with low response rate did not lead to different results.

Table 5.2: Distribution of the number of questionnaires per academic year. Percentages in the 4th and 5th columns are referred to the total of each year.

Academic Year	# Quest. (%)	Mean	# Q. <5 (%)	# Q. >20 (%)
2012/2013	21041 (34.35%)	11.42	730 (39.61%)	312 (16.93%)
2013/2014	19583 (31.97%)	10.63	723 (39.23%)	293 (15.90%)
2014/2015	20628 (33.68%)	11.19	640 (34.73%)	287 (15.57%)

5.2.2 Schools

The University of Padua has eight schools in which the different degrees and courses are grouped. For privacy reason, we cannot name them and thus we will refer to them using numbers from 1 to 8. Just as the variables that will be expressed in the following subsections, schools covariate is time-invariant. In alphabetical order, the eight schools are:

- Agricultural Sciences and Veterinary Medicine,
- Economics and Political Science;
- Engineering;
- Human and Social Sciences and Cultural Heritage;
- Law;
- Medicine;
- Psychology;
- Science.

Figure 5.2 expresses the proportion of courses grouped by school. Schools 4 and 5, respectively 20.29% and 25.23% of the total, are the most numerous (in terms of number of courses provided) and attended ones. Schools 1, 7 and 8 are the ones in the middle, with a proportion between 12.10% and 17.47% each. The other schools, namely 2, 3 and 6, are the smallest, including only $3-4\%$ of the total classes. Thus, the differences are evident, at least with regard to the distribution of the courses

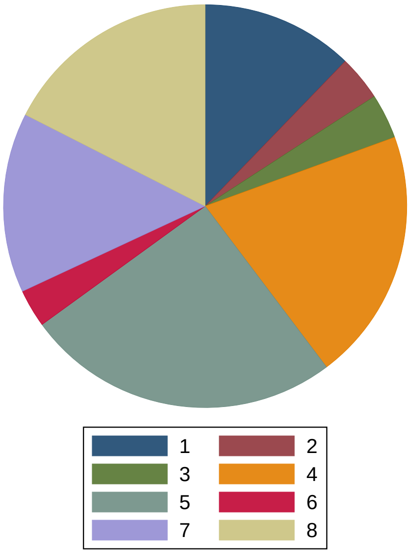

Figure 5.2: Pie chart of the courses per school.

among the eight schools.

In modelling, we wanted to understand whether the school influences students' satisfaction. In particular we chose school 5, the most numerous one, as the baseline. Thus, we verified if the other seven schools affected the growth factors differently than the fifth one. The variable was introduced in the conditional models through seven dummies, one for each school except the reference one. As we will explain later on, the dummies were not sufficiently significant to justify their use in the models.

5.2.3 Kind of degree

Our sample was collected excluding in advance the answers of students who paid to attend only a single course or who come from abroad, i.e. Erasmus students. Thus there are only three categories of degree available, namely bachelor's, master's and five-years-long degrees. Figure 5.3 shows the division of courses among the different kinds of degrees. It is notable that more than half of the courses are taught during bachelor's degree (57.25%), followed by master's (25.50%) and then five-years-long degree (the remaining 17.25%).

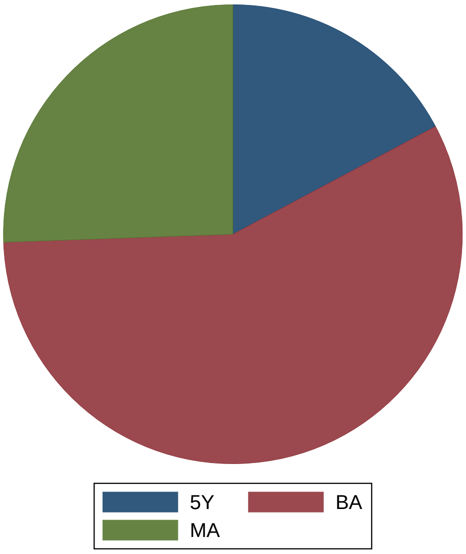

Figure 5.3: Pie chart of the courses per kind of degree.

In our conditional models, we took as a reference the largest category, in this case Bachelor. Therefore, through those models we will show different patterns in the change of satisfaction, according to the kind of degree the course is included into.

5.2.4 Borrowed courses

A borrowed course is a course available within a certain school, even though it is taught in another one. Usually it is more common to have students participating in classes from a degree within the same school of their own degree course. However, in our study we did not consider this situation. In our sample there are several borrowed courses, in the way we defined them. They are 206 out of the total 1843, more than 11%. They are not divided equally among the schools and the kinds of degree, as Table 5.3 shows. School 4 has the largest number of borrowed courses, a third of the its total. All the other schools have proportions of borrowed courses below 10%, with three of them not having any course of this kind (namely 1, 3 and 6). In modelling, if the covariate indicating a borrowed course is significant, it could even influence the significance of the school covariate. Noting the proportions, it could happen especially for school 4.

Table 5.3: Borrowed courses per school and kind of degree. Percentage is referred to the total of each school/degree.

	School								**Kind of degree**		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**BA**	**MA**	**5Y**
Freq.	0	3	0	126	39	0	7	31	147	0	59
Percent	0.00	4.23	0.00	33.69	8.39	0.00	2.61	9.63	13.93	0.00	18.55

The empirical analysis of the data set makes it clear that borrowed courses are mainly present in five-years-long and then Bachelor degrees. In the present study, none of the borrowed courses belongs to Master degrees.

5.2.5 Hours and ECTS

The hours of didactic activity represent the total amount of time dedicated to teaching in a course. Thus they are the sum of the hours of lecturing in the classroom, of practical lessons in laboratories and of all the other teaching arrangements. One ECTS corresponds, in Italy, to twenty-five hours of total workload. Total workload means in this case both teaching and individual study. It is therefore obvious for these two variables (hours and ECTS) to be strictly related to each other. Figure 5.4 expresses clearly their strong connection. From the scatter plot, we can notice a linear relation, with some outliers in the middle and right parts of the graph. In the

sample, ECTS variable ranges from 1 to 15, while the number of hours varies from 8 to 348. A wide range of hour variability is especially for courses with 7 to 9 and with 15 ECTS assigned. In the first case, such a variability might be due to the high number of courses with a "middle" number of credits. These activities could be also traineeships, often required at the end of a degree and before graduation. Activities of this kind usually require hundreds of hours to be accomplished. It is the same for activities with 15 credits, which are less common though.
The strong relation between the two covariates makes the simultaneous use of them redundant and unnecessary. Actually, this could lead to biased estimates and therefore to misinterpretation of results. To avoid this situation, we decided to keep only one of them as a covariate for the next latent growth models. We chose to select hours variable, since it has a larger range and is thus nearer to a continuous variable than ECTS.

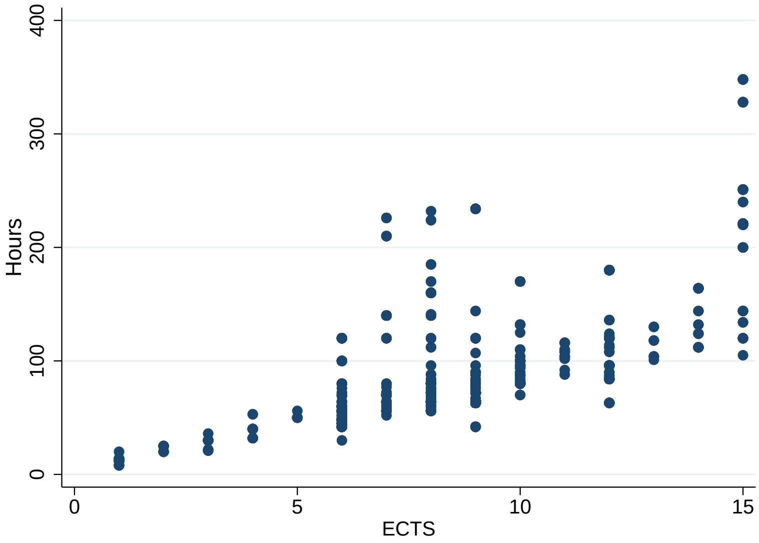

Figure 5.4: Scatter plot of the relation between ECTS and hours of didactic activity.

5.2.6 Role of professors

In the University of Padua there are mainly four different roles of professors. Teachers can be full, associate, assistant or external professors. In the data set, other (very rare) roles were initially present and were categorised as "Others". We decided to exclude them to conduct latent growth analysis, because it was not worthy to keep a category of professors including only eleven observations.
Table 5.4 displays the distribution of professor roles within each school and kind

of degree. It is clear that the roles are not equally divided according to these variables. In particular, seven schools have less than 15% of courses taught by external professors (i.e. individuals who are external to university, but recruited temporary as professors). Only the fifth school, the most numerous one, has a significant proportion of external professors (32.26% of the total school courses). Almost half of the courses in school 3 are taught by full professors. Nearly the same proportion represents the courses conducted by associate professors in school 2. These percentages are quite different from the other ones for the same categories of professors. However, we should remind that schools 2 and 3 are smaller than the others. Thus, the bias might be due to the difference of schools dimension in the sample rather than to real differences among schools.

Table 5.4: Role of professor per school and kind of degree. Percentage is referred to the total of the school/degree.

	Full Prof.		**Assoc. Prof.**		**Assis. Prof.**		**Ext. Prof.**	
	Freq.	%	**Freq.**	%	**Freq.**	%	**Freq.**	%
School								
1	64	28.70	81	36.32	68	30.50	10	4.48
2	9	12.68	33	46.48	22	30.99	7	9.85
3	31	49.20	16	25.40	12	19.05	4	6.35
4	111	29.68	122	32.62	117	31.28	24	6.42
5	82	17.63	112	24.09	121	26.02	150	32.26
6	16	28.07	20	35.09	15	26.31	6	10.53
7	61	22.76	104	38.80	79	29.48	24	8.96
8	63	19.57	97	30.12	121	37.58	41	12.73
Degree								
Bachelor's	190	18.01	336	31.85	319	30.24	210	19.90
Master's	138	29.37	139	29.57	148	31.49	45	9.57
Five-years	109	34.28	110	34.59	88	27.67	11	3.46

The noteworthy features concerning kinds of degree are mainly referable to bachelor's. This type of degree has more courses taught by external professors and less by full professors than master's and five-years-long degrees. The other professor categories are evenly divided in our sample among the degrees.

In the conditional models we are going to show in the following section, we chose full professor as the reference category.

5.3 Comparison between one- and two-class latent growth models

The unconditional model shown in Section 5.1 did not fit the data well. Furthermore it was not realistic, since it did not consider any covariate affecting growth factors or the possibility to have more than one cluster, with different growth trajectories. Two extensions of this model will be presented in the present section, in order to study more accurately the change in satisfaction. The first extension is a conditional second-order latent growth model, which only introduces some covariates in the previous unconditional model. Hence, it also requires all the observations to come from the same population, with only one latent trajectory explaining the change process (homogeneity). The second extension is a conditional second-order latent growth mixture model. According to this model, observations are divided into more groups, each with its specific latent trajectory (heterogeneity). In our case, the optimal solution provides for two latent groups. The first one is smaller, involving only 116 courses out of the total of 1,843 observations ($6,29\%$).

Showing simultaneously the estimates of these two models, we want to demonstrate which one of them is preferable. Thus, the comparison might make it clearer whether a single path is sufficient to describe adequately the evolution process of satisfaction in the whole sample. Table 5.5 includes the estimates of the measurement part of the models, i.e. the one referring to the items and the first latent level. In line with the previous unconditional model, factor loadings were considered time- and class-invariant. Loading estimates of both models confirm the results obtained through the unconditional model. The residual variances of first-level latent constructs are still high and significantly different from zero. However, they show a little improvement with respect to the previous model. The only available time-variant variable is the number of questionnaires filled in by students every year, a proxy of the real class size. If significant, it may influence differently each latent construct η_t. Nevertheless, in our analysis we decided to constrain this variable to be equal in the three years. This choice was made noting that the number-of-questionnaire estimates were all slightly negative and not statistically significant in the three time periods analysed. Moreover, in the previous Subsection 5.2.1 we showed a very similar distribution of the questionnaires filled in over time. After having set the constraint, both models confirmed the non-significance of this covariate. Thus, in our data set evidence shows that student's satisfaction with the course is not influenced by (the proxy of) class size.

Table 5.5: Estimation of the measurement part of conditional one- and two-class models. Levels of significance: p-value<0.001 $***$; 0.001<p-value<0.01 $**$; 0.01<p-value<0.05 $*$

	One-latent-class model			**Two-latent-class model**		
	Estimate		**S. E.**	**Estimate**		**S. E.**
Loadings						
OA	1.000		$---$	1.000		$---$
ED	1.188	$***$	0.008	1.188	$***$	0.012
I04	0.896	$***$	0.010	0.897	$***$	0.014
I10	1.075	$***$	0.010	1.075	$***$	0.015
I11	0.989		0.011	0.989		0.015
Residual Variances						
θ_1	0.469	$***$	0.044	0.428	$***$	0.131
θ_2	0.599	$***$	0.028	0.697	$***$	0.052
θ_3	0.575	$***$	0.050	0.475	$***$	0.087
Covariate on η_t						
# of questionnaires	-0.002		0.001	-0.003		0.002

The structural part of the two models is given in Table 5.6. It is assumed that time-invariant covariates, expressing individual differences among the courses, affect both latent growth factors α and β. Let us deal with one-latent-class conditional model at first. We run initially this model including school dummies (fifth school was the reference one), but we decided to remove all of them due to their low significances. In fact, the only significant (and negative) coefficient was the one referring to school 4 and associated to the latent intercept α. This would lead to conclude that this school has a lower starting point than the others (and therefore lower satisfaction at the beginning), but a growth rate similar to theirs. However, we noticed that the significant negative coefficient for school 4 was very close to the one of borrowed courses. We must remind that school 4 was the one with the higher proportion of borrowed courses. Therefore, it is reasonable to assume significance of school 4 coefficient to be due to differences between borrowed and not-borrowed courses, rather than to effective differences among schools. For this reason, we can affirm that school membership does not affect the change in satisfaction. All the other covariates have direct effects on the parameters determining the latent growth trajectory. As it was for the unconditional model, the variance of the latent intercept α is the only one significantly different from zero. Hence it confirms the hypothesis of different initial levels of satisfaction, but the same growth rate among the courses. This is in apparent conflict with the significance of covariates on the slope factor. Actually, it means that the constant parameter in the trajectory slope is not statistically significant, but the same slope factor does vary as a function of the significant covariates. Looking at the estimates, it is clear the effect that the kind of degree has on the latent intercept. Considering bachelor's degree as the baseline, the initial level of satisfaction of both the other degrees is significantly higher. However, while the degrees have different starting points as the growth curve begins, they all share the same linear growth rate (i.e., degree covariate has not influence on random slope). The highly negative estimate of borrowed parameter on the intercept confirms, as we said previously, that borrowed courses are less appreciated initially than normal courses. Nevertheless, their growth rate in time is equal to the one of standard courses. Being the hours range quite large, the estimates (and standard errors) associated to this variable are small as a consequence. Obviously, when we report a value of 0.000 we imply that the same value is smaller than 0.001. The hours of didactic activity have not a direct impact on the latent intercept, but they affect the latent slope. Thus, a course with many hours per week or lasting one academic year presents a negative growth rate, while it has the same latent intercept of short courses. For what concerns professors role, it is noteworthy that only associate professors have a slightly negative intercept with respect to the baseline (full professors). The other two roles have not different starting levels than the reference category. The growth rates of the three roles differ substantially from the one of full professors and appear to be quite similar to each other. With a positive value close to 0.10 on the slope, associate, assistant and external professors have a latent trajectory growing faster than the one for the baseline. This means that in the long term courses taught by full professors are considered by students less satisfactory than the others.

As we previously stated, the mixture model divided our population into two distinct groups. The first group is the smaller one, containing only the 6% of the total of the courses. From now on, we will refer to it as "Class 1". The second one will subsequently be called "Class 2". A higher number of latent classes would lead to even smaller clusters and thus we preferred to keep this two-class solution. The variance of the latent intercept is significantly different from zero, as usual, even if its significance has slightly reduced. The variance of latent slope and its covariance with the intercept are again statistically not significant. The estimates of the covariates in this model are rather curious. While in the conditional one-class model we had different significances, in this case they almost disappeared. For example, in Class 1

Table 5.6: Estimation of the structural part of conditional one- and two-class models. Levels of significance: p-value<0.001 ***; 0.001<p-value<0.01 **; 0.01<p-value<0.05 *

	1-latent-class model			**2-latent-class model**					
				Class 1			**Class 2**		
	Estimate		**S. E.**	**Estimate**		**S. E.**	**Estimate**		**S. E.**
Class %	1		–	0.06		–	0.94		–
Intercept α									
Master	0.205	***	0.051	0.151		1.281	0.267	**	0.088
5-year	0.149	*	0.062	−0.055		0.877	0.119		0.122
Borrowed	−0.367	***	0.068	−0.800		1.651	−0.196		0.107
Hours	0.000		0.001	0.014		0.009	0.000		0.001
Assoc. Prof.	−0.135	*	0.054	−0.814		1.003	−0.035		0.139
Assis. Prof.	0.031		0.055	0.036		0.636	0.069		0.074
Ext. Prof.	−0.021		0.070	0.057		0.801	−0.031		0.084
Slope β									
Master	−0.049		0.029	−0.220		1.633	−0.054		0.047
5-year	−0.075		0.039	0.991		0.912	−0.099		0.062
Borrowed	0.036		0.041	0.118		1.883	−0.009		0.059
Hours	−0.001	***	0.000	−0.018		0.011	0.000		0.000
Assoc. Prof.	0.100	**	0.030	0.516		1.055	0.019		0.139
Assis. Prof.	0.103	**	0.030	0.005		0.329	0.050		0.049
Ext. Prof.	0.096	*	0.040	−0.147		0.680	0.092	*	0.042
Covariance									
ψ_α	0.375	***	0.042	0.275	**	0.099	0.275	**	0.099
ψ_β	0.029		0.022	0.020		0.061	0.020		0.061
$\psi_{\alpha\beta}$	0.006		0.025	−0.022		0.121	−0.034		0.033

the covariates do not have a significant effect on any of the growth factors. This is due to very high standard errors, which compromise the significance. The cause of such high errors might be attributable to the small class size. Class 2 presents few significant effects, since the estimates have decreased while the standard errors have increased in comparison with one-class model. Master's degree continues to have a strong positive impact on latent intercept. The same can be said for the effect of external professors' teaching on the latent slope, similar to the one in the previous model. Even in this more numerous group, any other estimate is approximately equal to zero.

In our analysis, we also tried to see whether the covariates could be directly related to the categorical latent variable indicating class membership. Thus, we assumed the covariates could affect both class belonging and class growth factors. The results we obtained (not reported for sake of brevity) did not lead to such a different clustering from the other model. We thought it was not worth to use a more complex model (the one with covariates affecting class membership) to have only little difference in the results.

The analysis of the structural parts of the two models presented in Table5.6 would lead to think that the simpler model, i.e. with only one latent class, is adequate to explain the effective trajectory of students' satisfaction over years. However, both AIC and BIC indicate the more complex solution as the one to be preferred. This fact requires a further exploratory analysis of the clusters obtained, in order to assess if the model discriminates two different latent trends for satisfaction. We will focus on such analysis in the following and final section of this work.

5.4 Analysis of the two latent clusters

The present section aims to study the two groups obtained through the latent growth mixture model. Thus, we will verify whether this solution highlights any difference in the evolution process of satisfaction with the course. We remind that Class 1 is composed of 116 courses, while the remaining 1,727 observations belong to Class 2. Despite its low significances, the latent growth mixture model was able to catch clearly two contrasting change patterns. All the five items of the reduced scale used in our models proved to have similar quartiles in the groups and over time. This is confirmatory of the consistency of the clustering. For the sake of brevity, we chose to show only the box plots regarding the mean of those five items. The differences between the two clusters are pretty evident in Figure 5.5. Class 2 contains courses with a high satisfaction level, which is quite constant in the three academic years considered. Class 1 is instead composed of the most "problematic" courses. Actually, this kind of courses show a lower items mean than the others already in the first year analysed. In the second year, the level of the items score is approximately the same of the first time period, in line with "good" courses in Class 2. It is in the last academic year that Class 1 courses experienced a great change, with an incredible fall. In fact, if the mean of the item scores was around 6.5 in the previous years, it collapsed far below 6 in the last period. Analysing the differences in the covariates between the groups, we came to the following considerations. Class 2 has averagely a superior number of questionnaires filled in compared to Class 1. This

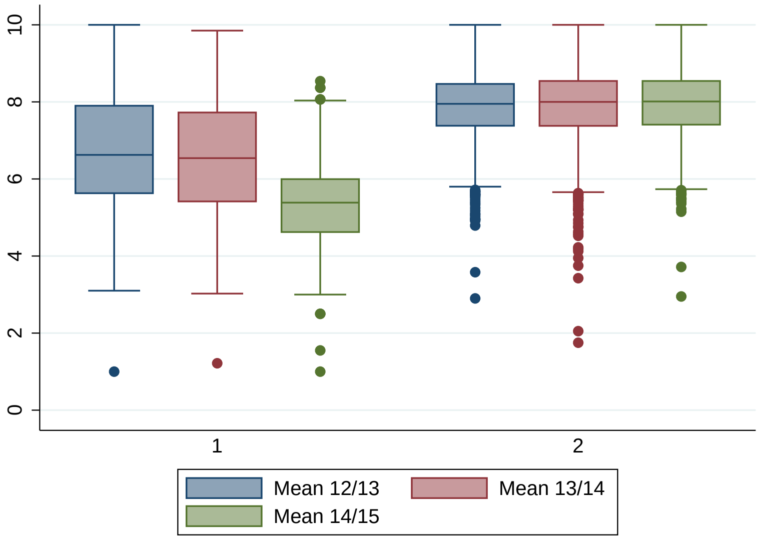

Figure 5.5: Box plots of five-item mean over time, comparison between Classes 1 and 2.

could be due to the word-of-mouth across students. Having heard about good experiences with a course from others, students may be more interested in attending the same course. For the opposite reason, “bad” courses have usually lower students’ attendance. The problematic group has longer courses on average and thus with more ECTS. This fact might be explained by students getting bored and giving lower scores as a consequence. As we explained before, the school covariate was not included in our latent growth models because of low significance of parameters. Studying this variable over clusters, we noticed that 79 courses out of the total of 116 (68.10%) come from schools 4 (39 courses) and 5 (40), which are the largest ones. In order to have more robust results, we excluded from the data set all the observations presenting less than three questionnaires filled in at least in one of the academic years. The reduced data set involved only 1,282 courses, with a reduction of 30.44% in relation to the original number of observations. The models estimated were very close to the ones obtained using the total data. For this reason we are not going to report such results. However, it could be interesting to know how clusters composition changes removing the courses with a low response rate, i.e. with only one or two questionnaires filled in. After applying the robust solution, Class 1 lost almost half of its courses (48.28%). It indicates that many courses with low ratings were in fact badly evaluated by only few students attending and not appreciating them. Being only one or two respondents, it is clear that those ratings cannot be considered completely reliable. However, Figure 5.6 shows a situation that is similar to the one with all observations considered (not-robust solution). It confirms again the consistency of our results, independently of the number of questionnaires filled in for each time period. The clusters maintain similar characteristics in both solutions. In the robust case, the "extreme" scores, i.e. really low or high ones, have reduced their number compared to the not-robust one. The average of items mean has slightly decreased in both classes, but for Class 2 it is again constant in time and still high. For what concerns Class 1, it is more visible the items mean falling year after year, even though the means in the first two years are still near.

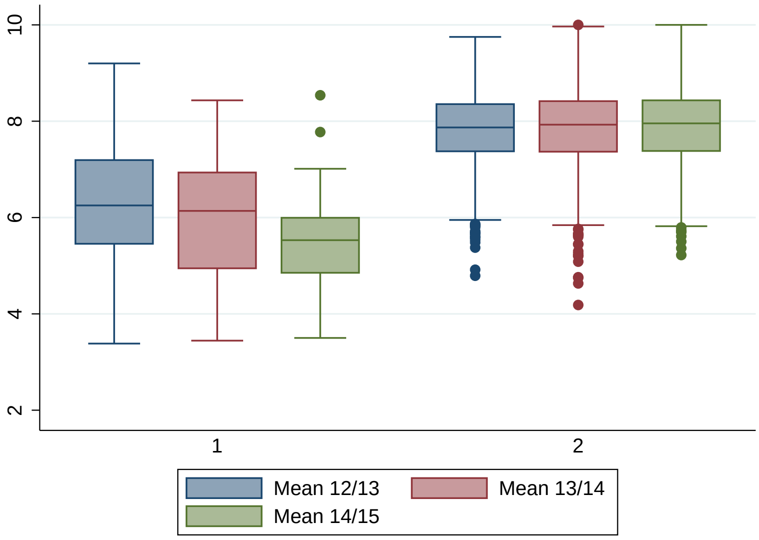

Figure 5.6: Box plots of five-item mean over time, comparison between Classes 1 and 2 (robust solution).

We already showed that covariates were not statistically significant in the two-class model. Just to confirm such non-significance, two logistic regressions were conducted. The first one was run on the whole data set, while the second one only on the reduced set of 1,282 courses. The dependent variable expresses the membership to the first or the second cluster. This dichotomous variable is equal to 0 if the course belongs to Class 1 and equal to 1 otherwise. The results are listed in Table 5.7. The only variable to be significant in both solutions is the one referring to hours, with a negative coefficient. Borrowed courses and the number of questionnaires in the second year are both significant at the 5% level. However, this significance disappears as we consider the robust solution. All the other estimates cannot be considered different from zero in both cases. These logistic regressions were meant to be only a confirmation of what we already expected. There are still many unknown aspects concerning satisfaction, which would allow a deeper comprehension of such an intangible and latent construct.

Table 5.7: Logistic regression on the class membership (original and robust solutions). Levels of significance: p-value<0.001 ∗∗∗; 0.001<p-value<0.01 ∗∗; 0.01<p-value<0.05 ∗

	Original solution			**Robust solution**		
	Estimate		**S. E.**	**Estimate**		**S. E.**
Master	−0.289		0.251	−0.404		0.354
5-year	−0.251		0.288	0.138		0.414
Borrowed	−0.665	∗	0.310	−0.629		0.381
Hours	−0.007	∗∗	0.002	−0.007	∗	0.003
Assoc. Prof.	0.125		0.259	−0.157		0.334
Assis. Prof.	0.377		0.276	0.518		0.394
Ext. Prof.	−0.089		0.315	0.226		0.482
# quest. 12/13	−0.002		0.018	0.001		0.196
# quest. 13/14	0.050	∗	0.023	0.043		0.026
# quest. 14/15	0.017		0.020	−0.006		0.210

A longitudinal analysis of the data was the final step of the present work and the aim of this chapter. The analysis was conducted by means of latent growth models. Initially a simple unconditional model was estimated. From that point, two further models were considered. Both included more covariates able to affect the latent growth factors. The first model considered only one latent trajectory for all individuals, thus homogeneity. The second one provided for two different latent classes (and trajectories), thus heterogeneity in the satisfaction growth. We preferred the second one, namely the latent growth mixture model. Even though the significances of covariates was absolutely not satisfactory, what our latent growth mixture model did is rather important. It managed to separate good courses, composing the largest group, from the bad ones, the small residual group. Our model is able to catch to some extent the differences in the evolution of satisfaction among the courses. Thus, most of the courses are going well, with students averagely and continuously satisfied. University of Padua should pay attention just to those courses whose satisfaction level has decreased in the last three years, as our analysis proved. Knowing which ones are the not-satisfactory courses, it will be easier to detect the causes of their low level of satisfaction. Consequently, it will be easier to have an improvement in the quality of didactics.

Conclusion

The main objective of the present thesis is to analyse the evolution of students' satisfaction with courses in the specific context of the University of Padua. We disposed of data involving students' responses to the questionnaire of didactics in the last three academic years. Only university courses present in all years were considered in the analysis as our observations. To ensure a proper comparison of students' ratings through time, only part of the original questionnaire was actually considered. This part is composed of twelve items (questions), the only ones which have been consistent in the three years. To understand the latent structure underneath the measurement scale, we ran a Factor Analysis on the eleven items dealing with specific aspects of the course. The twelfth item is the one concerning overall satisfaction and was thus excluded from the analysis. At first, we focused our attention only on the last year available, the one of major interest. The results achieved were rather different from the ones in previous researches. We could have expected such a discrepancy, since the scales adopted are not identical. Factorial model confirmed that there is only one latent dimension behind the eleven items and we can simply name it "satisfaction". However, eleven measures for a single factor were far too many and were compromising the fit of our model. Two reductions of the items set have been proposed, each of them consisting of five measures. The same factorial model on the last year was then estimated, according to those reduced scales. We established which one of the reduced scales was fitting the data best with the help of many different goodness-of-fit indices. After that, we used the preferred solution to explore cross-sectionally the level of satisfaction in each year considered. Once the latent component had been adequately verified, we proceeded with a longitudinal analysis of the data. This was possible by means of latent growth modelling. The first model we estimated had no covariates in it and assumed all the courses to have the same latent trajectory defining the evolution process of satisfaction. Considering this model too simple and not very realistic, we used two extensions including some covariates (teacher- and course-related variables). These covariates were supposed to affect directly the latent growth factors, which characterise in fact the latent trajectories. As the first model, one of the extensions tried to explain the evolution of satisfaction considering homogeneity in the population, thus one single growth path common to all the courses. The second one divided our observations into two different classes, each one with its own growth trajectory. Even though the significances of covariates were not relevant in the last model, the clustering obtained through it led us to make few remarkable thoughts. In an ideal situation in which courses are very good, we would expect to have one between: a constant level of satisfaction, for courses with a high satisfaction since the very beginning of the period analysed (e.g., courses taught by professors with many years of experience); a growing level of satisfaction, for courses in which new professors gain some experience over time or change efficiently their teaching methods, which were previously not good. In our case, the modelling was able to detect two really different behaviours within the courses taught in Padua. We found that most of the courses, grouped in one cluster, have not experienced a relevant change in the satisfaction level. However, this level was already high in the first academic year of our analysis. Such courses are therefore good and there is no need to worry about them. The second cluster contains only few courses with a decreasing latent trajectory. These "troubled" courses deserve special attentions, as they show problems that have to be fixed. Despite the low significance of school covariate, we found that the major part of the problematic courses (almost 70% of the total) belong to schools 4 and 5. The limits of our model are mainly due to the non-significance of the covariates available. That does not allow to say immediately what is wrong with bad courses. However, it leads us to think that many other variables may affect students' satisfaction, which is such a hard construct to understand completely. From the literature and personal observations, we could suggest to whom it may concern to collect other information about the courses. For instance, the gender of both students and professors might influence directly satisfaction with the course (Spooren, 2010). In the same way, the age could play another important role in defining satisfaction, just as professors' experience in a particular course. An adequate proxy for the last variable could be the number of years spent by professors in teaching those subjects.

Concluding, through our model we revealed a meaningful instrument able to divide good courses from the bad ones. At this point the cause of such dissatisfaction is not clear. Knowing which courses present evident problems, it should be easy for university management to verify what is making those courses so bad and guarantee an improvement of both teaching and learning processes.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Athiyaman, A. (1997). Linking student satisfaction and service quality perceptions: the case of university education. *European Journal of Marketing*, 31(7):528–540.

Baker, J. N. (2008). *Confirmatory Factor Analysis on the Nonverbal Literacy Assessment (NVLA): A Measure of Early Literacy Skills for Students with Significant Cognitive Disabilities*. ProQuest.

Bassi, F., Clerici, R., and Aquario, D. (2016). Students' Evaluation of Teaching at a Large Italian University: Measurement Scale Validation.

Bassi, F. and Dias, J. G. (2013). Longitudinal patterns of financial product ownership: A latent growth mixture approach. In Grigoletto, M., Lisi, F., and Petrone, S., editors, *Complex Models and Computational Methods in Statistics*, pages 27–36. Springer.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, 107(2):238–246.

Beran, T. and Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education*, 30(6):593–601.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Publications.

Browne, M. W. and Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2):230–258.

Bryman, A. and Cramer, D. (2005). *Quantitative data analysis with SPSS 12 and 13: a guide for social scientists.* Psychology Press.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.

Chen, C.-Y., Chen, P.-C., and Chen, P.-Y. (2014). Teaching quality in higher education: An introductory review on a process-oriented teaching-quality model. *Total Quality Management & Business Excellence*, 25(1-2):36–56.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of marketing research*, 16:64–73.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Dalla Zuanna, G., Bassi, F., Clerici, R., Paccagnella, O., Paggiaro, A., Aquario, D., Mazzucco, C., Martinoia, S., Stocco, C., and Pierobon, S. (2015). Tools for teaching assessment at Padua University: role, development and validation. PRODID Project (Teacher professional development and academic educational innovation) - Report of the Research Unit n. 3.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

DeVellis, R. F. (1991). *Scale development: Theory and applications.* Sage Publications, Inc, London.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S. M., and Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing scholarship*, 39(2):155–164.

Dziuban, C. D. and Shirkey, E. C. (1980). Sampling adequacy and the semantic differential. *Psychological Reports*, 47(2):351–357.

Elliott, K. M. and Shin, D. (2002). Student satisfaction: An alternative approach to assessing this important concept. *Journal of Higher Education Policy and Management*, 24(2):197–209.

Fabbris, L. (2002). La misura della student satisfaction per la valutazione della qualità della didattica. *F. Delvecchio e L. Carli Sardi — Indicatori e metodi per l'analisi dei percorsi universitari e post-universitari, Cleup, Padova*, pages 1–20.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272–299.

Fornell, C. and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, 18(1):39–50.

Gerbing, D. W. and Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, 25:186–192.

Guido, G., Bassi, F., and Peluso, A. M. (2010). *La soddisfazione del consumatore: la misura della customer satisfaction nelle esperienze di consumo*. F. Angeli, Milano.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2010). *Multivariate data analysis: A global perspective (7th Ed.)*. Pearson Prentice Hall, New Jersey.

Hancock, G. R. and Buehl, M. M. (2008). Second-order latent growth models with shifting indicators. *Journal of Modern Applied Statistical Methods*, 7(1):39–55.

Harvey, L. (2006). Understanding Quality. In *Bologna Handbook: Making Bologna work*. Brussels, European University Association and Berlin, Raabe.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2):139–164.

Hooper, D., Coughlan, J., and Mullen, M. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *Electronic Journal of Business Research Method*, 6(1):53–60.

Hu, L.-T. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55.

Iasevoli, G. (2007). La definizione e la misurazione della soddisfazione del consumatore. In Hoffman, K. D., Bateson, J. E. G., and Iasevoli, G., editors, *Marketing dei servizi*, pages 274–276. Apogeo, Milano.

Jolliffe, I. (2002). *Principal component analysis.* Wiley Online Library.

Jung, T. and Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, 2(1):302–317.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20:141–151.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1):31–36.

Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3):767–778.

Massingham, P. and Herrington, T. (2006). Does Attendance Matter? An Examination of Student Attitudes, Participation, Performance and Attendance. *Journal of University Teaching & Learning Practice*, 3(2):82–103.

Muthén (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In Kaplan, D., editor, *Handbook of quantitative methodology for the social sciences*, pages 345–368.

Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of marketing research*, pages 6–17.

Pirsig, R. M. (1991). *Lila. An Inquiry into Morals.* Bantam Books, New York.

Salgueiro, M. F., Smith, P. W. F., and Vieira, M. D. T. (2013). A multi-process second-order latent growth curve model for subjective well-being. *Quality & Quantity*, 47(2):735–752.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Segars, A. H. (1997). Assessing the unidimensionality of measurement: A paradigm and illustration within the context of information systems research. *Omega*, 25(1):107–121.

Selltiz, C., Wrightsman, L. S., and Cook, S. W. (1976). *Research Methods in Social Relations.* Holt, Rinehart and Winston, New York.

Spooren, P. (2010). On the credibility of the judge: A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in Educational Evaluation*, 36(4):121–131.

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching the State of the Art. *Review of Educational Research*, 83(4):598–642.

Srikanthan, G. and Dalrymple, J. (2003). Developing alternative perspectives for quality in higher education. *International Journal of Educational Management*, 17(3):126–136.

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, 12(1):55–76.