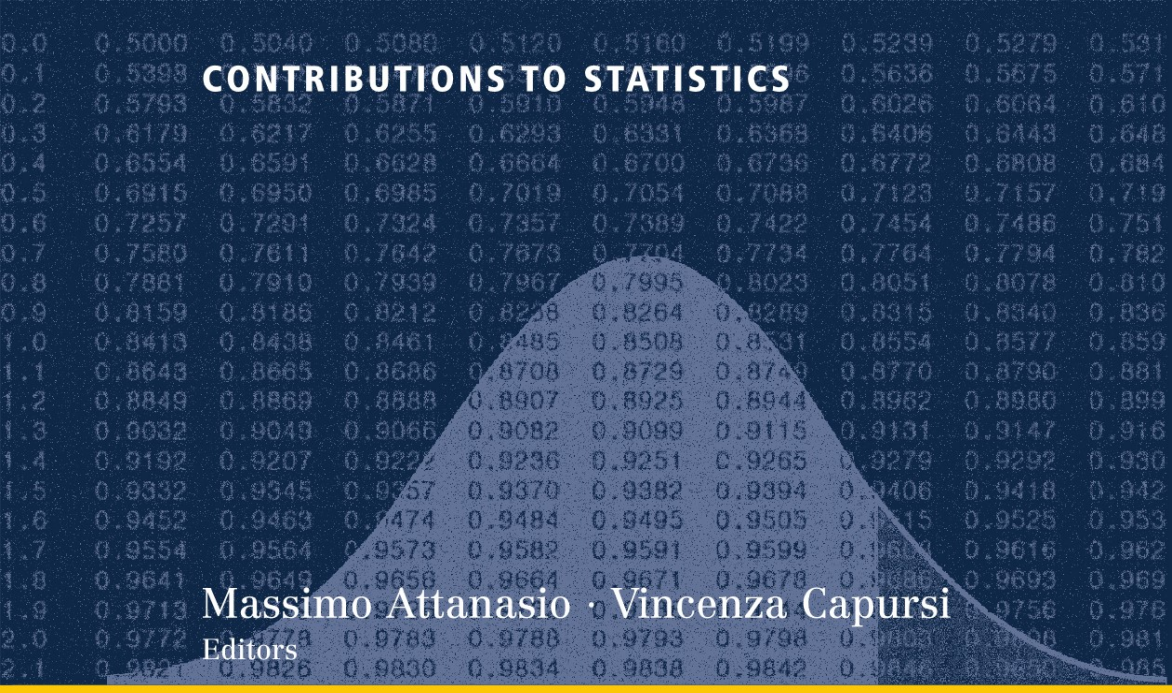


CONTRIBUTIONS TO STATISTICS

Massimo Attanasio · Vincenza Capursi
Editors

Statistical Methods for the Evaluation of University Systems

Physica-Verlag
A Springer Company

Contributions to Statistics

For further volumes:
http://www.springer.com/series/2912

Massimo Attanasio · Vincenza Capursi
Editors

Statistical Methods for the Evaluation of University Systems

Physica-Verlag

Editors
Massimo Attanasio
Università di Palermo
Dip. to Scienze Statistiche e Matematiche
Viale delle Scienze
90128 Palermo
Italy
attana@unipa.it

Vincenza Capursi
Università di Palermo
Dip. to Scienze Statistiche e Matematiche
Viale delle Scienze
90128 Palermo
Italy
capursi@unipa.it

ISSN 1431-1968
ISBN 978-3-7908-2374-5 e-ISBN 978-3-7908-2375-2
DOI 10.1007/978-3-7908-2375-2
Springer Heidelberg Dordrecht London New York

Cover design: Integra Software Service Pvt. Ltd., Pondicherry

Printed on acid-free paper

Physica-Verlag is a brand of Springer-Verlag Berlin Heidelberg
Springer-Verlag is part of Springer Science+Business Media (www.springer.com)

Preface

During the last years social-economic systems have been involved in a rapid change not only at the European, but also at a wider international level. This process has triggered a renewed interest for topics related to Education and Knowledge development. Particularly "higher education has been affected by a number of changes, including higher rates of participation, internationalisation, the growing importance of knowledge-led economies and increased global competition". Thus, the Bologna Process [1999] and the Lisbon Strategy [2000] aim at making the European Union "the most dynamic and competitive knowledge-based economy in the world capable of sustainable economic growth with more and better jobs and greater social cohesion". In the last decade, the Italian University System (IUS) has changed in several directions as a result of those two European policy processes. The principles that inspired such changes lay in the improvement of the whole University service, of the competitiveness and of accountability, and in the aim of getting closer to the other European University systems. The reform – introduced by several bills, first D.L. 509/1998 and 270/2004 and after DD.MM. 544/2007 e 362/2007 – has determined that the propensity to accountability of all the IUS's processes, the accuracy in the management of financial and human resources, the monitoring and the evaluation of the main processes of the US have become the principles of the governance of the IUS. After 7 years, the reform is not still fully "implemented", and its approaching effectiveness has already provoked questions on its effects and results. For certain, two big issues have been claimed in the last years: the "shortage" of monetary funds to cover the new "tasks" introduced by the reform and the absence of consistent and clear evaluation policies. In fact, a structured framework for the evaluation of the quality in the IUS is still in progress. The main obstacle is given by the ambiguities of the IUS, because the processes and the functions of the didactics and the research activities have not been defined clearly. In the last years the University Ministry bills and indications have been very fragmentary and without a long-term perspective. These indecisions have not helped the universities in the process of defining and establishing common evaluation methods and models in order to improve the research and didactics activities. For instance, a nationwide research evaluation program was conducted just once in 2001–2003 and another one will be likely carried out in the next years. For the didactics activities, Students Evaluation of Teaching (SET) results are utilized just to organize teaching activities in terms of quality in

a few universities, even if this survey is mandatory since 1999. In this direction, the first input is given by the Ministerial Decree 544/2007, in which there is an indicator to measure the quality of teaching. This indicator is used to give funds to the universities and it does not take into account the students assessments. In fact, it is merely quantitative because it is given by the number of the courses, in which SET is conducted, over the total courses offered, so the university takes "money" if that indicator is greater than the median (which is computed nationwide), otherwise it takes less funds. On the other hand, on behalf of the Ministry of Education and Scientific Research, it was instituted a National Committee for the Evaluation of the University System in 1999. It has produced either several reports on descriptive analysis of the IUS or some research topics on evaluation of the IUS. At the same time, several Italian statisticians have published scientific papers and reports either about the criteria, the modalities, the contents of the IUS activities, or the construction of measures (and indicators) devoted just to the IUS. In general, the aspects covered by the Italian literature concern the IUS organization, its functionality and, in particular, the political and cultural meaning involved in the whole evaluation process. Both the institutional and the scientific works are the result of the Bologna Process, whose "core" is the quality, as demonstrated by the European quality assurance developments. The Standards and Guidelines for Quality Assurance in the European Higher Education Area and the annual European Quality Assurance Forum are a concrete proof of the crucial role of "Quality Shared" as a success factor to engage a real change in the policies. The difficulties encountered in classifying the most appropriate evaluative practices for the IUS can be found in the statistical literature evaluation processes. Indeed, the statistical literature, that analyzes the main processes of the IUS, is very vast and therefore it is hard to classify it. But some trends in methods and statistical models to analyze the results of the evaluation of the IUS can be pointed out. At first the scientific developments concentrated on the construction of simple and/or composite indicators, with the aim of providing statistical tools easily interpretable to policy makers. However, the inadequacy of indicators to capture the complexity of the assessment processes led to the need for the development of a modelling approach for the analysis of such processes. The modelling approach is vast too, because the IUS covers such different issues that we are able to mention (and partially cover in this book) just some of them. For instance, the most used models to measure individual opinions about the quality of teaching activities allow us to take into account factors that can lead to a misleading interpretation of the latent structure of the examined phenomenon. So multivariate analysis techniques or covariance models are extensively used for the individualization of latent dimensions. A further approach is based on Rasch Model, able to measure individual opinions and to test a questionnaire. Moreover, fixed or random effects models are used to analyze longitudinal individual data because they allow to test the presence of heterogeneity of behaviour among sub-populations and they are appropriate to study the student's career performance and occupational outcomes. Finally, in a stochastic process view, the students' careers and occupational outcomes are analyzed by "Markov chains", that allow to model durations and transitions from one state to another of the IUS in an appropriate way. This book, which

collects a selected range of refereed chapters, is the results of the projects funded by the Italian Ministry of Education (Prin, 2005) entitles "Construction Indicators for Public Decision-making Processes Between Measurement Issues and Opportunities Knowledge" and (Prin, 2008) "Measures, statistical models and indicators for the assessment of the University System". It contains four parts, each of them devoted to the following topics:

- Introduction: Different Perspectives of the Evaluation of the Italian University System
- The Evaluation in the Italian Universities. Student Teaching Evaluation
- The Evaluation in the Italian Universities. Statistical Methods for Careers and Services Evaluation
- Research Design and data for Evaluation: University between the High School and the Labour Market

The first part includes three chapters: two of them devoted to the assessment of university teaching by students from two different points of view. One makes a critical analysis of the practice teaching evaluation in the IUS and the other one investigates the assessment process in organizations. The third chapter examines a method, currently used by the Italian research institute Censis, for ranking Italian universities on indicators related to educational quality and proposes alternative ways to get the rankings. The second part is dedicated to chapters developing different statistical models, almost all attributable to GLMM, aimed to investigate the determinants of the evaluation teaching process by students. In this part relevant issues, such as the quantification of the impact of individual covariates on the perception of the main aspects of University teaching and the study of how perception changes with the profiles of students, are analyzed. The third part is devoted to models to analyze the quality of services and careers of students. This part contains some methodological and substantive results of interest: the estimates of the satisfaction levels for specific feature of students and/or service by the extension of the logistic binomial GLMM, the estimates of the systematic changes over time of this student career delay indicator by the Latent Curve Model and the estimates of the student's perception of the quality of the management. Finally, the last part includes different interesting applications. These chapters may appear fragmentary in the content. However, the results span from the assessment of primary and secondary effects in secondary school choices in Italy to the labour market outcomes for PhDs. So it is possible to give a track, through the use of different statistical models, of the difficult transition from the Italian higher education till the entry into the labour market. We would like to thank all the referees for their invaluable effort in reviewing more than 25 papers. They are:

Silvia BACCI – Università di Firenze, Italy
Erich BATTISTIN – Università di Padova, Italy
Gianni BETTI – Università di Siena, Italy
Sergio BOLASCO – Università "La Sapienza" Roma, Italy

A special thank goes to Rosalinda Allegro e Giovanni Boscaino who helped us for the preparation of this volume. A final acknowledgment goes to the Italian Ministero dell'Università e della Ricerca Scientifica, which funded the projects "Construction Indicators for Public Decision-Making Processes Between Measurement issues and Opportunities Knowledge", 2005 and "Measures, statistical models and indicators for the assessment of the University System", 2008, and to the Rectorat of the University of Palermo, which hosted the final workshop of the project and contributed to publishing expenses.

Palermo
April 2010

Massimo Attanasio
Vincenza Capursi

Contents

Contributors

Fabio Aiello Facoltà di Scienze Economiche e Sociali, Università di Enna "Kore", Enna, Italy, fabio.aiello@unikore.it

Rosalinda Allegro Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università degli studi di Palermo, Palermo, Italy, rl.allegro@libero.it

Simona Balbi Dipartimento di Matematica e Statistica, Università di Napoli "Federico II", Napoli, Italy, simona.balbi@unina.it

Simona Balzano Dipartimento di Scienze Economiche, Università degli Studi di Cassino, 03043 Cassino, Italy, s.balzano@unicas.it

Lorenzo Bernardi Dipartimento di Scienze Statistiche, Università di Padova, Padova, Italy, lorenzo.bernardi@unipd.it; bernardi@stat.unipd.it

Paola Bolzonello Dipartimento di Scienze Statistiche, Dipartimento di Sociologia, Università di Padova, Padova, Italy, paola_bolzonello@libero.it

Stefano Campostrini Dipartimento di Statistica, Università Ca' Foscari Venezia, Venezia, Italy, s.campostrini@unive.it

Dalit Contini Dipartimento di Statistica e Matematica Applicata "Diego De Castro", Università di Torino, Torino, Italy, dalit.contini@unito.it

Marcella Corduas Dipartimento di Scienze Statistiche, Università di Napoli Federico II, 80138 Napoli, Italy, corduas@unina.it

Paola Costantini Dipartimento di Economia, Università di Cassino, Cassino, Italy, p.costantini@unicas.it

Corrado Crocetta Dipartimento di Scienze Economiche, Matematiche e Statistiche (DSEMS), Università di Foggia, Foggia, Italy, c.crocetta@unifg.it

Antonella D'Agostino Dipartimento di Statistica e Matematica per la Ricerca Economica, Università di Napoli Parthenope, 80133 Napoli, Italy, antonella.dagostino@uniparthenope.it

Béatrice d'Hombres Econometrics and Applied Statistics Unit, CRELL, Institute for the Protection and Security of the Citizen European Commission, Joint Research Centre, Ispra, Italy, beatrice.d'hombres@jrc.it

Davide Ferrari Dipartimento di Economia Politica, Università di Modena e Reggio Emilia, Modena, Italy, davide.ferrari@unimore.it

Patrizio Frederic Dipartimento di Economia Politica, Università di Modena e Reggio Emilia, Modena, Italy, patrizio.frederic@unimore.it

Simone Gerzeli Dipartimento di Statistica Applicata ed Economia "Libero Lenti", Università di Pavia, Pavia, Italy, simone.gerzeli@unipv.it

Giulio Ghellini Dipartimento di Metodi Quantitativi, Università di Siena, Siena, Italy, ghellini@unisi.it

Ornella Giambalvo Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università di Palermo, Palermo, Italy, giambi@unipa.it

Luigi Enrico Golzio Dipartimento di Economia Aziendale, Università di Modena e Reggio Emilia, Modena, Italy, luigienrico.golzio@unimore.it

Maria Iannario Dipartimento di Scienze Statistiche, Università di Napoli Federico II, Napoli, Italy, maria.iannario@unina.it

Michele Lalla Dipartimento di Economia Politica, Università di Modena e Reggio Emilia, Modena, Italy, michele.lalla@unimore.it

Vito M.R. Muggeo Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università di Palermo, Palermo, Italy, vito.muggeo@unipa.it

Domenico Piccolo Dipartimento di Scienze Statistiche, Università di Napoli "Federico II", Napoli, Italy, dopiccol@unina.it; domenico.piccolo@unina.it

Mariano Porcu Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Cagliari, Italy, mrporcu@unica.it

Maria Francesca Romano LEM, Management e Innovazione Scuola Superiore Sant'Anna di Pisa, Pisa, Italy, m.romano@sssup.it

Andrea Scagni Dipartimento di Statistica e Matematica Applicata "Diego De Castro", Università di Torino, Torino, Italy, andrea.scagni@unito.it

Isabella Sulis Dipartimento di Ricerche Economiche e Sociali, Università degli Studi di Cagliari, Cagliari, Italy, isulis@unica.it

Stefano Tarantola Econometrics and Applied Statistics Unit, CRELL, Ispra, Italy, stefano.tarantola@jrc.it

Laura Trinchera Dipartimento di Studi sullo Sviluppo Economico, Università di Macerata, Macerata, Italy, laura.trinchera@unimc.it

Arjuna Tuzzi Dipartimento di Sociologia, Università di Padova, Padova, Italy, arjuna.tuzzi@unipd.it

Daniël Van Nijlen Centre for Educational Effectiveness and Evaluation, Katholieke Universiteit, Leuven, Belgique, daniel.vannijlen@ped.kuleuven.be

Maria Prosperina Vitale Dipartimento di Economia, Università di Salerno, Fisciano (SA), Italy, mvitale@unisa.it

Susanna Zaccarin Dipartimento di Scienze Economiche e Statistiche, Università di Trieste, Trieste, Italy, susanna.zaccarin@econ.units.it

Emma Zavarrone Dipartimento di Economia Politica, Università di Milano-Bicocca, Milan, Italy, emma.zavarrone@unimib.it

Part I
Introduction: Different Perspectives of the Evaluation of the Italian University System

Chapter 1
TES – From Impressionism to Expressionism

Lorenzo Bernardi

1.1 Foreword: Excusatio Non Petita

A speaker addressing the final session of a conference risks being clumsily repetitive, for two reasons: on the one hand, because of a natural tendency to ride old hobby horses, but more especially, because much of the ground already covered by previous speakers is likely to be repeated. Whilst I make no apology for the first (I have noticed – and perhaps this is more evident in the political life of our country – that the simulacrum of credibility is definable by the regularity with which people uphold their positions), in the case of the second, I can only invite listeners to appreciate the convergence of opinions that has developed over time among many observers of the Italian experience in the area of student-focused Teaching Evaluation Surveys (TES). This convergence appears to be reached, even allowing for the diversity in stylistic expression and depth of feeling, in proportion to the intensity and directness of exposure to the subject matter, and the strength of the documentation and scientific arguments.

Before proceeding, I must make two more short points: firstly, the personal commitments of my recent professional life have been rather institutional, that influences my approach to the subject. My aim is to develop arguments on the cultural and political use of TES, largely ignoring the methodological and scientific methods by which it should be driven; on the other hand, I feel I should explain the title of this talk: those who know me will be familiar with my habit (sometimes paroxysmal or inapt) of exploiting the ample opportunities for expression afforded by the Italian language: likewise on this occasion, at the time of responding to the request of the organizers, I gave in to temptation; I must confess to have dared, although in overcoming the difficulties I had some help from a vocabulary of the Italian language (more exactly: A. Gabrielli, *Il grande italiano 2008*, Hoepli 2007[1]); this gives me

L. Bernardi (✉)
Dipartimento di Scienze Statistiche, Università di Padova, Padova, Italy
e-mail: bernardi@stat.unipd.it

[1] I must honestly admit, however, that I did not have the same luck with other dictionaries consulted.

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_1,

access to two definitions which, with a degree of licence, can be seen as consistent with the considerations I intend to present.

Impressionism: "representation of a reality through images that are *immediate, detached, selected according to the impression of the moment*". *Expressionism*: "tendency characterized by a predilection for the more dramatic and intense aspects of the human experience, represented through language that is *dramatic, full of conflict and unsettling*".

Now, in the first definition I want to focus attention on the words in italics, which would seem to sum up my critical appraisal of what TES has tended to be thus far: *immediate* – swiftly formulated and concluded, lacking solid theoretical constructs; *detached* – occasional, ephemeral, not homogeneous in space and in time; *according to the impression of the moment*, fragmented, without any process of accumulation and growth. The transition to an expressionist state, by contrast, must be *worrying*, not only for those who design and conduct the TES – an aspect with which everyone here will be fully and practically familiar – but also for those who may be affected according to their tasks by the results and invited, implicitly or explicitly, to adopt a different approach in their work. And with this end in view, it needs to be *dramatic* and to highlight the *conflicts* deriving from the great diversities of the domains in which it operates – thematic and territorial – also from the natural yet stimulating clashes of interpretation still possible (despite being no longer *in statu nascendi*).

After this lengthy introduction, we need something about the structure of my address, which is in four parts: the first is dedicated to a brief but necessary *destruens* consideration, keeping in mind the many invaluable contributions in which solid criticisms have been launched to the several factors of TES; secondly I shall then try to define principles and coordinates for a *construens* hypothesis outlining a more advanced and meaningful approach to the survey; thereafter, there will be an acrobatic attempt to construct a Utopian model intended purely as a goal (a far-off goal . . .); and finally, returning to the real world and more especially to the principle of progressive refinement, I will suggest a few feasible ideas for improving the current structure of the survey.

1.2 Pars Destruens: Accusatio Manifesta

As already anticipated, I will offer just a few words on this aspect of the question, obviously so as to minimize any repetition of what has already been said during the two days of the conference. Accordingly, my opinion can be summed up in four adjectives: notwithstanding its 10-year existence, the survey continues to be *impertinent*, *unsuitable*, *inconsequential*, and *inefficient*.

Whilst in the Italian language, *impertinent* can also mean "not pertinent", I use the word here in its everyday sense of impudence and disrespect: the survey has done no more than menace the conformism of tradition, making noise as a naughty

boy, a touch rebellious but with no real and firm intention of upsetting the tables and changing the rules; indeed there is even a quirky affection for the survey, though mixed with indifference proportional to the complexity (sometimes abstruseness) of the construction and analysis models that have been employed; there have been no actual refusals or calls for abolition – not institutionally at least, albeit beneath the surface many colleagues might wish it so. For threats to the institutions and the pax *academica*, and strategic plans of action, one must look elsewhere. The question to pose is, would it be possible, and under what conditions, to apply the other meaning of pertinent: *relevant and useful*[2].

On the accusation of *unsuitability* I must dwell slightly longer, as this is linked closely to the concept of usefulness; there are three levels of reasoning to express our opinion:

I. since the mandatory inception of TES, its purposes were not made clear; in effect, there are many objectives that could have been identified, and from these, the setup and content of the survey could have been delineated more clearly and explicitly: quite simply, does one measure satisfaction, or the sense of responsibility displayed by teachers, or the dependability of organizational processes, or the gap between student expectations and effective attainments, or does one assess the overall quality of courses and tuition, or gauge the relationship between value of teaching and other system evaluation components, or indeed all of these? In literature, conventionally, the technique and structure of evaluation are not armed and deployed without first establishing the battleground and the field of conquest; the impression is that the universities have armed themselves with a sabre and a blindfold: the swordsman can choose to strike low and blind, or simply wander around proudly armed but essentially harmless;
II. the *design of the survey* was determined *locally*: the organization, the responsibility, the timing, the nature of the indicators provided (apart from a small *set* proposed belatedly by CNVSU), and the method of dissemination are established by each local centre, effectively disallowing comparison of the results. Instead, we are convinced there can be no competition – teaching and scientific – when the documentation submitted for comparison and selection is put together on the basis of self-determined rules;
III. there were no "hot to use results" rules, either at national level or in general even at local level; neither was there any explicit mention of the areas of academic life targeted for change.

These first two considerations already provide grounds on which to justify the *inconsequence*, which can also be supported by previous research conducted as part

[2] The Gabrielli dictionary again.

of the Dottor Di.Va.Go project,[3] as well as by the single experiences of almost all universities; in particular, it will be remembered that the results of the survey:

I. are not widely disseminated, indeed officially (by resolutions of the Academic Bodies) their circulation is limited (normally to the Deans only, who are invited – though not bound – to use the findings as they wish);
II. are rarely the subject of comparison and public discussion, whether of the method or of the indications given by the findings;
III. are not translated into decisions for change in respect of the failings reported.

All these factors have their origin, firstly, in the attitude of most faculties, secondly in that of the universities themselves, and to a lesser extent in the conduct of the ministry, which nonetheless was in favour of the survey and stands to gain from it.

Finally, I believe it can be stated that in view of the various limits mentioned, the survey is also *inefficient*: in fact it is costly, whatever the basis for the design of the survey adopted in various situations; almost as a direct consequence, there are no cost-benefit analysis studies on its implementation, which could easily lead to abandon the whole survey; it is to some extent a source of distorted information, due principally to its methodological weaknesses (survey conducted only on occasional attendees, missing the opinions of greatest interest, that is to say those who have elected not to attend lessons or lectures; neglect of the "consistency of teaching/verification of learning" equation; etc.); the tired and mechanical repetition of the ritual over time has transformed the survey into a purely bureaucratic exercise, which among other things generates a tendency to become more and more careless or to fill in questionnaires unthinkingly or derisively. To conclude on this aspect, a simple aphorism: the social researcher knows that a survey perceived as statistically harassing (due to its intrusiveness, and especially to its uselessness) will doubtless give results that are statistically (and cognitively) embarrassing.

1.3 Pars Costruens: Non nova, Sed Nove

1.3.1 Guiding Principles

Before moving into this dimension of this chapter, which in many ways is based on subjective perceptions and Icarus-like attempts[4] – on wings of insubstantial feathers attached insecurely to a fragile frame – I ought to mention the principles by which it

[3] Related to this, see Gerzeli S, Parise N, Campostrini S, Magni C, Bernardi L in Capursi V, Ghellini G (2008) pp 70–89 and Bernardi L, Campostrini S, Parise N (2007) cicl.

[4] This image is preferred to the older and more traditional Pindaric digression ("Pindaric flight") because that the latter is by now associated with fatuous vacuity; in short, better to try and fail rather than be accused of emphatic and rhetorical but pointless lyricism.

is inspired. The first principle compasses two related affirmations: (a) whilst accepting the autonomy attributed to the system and to the single universities, as regards the logic and importance of evaluation, there needs to be a uniform approach to implementation of the TES; (b) this derives from the belief that autonomy in the area of evaluation (and perhaps not only in this area) should be the prerogative of the university system overall, rather than the individual university. These two opinions are comparatively strong, and while not original, may not easily be shared. What is more, they originate from the assumption (purely political, ideological, and cultural in nature) that it is the academic system that must share the duty of responding collegially to the expectations of the nation. Consequently it has to ensure it possesses the right tools and the most rigorous procedures, so that it can be judged on the basis of fully and correctly ordered arguments, shunning solipsistic whims, expedients, obfuscations. The reasons underlying this principle are quite trivial:

I. evaluation is worthwhile only if it produces comparative capacity adopting homogeneous plans of analysis;
II. homogeneous methods should be a guarantee to give to the survey a wider political significance not easily opposed by objectors, whether secretly or openly;
III. "true" rules and conditions of comparison generate a greater sense of responsibility both in the managers of the survey and in users of the findings that emerge from it;
IV. finally, this is the one condition capable of legitimizing national policies that are not necessarily or exclusively punitive. In effect, the teaching evaluation approach would be the one to favour, designed as it is to reward positive trends, changes of direction and effective emulations that could come about over time.

It should not be forgotten that from the outset of the initiative, there was real focus on the idea of uniformity at national level, as witnessed by three documents (the first two from the National Committee and the third from the Body that replaced it, namely CNVSU[5]) in which the authors seek to suggest ideas for a common approach, albeit they are concerned mainly with the content of the questionnaire, as the formulation of the questions and the number of multiple choice answers. Only in the first of these documents there are reflections on the design of the survey and on its functionality, also on the limits expected and, for various good reasons, accepted. To move in the direction indicated, however, common rules would need to be imposed in many aspects of the survey's design. This, however, is not the place to offer exact solutions, and moreover, it must be remembered that these aspects are

[5] See the following documents on the CNVSU site: RdR 1/98 (1998) Valutazione della didattica da parte degli studenti [Evaluation of teaching by students]; RDR 1/00 (2000) Questionario di base da utilizzare per l'attuazione di un programma per la valutazione della didattica da parte degli studenti [Basic questionnaire for use in the implementation of a programme for evaluation of teaching by students]; Doc 9/02 (2002) Proposta di un insieme minimo di domande per la valutazione della didattica da parte degli studenti frequentanti [Proposals for a minimum set of questions for the evaluation of teaching by attending students].

prejudicial to the end in view. In effect, univocal answers need to be given for each one of these operational choices:

I. to state responsibilities/tasks for the design;
II. to state responsibilities/tasks for organization and administration;
III. to state times, procedures and operators for the purposes of carrying out the work in the field and administering the questionnaires;
IV. to state procedures for diffusion of the findings, in order to present results consistently with their nature, responsibility, capacity for interpretation, and expectations of different recipients;
V. finally, agreeing on clear opportunities and forms for discussion, especially comparisons (local and national).

Just as important, and obvious in our view, is the second principle: TES is just one part – necessary but far from self-sufficient – of a complex evaluation mosaic that must appear more vivid, especially from the outside. In effect, the survey is effective only if integrated with other initiatives systematically and logically connected one with another. Looking only at the dimension of enhancing the teaching function of the universities – and therefore not, in this work, at that of evaluating their research activity and their management-administrative-accounting organization – an organic design should include coordinated surveys on:

I. *freshmen* – to assess their educational, cultural and social background, their reasons for going to university, and their expectations. Knowledge of these factors is almost indispensable when seeking subsequently to interpret characteristic differences in their career paths;
II. *careers within study courses* – a survey which is increasingly easier by the computer system now present in every higher education organization, capable of responding not only to the need for different analyses between faculties and universities[6] but also of indicating the nature and timing of circumstances and factors by which the educational process can be stalled;
III. *drop-outs* – including consideration of the conditions, timings and reasons, and the possible prospects for social and/or professional advancement that may have forced or prompted the decision to quit;
IV. *undergraduates* – to gather general assessments on the life and learning experience during the study period, considering the whole university experience, trying to keep apart single events, in order to indicate/highlight the value and limit of each career phase;
V. *the social and occupational destiny of graduates* – information needed, obviously, for a better insight into the effectiveness of the system[7];

[6] Interestingly the importance being assumed by indicators relating to this subject, as with regards the carrot-and-stick financing of single universities.
[7] On this aspect, for many years work was done first on an individual basis by many universities, then using general surveys proposed by associations of universities (the two associations including

VI. *the reactions from stakeholders* – with regard to the quality, pertinence and potential of the training received, with a view to its usefulness that should be more than contingent or quickly obsolescent;
VII. with a complementary function deriving from the principle that, for completeness, each evaluation must collect the views of all players in a process or a service, as regards the *attitudes, expectations and behaviors of the teaching staff*, and offer the facility of measuring the distance of opinions and goals assigned by each active component;
VIII. finally, and with the same objective as the previous point, *on the reactions* of all the *Governing Bodies* (University, Faculty, Study Course) to the findings of the survey, with special attention given to the methods and mechanisms of redesigning and upgrading the tools and the content of the didactics.

There is no doubt that for the first five, in particular, of these surveys envisaged as complementary to the TES, it is almost essential that studies should be conducted on an individual basis, allowing linkage of the information relative to each student, since this is the ideal condition for giving consistency and systematicity to an organic evaluation process. We think that, in effect, even studies on aggregate data alone – relative to single survey – can provide particularly interesting elements for analysis. The aim must be to ensure that the various surveys dialogue effectively, on the same subjects and on events and considerations that are different yet integratable.

Before moving on in detail to new proposals, it will be noted that certain of the surveys listed above are already in use at many universities, although one has the impression that they are often implemented in parallel, with no stated aim of aiming them toward a single goal, and that just as often they are entrusted to different administrators, all jealously guarding their respective areas of independence and unwilling to see their labors coordinated – whether on design, concepts, techniques and methods of delivery, or scheduling, or necessary linkages – within the scope of an Overall System Evaluation Plan.

With these two basic principles in mind, accordingly, one can reasonably identify the nature of the product that ought to emerge from the evaluation process, precisely establishing (1) *the object/s of the survey;* (2) *the operating conditions conducive to its implementation;* (3) *the essential references that can define it*, while at the same time stating the inherent advantages and limits. As regards the first element, I feel we have to accept a long-term goal, still too far off to be achieved by all the universities, which stems from the desirability of setting out an integrated and entirely coherent analysis for each *study course*; a transitional solution might be provided by the temporary need to render the analysis of each *faculty* more important and pertinent, while able nonetheless to allow for the effects of variability between the study courses offered. As regards the *second* element, it will be perhaps useful to

---

the largest numbers universities are Almalaurea and Stella). However, one cannot remain silent, just as on a matter of such great responsibility and importance one cannot agree with the distorted conception of independence that in effect sanctions individual solutions not comparable one with another, differing in design and in the nature of the information gathered.

state some imperative aspects more forcefully; in order to make the analysis of the selected object incisive and meaningful, it is essential:

I. to adopt the traditional quantitative approach connected with the use of the sources and the surveys conducted, although accomplishing a correct coordination of the single components, which currently are working separately;
II. to accept and develop the qualitative approach, integrating it with point I. and taking into consideration written materials,[8] in-depth interviews, consultation of qualified witnesses, reports of debates and comparisons drawn on the findings of the evaluation process;
III. to make mandatory the production of an annual self-assessment report for the structure in question, the content of which to be prescribed in an explicit and ordered fashion[9];
IV. to favour the notion of *peer reviews*, not least to enable discussion of didactics approaches and the relative operating mechanisms, and the respective results;
V. to be aware of the need for access to substantial investments – intellectual as well as financial and material;
VI. to be equally aware of the need to maintain centralized responsibility for the definition of evaluation procedures, and likewise for testing of the final product.

In the case of the *third* element, it will be as well to move on directly to the next paragraph.

1.3.2 The Proposal: A First, Almost Utopian Design

Following the stated principles and taking into account the conditions just mentioned, the methodological criteria that should be adopted in an ideal design, careful to the method but also to cognitive/cultural aspects, could in my opinion be these:

I. to get longitudinal data (linking the information provided by each student census from enrollment to graduation, and possibly *post-graduation*: this means, firstly, coordinating surveys conducted currently in different ways[10]); a design integrating different surveys would, amongst other things, avoid the need to ask repeatedly for the same information and could encourage students to participate more willingly[11];

[8] Minutes, committee reports, resolutions, etc.

[9] To this end, it might also be possible to agree on a principle of rotation in identifying the courses/faculties to consider, not least in order to avoid rendering the business of analysis bureaucratic, formal and excessively burdensome.

[10] Often these surveys do not "speak to one another", designed as they are to provide analyses on aggregated data but proving unsuitable for the examination of pathways and intervening factors and for characterizing student careers.

[11] One is aware obviously that this notion rests on the need to be in possession of the names of individuals taking part in the survey so as to be able to link items of information deriving from

II. to conduct a statistical survey, either on students or on events characterizing the university experience; to this aim, in our view, a decidedly novel strategy for the TES might be to conduct the survey just once a year, typically during the period of transition from one academic year to the next. In this way it is possible to gather information relating to attendance at all the lessons, participation in final exams, passes or failures, so that the census would gain the perspective of differential approach in its analyses, according to the specific behavior of the single student. On one hand this proposal is cost-reducing for the administrative aspects regarding the questionnaire collection procedure, and, on the other hand it would also guarantee opportunities to examine the differential capacities for judgment shown by students with regard to the courses[12];
III. to ensure the involvement of students not only by giving notice of the survey but above all by seeking suitable forms for feedback of the results and for linked comparison between stakeholders;
IV. to keep the survey under a central management structure, thereby ensuring procedural correctness, uniformity of conduct, and comparability of final findings;
V. to promote "official survey days" with the aim of emphasizing the importance attributed "officially" to TES initiatives;
VI. to promote and standardize Faculty and University days, open to the public, for presentation and discussion of the analyses and the reports with a notice posted in advance;
VII. to construct a National Evaluation Report on the basis of material generated at every university, integrating the teaching dimension with other components of academic life, as a way of demonstrating the collective sense of responsibility toward the subject and taking the initiative away from external subjects or agencies, often business or scandal oriented, which in recent years have been more inclined to indulge in *épater le bourgeois* than to present measured and reliable system evaluation.

Many advantages could come from this overall approach. Looking first just at some political and cultural aspects, it is possible to mention:

I. the existence of an entity exercising general control over the process of shaping the entire evaluation activity, and in particular, ensuring balanced management and treatment of the TES data procedure;
II. a declared and direct acceptance of responsibility by the Governing Bodies of the single academic institutions;
III. affirmation of the principle of transparency, an indispensable element of the evaluation process;

---

them separately; moreover, the feeling is that increased knowledge of the purpose, and particularly the usefulness and effective utilization of data, could ensure a greater willingness to provide the elements required in order to link the surveys, if necessary adopting suitable measures to safeguard privacy.

[12] Likewise in this instance, various polling procedures and techniques would be considered (and experimented initially) in the quest to maximize the participation of potential respondents.

IV. the objective advantage of comparison based on uniform working conditions;
V. more generally, the soundness of action policies (local and national) is supported by principles of shared responsibility and uniform methods of observation.

The sharing of all the process of the TES evaluation could provide advantages of operative and cognitive types. A good practical solution consists in just one annual survey including the entire experience[13] of the academic year just concluded:

I. less organizational and financial strain of the survey, hence a reduced effort in polling (and harassing) the students[14];
II. gathering opinions from students who have experienced various forms of participation in learning activities: attendance at lessons or not, attendance at other forms of tuition (workshops, practical sessions, etc.) or not, taking exams or not, either passing or failing. Some cognitive value would be gained from the possibility of making up a differential analysis according to different profiles;
III. with similar cognitive aims in view, the chance to measure the differential capacities to evaluate shown by students, according to their personalities and career choices, and to the type of the teaching activities surveyed[15];
IV. the guarantee of being able to conduct differential analyses, along time (comparison between years) and space (between universities, faculties, similar study courses);
V. linkage with further evaluation with the job placement data.

1.3.3 A Possible Design

We are fully aware that the question is delicate and that structural modifications with excessive strength could even have the effect of stalling the system now in action.

[13] Entire experience, in the sense that the student may have attended lessons and lectures and sat the relative exams, or attended lessons without sitting the exams, or possibly enrolled for a given course but neither attended the lessons nor taken the exams.

[14] The comment of note 15 applies here too, as concerning the need to continue experimenting with different methods of polling in order to secure convincing levels of response. In recent years moreover, this approach has already been adopted in some universities, trialing web survey techniques, for example.

[15] These dimensions are totally neglected by the existing survey design, and this allows doubters (understandably) to deny the very value of TES, insisting on the one hand that students are ill-equipped to make judgments, and on the other, that the judgments they form may be too vague, with an underlying attitude (positive, negative, derisive) that colors all of their answers indiscriminately. The possibility of separating out individuals who are motivated, aware and responsible from those who treat the survey as a chore – likewise the possibility of linking each survey to the characteristics of the students – would seem to represent a step up in value that renders the survey particularly effective and gives sense to the exercise.

We feel it will be useful to outline the minimum or preliminary conditions for a move toward a more organic design (just on the basis of my own personal ideas). Here too, I should begin by stating the principles we see as desirable: (a) we are convinced as to the importance of centralizing responsibility for the definition of rules and procedures determining organization of the eventual national evaluation programme[16]; (b) we discard the "all at once" approach in favor of a more relaxed (and practicable) "layer by layer" strategy; (c) we feel it is essential to encourage experimentation, both as regards the procedures, tools and techniques adopted in conducting surveys, and more especially as regards the methods of analysis used on the collected data, which should in time become demonstrably more focused and selective, producing information that will be pertinent and consequential; (d) we see it as essential to develop, on the one hand, the ability for internal and external communication not only of results but also of their effective value, which will always be sharply determined by the singular operating conditions adopted, and on the other, the need to engender osmotic interaction between institutions, so that experiences and the results of innovative trials can be shared. In short, these considerations of mine would lead ideally to acceptance of the principle of steady progress in the process of consolidating the TES, acknowledging its intrinsic weaknesses but welcoming its usefulness as an indicator for the future. To conclude, then, I feel that in the current scenario – not least given the lengthy period during which TES has been adopted and since become supinely mechanical, betraying a slow but clear loss of interest, and giving too few useful results – we are prompted to make certain initial adjustments that should focus on:

I. standardizing the questionnaires for the different types of survey more fully and widely, and likewise the times and methods of polling, imposing common procedural rules and uniform checks[17];
II. regularizing the indicators for comparison, adopting standard methods of production;
III. trialing survey approaches that will also collect information and judgements on the taking of exams;
IV. widening survey aggregates to include enrollees, drop-outs, undergraduates, teaching staff and outside contacts;
V. preferring and rewarding initiatives that envisage the integration of surveys;
VI. illustrating the decisions taken as a result of evaluation, and thereafter, rechecking the effects of any changes introduced.

[16] I have no fixed ideas on this, but would give added value to the notion that the main active player might be the CRUI, willing nonetheless to accept the cooperation of MUR and its agencies in a watchdog role.

[17] In particular, I would emphasize the notion that the survey should be conducted by a centralized institution and not entrusted to single faculties or study courses on an ad hoc basis, given the associated risk of irreconcilable variations in information-gathering.

It will be appreciated that this is an almost minimalist outlook and that the primary concern is to raise the credibility of the evaluation; in effect, the ultimate purpose of TES is to be useful, even if the ways in which it is packaged cannot be regarded as optimal. This, in the absolute conviction (and hope) that the application of TES will continue to *unsettle* and promote *conflict* by virtue of the genuinely *dramatic* role it must play to ensure the world of higher education remains constantly in pursuit of improvement.

Chapter 2
The Assessment of University Teaching by Students: The Organizational Perspective

Luigi Enrico Golzio

2.1 Assessment in Organisations

Assessment in public and private organisations is a process or series of activities concerning the planned activities (of the organisation, group or individual) carried out in a formal manner, for the purpose of reaching an informed judgement, based on research, data processing and the interpretation of verifiable information, communication and negotiation between the organisational actors involved in the process. In public and private companies, assessment is a process adopted in the management of:

- systems of planning and monitoring (budgeting). The assessment consists of an evaluation of the efficient use of the resources assigned to the various organisational units in relation to management objectives laid down in advance;
- systems for the assessment of individual performance. In this case the assessment is an evaluation of the achievement of individual objectives laid down in advance, linked to the allocation or withholding of pre-defined incentives.

An assessment may therefore be classified on the basis of the objectives that give rise to and justify it.[1]

Organisational assessment concerns two alternative objectives: the development and the monitoring of the organisational behaviour of actors. *Monitoring assessment* is an evaluation of the performance delivered in relation to the expected level of performance, aimed at verifying compliance with the agreements, rules and responsibilities of the individual actors or organisational groups, and resulting in the allocation or withholding of resources, incentives or sanctions.[2] On the other hand, in terms of training and development, *training and development assessment* results in an evaluation of the services provided in order to enable the individual undergoing

L.E. Golzio (✉)
Dipartimento di Economia Aziendale, Università di Modena e Reggio, Emilia, Modena, Italy
e-mail: luigienrico.golzio@unimore.it

[1] On organisational assessment see [1].
[2] On types of organisational assessment see [8].

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_2,

assessment to gain insight into his or her shortcomings, with a view to improving performance in the future. Table 2.1 below provides an overview of the characteristics of the two types of assessment.

The two types of assessment may be further distinguished by the type of contract relating to the objectives agreed between those performing and undergoing the assessment, the salient characteristics of which are shown in Table 2.2 below.

The organisational aims and the types of organisational contracts are the two elements that mark the distinction between monitoring assessment and training and development assessment. Monitoring assessment is imposed on the individual making the assessment by the organisation, as a hierarchical responsibility, and this is reflected in relations with the individual who is subject to the assessment. This individual may fear the assessment because the (uncertain) outcome has implications in terms of rewards and sanctions. This type of assessment may also be problematic for the person carrying out the assessment, who is required to play the part of the judge. In the case of a negative outcome, it may have an impact on relations with the person subject to the assessment: these relations may be ongoing and may have an impact on the performance of the organisational unit on which the person carrying out the assessment will subsequently be judged.

Table 2.1 Types of organisational assessment

Organizational variables	Monitoring assessment	Training and development assessment
Aims	Monitoring of performance	Improvement of performance
Contract	Transactional contract	Relational contract
Game theory	Zero sum, win/lose	Non zero sum, win/win
Communication	Defensive	Open, problem-based
Person carrying out assessment	Judge	Mentor
Person subject to assessment	Defendant	Partner
Timeframe	Past	Future

Source: author's own data.

Table 2.2 Types of organisational assessment

Contract characteristics	Transactional contract	Relational contract
Goods exchanged	Economic goods	Economic and emotional goods (trust, esteem)
Obligations	Specific	Generic, ambiguous (flexible)
Timeframe	Predetermined, short-term, static	Indeterminate, medium-to long-term, dynamic
Area of acceptance	Narrow	Broad
Involvement of the parties	Limited in material	Pervasive and comprehensive
Control mechanisms	Objective	Objective and trust-based
Performance	Objective and observable	Subjective and internalised

Source: author's own data.

Training and development assessment, although initiated by the organisation, is not imposed, but left to the discretion of the actors involved. Such assessment is seen as desirable by the person subject to it, who is prepared to take well formulated criticism from the person carrying out the assessment in order to improve his or her performance and to acquire new skills. The assessment is also accepted by the person performing it, because in relations with the person subject to the assessment, the role is that of mentor (providing support and assistance), protecting and improving relations with the person subject to the assessment and therefore also his or her contribution to the performance of the organisational unit.

The two types of assessment are distinct and alternative, but at the same time, they may take place in parallel. Training and development assessment is, albeit only in part, a performance assessment (the judgement expressed concerns individual merit); performance assessment is necessarily also a form of training (the evaluation is useful for learning and improvement). The element that distinguishes the assessment and legitimates its alternative function is the underlying organisational objective (why it is carried out) that characterises all the remaining organisational variables.

2.2 Assessment by Students in Italian Universities

Assessment by students presents certain specific characteristics that it is worth examining. The procedure takes the form of a monitoring assessment. This is required of all Italian Universities, as laid down by Act no. 370/1999, Article 1(1) of which requires Italian Universities *to set up an internal system of assessment of the teaching programmes*. Article 1(2) requires the assessment unit *to carry out a periodic survey of the opinions of students about the teaching programmes and to submit a report to the Ministry of Education, Higher Education and Research and the national assessment unit (CNSVU) no later than 30 April each year*. The Ministry uses student evaluations for decision-making in relation to two matters: the setting up of courses and the allocation of funds (the 3-year planning fund). In particular, student evaluations are used as:

- a quality assurance instrument, specifically as an indicator of effectiveness (the level of satisfaction of the students in relation to specific courses, pursuant to Article 1(2), Act no. 370, 19 October 1999, for the approval of courses to be implemented (Ministerial Decree no. 244/2007, Ministerial Decree on the necessary requisites for the setting up and implementation of courses);
- a quality indicator (Article 11(3), *the percentage of courses in which the evaluation of the students is above the national average, in relation to the faculty groupings defined in relation to the provisions of Annex A.2, Ministerial Decree no. 362/2007*) for the (ex post) evaluation of the results of the implementation of the University programme for the 3-year period 2007–2009 (Ministerial Decree no. 506/2007 (Indicators for the assessment of the results of the 3-year programme 2007/2009).

The assessment by the students concerns the quality of the service provided by the lecturer, the Faculty Council and the Academic Senate. It seems appropriate and useful to refer to the concept of service (and to the models proposed by the related scientific research) since teaching is a set of intangible activities, utilised by the client (the student) who pays for the service in a regulated market (in which the academic qualification has a legal value).[3] As a result the judgement expressed by the student is necessarily subjective, but no less reliable for this reason. The service consists of the performance of certain activities supplied by organised actors. In the specific case, the evaluation by the students concerns the package of services supplied by a range of actors, as individuals and groups.
To be precise:

- the individual lecturer is subject to assessment with regard to the delivery of the *principal service* consisting of teaching (for example, with regard to the planning of the course (contents and teaching methods), the amount and quality of the teaching materials, the clarity of the explanation, the level of interest aroused in the students, supplementary teaching activities, availability);
- the faculty members in the Faculty Council with regard to *teaching resources*, the use of which is of central importance in the individual Faculties (teaching programme, lesson timetable, number of examinations, tutorial services, accessibility of the library, and so on);
- the academic and administrative staff who serve on the Academic Senate with regard to *auxiliary resources* (the teaching facilities, such as the number of lecture rooms, laboratories, computer workstations and libraries and their quality, the services for providing support for students (bursaries), the administrative facilities (student registration offices, placement, career guidance, and so on). The quality of these auxiliary services, together with that of the central services, is to induce the student to choose a particular university rather than those with which it is in competition;
- the group of academic and non-academic staff who make up the internal assessment unit, which, pursuant to the legal provisions, is responsible for the quality of the services for the data collection, processing and dissemination of the student evaluations in relation to the student body, the University, and the Ministry.

In short, the assessment by the students is an organisational process that is intended to evaluate the services provided by multiple actors in the University (individually and in groups).

2.3 Assessment by Students as an Organisational Process

The assessment by the students considered as an organisational process may be examined in three analytical perspectives: the *measurement*, the *cognitive*, and the

[3] On the concept of services and management of services see in particular [9, 15].

strategic perspective.[4] Priority given to one of the three perspectives will have implications for the quality of the assessment, the planning of the process, the organisation of data collection, the processing of the assessment data, and the dissemination of the results of the assessment.

The *measurement* perspective conceives of the assessment as individual decision-making by the person performing the assessment (the student), who is required to formulate an accurate assessment with an adequate instrument of measurement, i.e. the assessment form or questionnaire. In other words the assessment is a problem of measurement that concerns in particular the scale of evaluation to be utilised for the purposes of the reliability of the assessment (with regard to the stability of the assessment by the students), as well as the type and number of questions, and the methodology for processing the data collected. This perspective, based on the assumption that the assessments carried out are reliable from a technical point of view, was found to be of limited interest when it was shown that the format does not influence the quality of the evaluation in a consistent manner, or that there is no particular format that is significantly better than others.[5]

The *cognitive* perspective places emphasis on the study of the cognitive processes of the person carrying out the assessment, because the quality of the assessment depends on these processes. The student has a perception of the teaching environment (the teachers on the individual courses, the other students on the course, the teaching rooms, and so on) and memorises these experiences in the form of cognitive structures (schemes, scripts, cognitive maps, prototypes, examples). They are utilised by the student in the perception of the stimuli transmitted by the lecturer. The assessment is the result of the codification, processing and interpretation of the stimuli transmitted by the lecturer that the student commits to memory. In a cognitive perspective the limited information about the performance of the lecturer (for example, the preliminary stages in which the teaching material is planned and prepared) and the limited powers of reasoning of the student, are overcome by the cognitive strategies adopted. They consist in the use of heuristic principles in decision-making (those pertinent to evaluation are *ready availability*, *representation* and *anchoring*) or the use of cognitive short-cuts giving rise to problem-solving in a simplified form, without having access to all the necessary information and the computational ability necessary to process it. The use of heuristic principles in decision-making may give rise to bias in the assessment (in the specific case the effects are *indulgence*, *strictness* and *proximity*).[6] In order to reduce or prevent the assessment errors depending on bias, the cognitive perspective proposes two measures to improve the reliability of the evaluation: the improvement of the capacity of judgement and the more efficient utilisation of the information held in the memory of the person carrying out the assessment. The first measure can be implemented

[4] For a survey of research perspectives relating to the assessment of performance in public and private companies, reference may be made to Fabbri [4].

[5] Landy and Farr [12].

[6] On heuristic decision-making devices and bias, see [16].

by means of training courses aimed at enhancing the understanding of the extent of the service to be assessed, and the proper use of the assessment scale. The second measure consists of the keeping of a diary by the person carrying out the assessment in order to make more effective use of the information available about the service to be assessed.

The cognitive perspective gives priority to the evaluation of performance in employment relations between managers and subordinate employees, whereas assessment by students presents particular characteristics distinguishing it from this situation, as noted above. With regard to assessment error, the research by Schein and Hall on assessment data collected from two groups of students attending undergraduate courses (with no work experience) and master's courses in management (with previous work experience) suggests that assessment by students is subject to limited bias, and as a result the quality of the evaluation is not undermined.[7] The two groups carrying out the assessment, differentiated in terms of experience and therefore also memory, provided convergent judgements on the qualities distinguishing a good lecturer from a bad one, that is to say, intellectual and communicative ability, energy and personal enthusiasm, and the level of commitment and responsibility in performing the teaching role (providing support for learning). In particular, the assessments of good lecturers (those from whom the students had learned the most) were more extreme than those for the bad teachers.

The limit of the two perspectives outlined above is that they conceive of assessment by students in isolation from the organisational context in which it takes place. As a result the lack of reliability of the judgements expressed by the students may be explained by technical shortcomings relating to measurement (for the *measurement* perspective) or by cognitive limits, in particular decision-making bias (in the *cognitive* perspective). In other words, the quality of the assessment may be undermined by unintentional factors which the person performing the assessment is unaware of, according to these two research perspectives. In fact, as already noted, assessment concerns the performance of actors and groups of actors with interests that are partly shared and partly distinct with regard to the results of the assessment expressed by the students and their dissemination. In the logic of monitoring assessment, the judgements of the students can be and are utilised as a means of influence or power among the actors concerned.

The *strategic* perspective considers assessment in terms of *organisational games*, including power issues.[8] This perspective conceives of any organisational system (including individual universities) as a political system, of an indeterminate type, never completely controlled or regulated, underlying its existence as a social system. It is a universe characterised by conflict in which actors make rational use of the sources of power at their disposal. In the organisational system there are no common objectives, but only shared objectives because the division of labour

[7] Schein and Hall [20].

[8] On the concept of games and the strategic analysis of power, reference may be made to Crozier and Friedberg [2, 6].

assigns to each actor/group a particular and limited objective. Each actor has an interest in considering limited objectives as general in order to give greater value to their contribution to the survival and development of the organisation. Reference may be made in this connection to claims by faculty members about the superiority of their courses or area of study.

Organisational rules delimit the area of uncertainty of individual and group behaviour, but are never completely binding on individual actors, who always maintain a certain degree of freedom, and the possibility to negotiate. The degree of freedom of individual actors is a source of uncertainty in relation to their behaviour in dealings other actors and the organisation as a whole. Every actor therefore has a degree of power over the other actors, that may be used to reduce the interdependence between the actor and the others. In this organisational context, on what conditions is it possible to achieve cooperation among the actors who carry out interdependent activities, enjoying a degree of freedom and pursuing divergent if not contradictory interests, for the realisation of shared objectives? In order to achieve negotiated cooperation between the actors, the *strategic* perspective proposes the concept of *organisational game*. This is the instrument devised by the actors to regulate cooperation between them because it conciliates freedom and constraint. Players remain free, but in order to win are required to:

- comply with the rules (because they assure a continuity of relations between the actors);
- partially satisfy the expectations of others. Each actor exerts power over others in a reciprocal manner, and allows others to exert power over him/her. Other actors become a limitation;
- adopt a rational strategy in relation to the nature of the game.

In conclusion, the game is always a matter of cooperation, and the outcome is the achievement of the shared objectives of the organisation. The rules of the game determine the possibility of winning or losing, delimiting the range of winning strategies that may be adopted by the actors. Each actor behaves simultaneously in order to limit the other actors, taking advantage of the opportunities in the *game* to improve their situation (offensive strategy); deal with their attempts at delimitation by widening their margin of freedom and their powers of action (defensive strategy). There is no irrational behaviour, but *strategic* behaviour, that is stable and autonomous, and the regularity of this behaviour needs to be identified and observed empirically in relation to the organisational context.

The crucial factors of uncertainty for the organisational system relate to four sources of organisational power available for each actor, that may be defined in connection with the organisational structure of each University as follows:

- possession of a particular skill (relating to research and/or teaching), either professional or contextual, concerning relations (regarding the specific organisational structure of the University and the higher education system) that would be difficult to replace;

- influence over relations between the University and its environment (local, ministerial, and so on) determine the balance of power, due to the indispensable role of the University as an intermediary and interpreter between different and at times conflicting agendas (suppliers, students and their families, businesses, institutions);
- the control of communications and the flow of information, since the place occupied in a network of communications and the means of transmission of information (that may be delayed, filtered, or manipulated) has an impact on the ability of the recipient of the information to act. Communication may take place in return for safeguards and favours;
- the existence and implementation of organisational rules. The normative framework limits the powers of those in a subordinate position, but also the arbitrary power of those at the upper end of the hierarchy, since they may not have the means to obtain from their subordinates any more than the rules provide for.

The *game* of assessment by students requires the involvement of various actors, both individual and collective, in the University. Each of them may count on sources of power giving rise to a degree of uncertainty in their relations with other actors, and to specific forms of behaviour (defensive or offensive) as summarised in Table 2.3.

For each actor the *game* of assessment by students represents an opportunity or a threat to their area of autonomy. The resulting organisational behaviour gives rise to alliances among those with common interests. The virtuous Faculties will attempt, in a unified fashion, to benefit from the allocation of resources, at the expense of the inefficient ones. The assessment unit should be able to count on the support of the Rector of the University and the student representatives to publish the results of the assessment, not just in aggregate terms but also course by course.

In practical terms the *strategic* perspective conceives of assessment as a process of measurement and communication that is rooted in and characterises the organisational relation between those carrying out the assessment and those who are subject to it. It is based on the assumption that all the actors involved in the process are active participants who take part in the *game*, and that within the regulations, pursue their personal agenda in a discretionary manner.

In the strategic perspective the quality and reliability of the assessment by the students of the performance of their teachers reflect the conscious choice of those carrying out the assessment to express or not to express their *secret knowledge*[9] about their teachers' performance.

In other words the strategic behaviour of the student not to express an opinion, or to supply an unreliable opinion, should be seen as a conscious choice that is a matter of convenience (because it safeguards or enhances the relationship with the lecturer), rather than due to a lack of skills or the ability to carry out an assessment. The failure to provide an assessment or to provide one that is unreliable reflects the position of conflict of the student, which, if collective, gives rise to the need for

[9] Expression Taken from R. Normann, op.cit.

Table 2.3 Actors, sources of power and strategic behaviour

Actors	Sources of power	Defensive strategy	Offensive strategy
Faculty member	Control over the the method of assessment of student performance (examinations)	Attempts to interfere with the collection of student assessments Opposition to the publication of individual assessment	Collusion with students in the formulation of assessments Improvement of teaching
Faculty Council	Control over the dissemination of the results of the assessment Faculty teaching regulations	Attempts to interfere with the collection of student assessment Opposition to the publication of individual assessments	Social control over the consensual rules relating to common teaching standards
Academic Senate	University teaching regulations Control over dissemination of the results of the assessment Allocation of resources	Formal compliance with legal requirements	Incentives for Faculties with positive results and sanctions for inefficient ones
Student	Compulsory collection of assessment by students Utilisation of assessments for ministerial approval	Failure to provide assessments Submission of unreliable evaluations	Improvement of teaching conditions (teaching programmes, resources, methods of assessment)
Internal Assessment Unit	Institutional intermediary with the Ministry Control over the dissemination of the results of the assessment	Organisation of the collection of the assessments by the students within the time limit	Publication of the assessments by the students for each course and each faculty member

Source: author's own data.

organisational strategies that recognise it and make it explicit. The collection and dissemination of assessments of teaching programmes by students may reduce this conflict and result in changes to the organisation.

2.4 The Content of Assessment by Students

Students are asked to carry out an assessment of the teaching services provided by the University at which they are enrolled. In order to understand the content of the assessment that is required of them, reference may be made to the hierarchical model proposed by Kirkpatrick,[10] that distinguishes between four types of content

[10] D.L. Kirpatrick [11]. With regard to this model see also [14, 17, 18].

to be evaluated by participants in training courses, corresponding to four different levels:

- level 1: response
- level 2: learning
- level 3: organisational behaviour
- level 4: organisational results

The model states that each change brought about by the training programme (level) in turn produces effects on the following level on the basis of a *cause and effect* relation, and a hierarchical order (from level 1 to level 4). Specifically a positive response influences the motivation to learn; learning in turn gives rise to new behavioural expectations, leading to better results for the organisation.

The response reflects the degree of satisfaction of the participant in relation to the experience of the course.[11] *The response may be defined as the degree to which the participants liked the course. An evaluation of the response is similar to the measurement of those taking part in a conference, that does not include the measurement of whether they have taken part in a learning process*. The level of satisfaction with the teaching programme therefore does not provide a guide to the effectiveness of the course. The response expresses an evaluation of different aspects of the course: the degree to which it meets the expectations and needs of the participants, the topics examined, the lecturer, the teaching material, the degree to which it was perceived to be a welcoming experience, and practical aspects (teaching rooms, laboratories, facilities), and the other participants on the course. The response may change over time in relation to the experience of the participant. The response may be positive in terms of the topics dealt with even if they are considered to be of limited utility to the participant. Hence the need to evaluate the next level.

Learning may be considered in terms of the development, thanks to the course, of the knowledge of the participants (knowing), their skills (knowing how to do), and their attitudes (knowing how to be).

In the specific case of university courses it should be noted that the assessment of behaviour concerns in particular the evaluation of the acquisition of *explicit* knowledge, in other words objective knowledge (the result of scientific research) that is abstract and may be codified, formalised and therefore transferred and utilised by the participant.[12] An evaluation of behaviour is functional to understanding the effectiveness of teaching methods utilised during the course. Learning does not necessarily lead to the automatic application of what the students have learned in class. Hence the need to evaluate the next level.

Behaviour consists of the transfer in the workplace of knowledge, skills (knowing how to do) and attitudes on the part of the students. Behaviour is situated in the workplace and not in the classroom; it is therefore influenced by the organisational

[11] Kirkpatick, op. cit, pag 3.

[12] With regard to the concept of knowledge in the process of organisational learning, see [5].

context that can facilitate or inhibit the types of behaviour expected by the organisation. Behaviour is difficult to measure because it is hard to predict when and whether it will take place. However, it is important to verify it in order to monitor the effectiveness of training. Hence the need to evaluate the next level.

The results achieved by an organisation consist in its overall performance. In the specific case the results of the activities of the University are of two types:

- efficiency in the use of resources (lectures, lecture theatres, technical and administrative staff) in supplying teaching services under the supervision of the Ministry and the national assessment unit (CNVSU);
- the contribution to the creation of value for the end-user (graduates, businesses, public bodies) by the means of the quality of the knowledge transmitted to the students.

Each of these elements of training subject to assessment requires the use of specific monitoring instruments, as shown in Table 2.4 below.

Kirkpatrick's model has been integrated by Hamblin,[13] who argues that the content evaluated at each level is useful insofar as it can be compared with a corresponding *initial objective* (teaching). The initial objective laid down in advance and relating to the response determines certain choices in teaching programmes, that will elicit a response that may be appreciated and compared with the initial objective. In the specific case in order to fully appreciate the student responses, the individual faculty members, the Faculty Council and the Senate should lay down the initial objectives in advance in terms of response, learning, behaviour and results, that is to say the teaching objectives or *descriptors*.

The positive aspects of the Kirkpatrick model may be summarised as follows:

- responses are recorded at a low cost at the end of the courses, since they are based on pencil and paper questionnaires;
- the evaluation of teaching is feasible when it is a matter of assessing practical knowledge and abilities, as in the case of technical education;
- the evaluation of organisational behaviour, specifically when the types of behaviour that are expected and the foreseeable exceptions in the interaction between persons and machines, is possible *at low cost*.

Table 2.4 Elements and instruments of assessment in training

Element to assess	Instrument adopted
(1) Response	End-of-course questionnaire
(2) Learning	End-of-course examination/final dissertation
(3) Behaviour	Ex ante and ex post performance evaluation
(4) Organisational result (University)	System of indicators of the efficiency and quality of teaching and benchmarking

Source: adapted from Kirkpatrick,1960.

[13] A.C. Hamblin [10].

The limits of the model are constituted by:

- the fact that the nature of the model is deterministic. It has not been proved scientifically that the positive outcome of an assessment at Level 1 determines the chances of success in the later levels;
- the final evaluation of the training provided (in class or in the workplace) conceals any critical aspects in the initial phases (analysis of educational needs and training) if these needs are not laid down at the planning stage of teaching programmes;
- the evaluation of behaviour and the organisational results is critical for the available resources, costs and the time difference between classroom teaching and the workplace.[14]

This overview of the positive aspects and the limits of the Kirkpatrick model makes it possible to make an informed evaluation of the positive aspects and limits of the responses to the teaching programmes undergoing monitoring. First of all it must be noted that, together with the learning results, the evaluation by students is the only formal assessment carried out in all Universities, that are not in a position (or do not deem it to be beneficial) to evaluate all the other aspects of the teaching process. These two aspects are useful for an evaluation of the teaching provided and the learning that takes place. The assessment by students is also the aspect of teaching programmes that is most widely measured among the four organisational levels due to its low cost, facility of implementation, speed of feedback provided by the participants, and above all, the fact that it is an indicator of quality in a perspective of customer satisfaction. The assessment is a matter of perception, and therefore subjective, that reflects the experience of the participants in a situation of cognitive dependence, concerning the aspects of the course that they are aware of from direct experience, in other words the *context*, the *how* and to a limited extent the *what* (that they will be able to fully evaluate after the course in the workplace). It is a judgement limited to the relation between the faculty member and the student, necessarily limited to the processes taking place in the lecture room, and it is of value to both parties as they seek confirmation of their respective roles and behaviour.[15] In a service management perspective, assessment by students is a useful form of feedback for the lecturers. This is all the more the case when the lecturers (and the organisation designing the questionnaire) state the objectives and the purpose of the evaluation by the students in advance.[16] Finally it may be argued that a positive response will have a positive impact on the atmosphere in the lecture room and on the later stages of the programme, although there appears to be no scientific evidence in support of this claim.

Student responses are the aspect that has attracted least attention from academic researchers, who tend to focus on the other levels. This explains the current value

[14] On the limits see [3].
[15] In this connection see [13].
[16] On these points see [19].

and longevity of the Kirkpatrick model. The scientific literature has identified three key aspects to explain the response in terms of satisfaction by participants at the end of the course: the perceived effectiveness of the course; the perceived utility of the course; the perceived effectiveness of the performance of the lecturer.[17] These three determinants in turn are explained or include further specific items that have an impact on them.

The perceived effectiveness of the course includes the course facilities (accessibility, coffee break facilities, suitability of the lecture rooms, air conditioning, acoustics, furnishings, teaching resources such as blackboards, whiteboards, and simulators, the chance to communicate, Internet workstations, and so on); the organisation of the course (timetables, number of sessions, teaching load, total length of the course); the quantity and quantity of the teaching material.

The perceived utility of the course of study may be explained by the perception of acquiring competences (knowledge and skills) necessary for performing work (currently in progress) in a more effective manner, and/or to improve one's role in the organisation (prestige, self-confidence, and so on); the perception of personal growth or development for the long term, either within or beyond the organisation; and the perception of a proper balance between theoretical and practical aspects of the course.

The perceived effectiveness of the performance of the lecturer depends on mastery and expertise in the topics examined; the teaching style adopted during lectures; a consistent and varied use of teaching methods (lessons, guided discussion, group work, role play, case studies, workshops) and effective time management (complying with the timetable).

A study by Giangreco, Sebastiano and Peccei aimed to verify in an empirical fashion the results of existing scientific research, and attempted to answer the question: which of the three factors (the perceived effectiveness of the course of study; the perceived utility of the course of study; the perceived effectiveness of the performance of the lecturer) identified in the scientific literature had the greatest influence in terms of the satisfaction of the course participants?

The study was carried out using 2,697 completed questionnaires of the 3,698 distributed, representing 72.9% of those taking part in the courses in the province of Varese funded by Fondimpresa, the bilateral inter-category fund (set up by the social partners Confindustria and CGIL, CISL and UIL), in the context of the PISTE programme (process innovation, new technologies, development of management systems, marketing). The questionnaires were filled in by high-school and university graduates, blue- and white-collar workers, and middle managers in 208 undertakings, of all sizes, from micro enterprises (less than 10 employees) to medium-sized to large companies (more than 250 employees). The period in which the courses were run was from March to December 2005, during which time 7,230 h of training were provided as part of 307 training modules.

[17] For an in-depth survey see [7].

With regard to the research methodology, the overall satisfaction with the courses is a dependent variable explained by three independent variables (the perceived effectiveness of the course of study; the perceived utility of the course of study; the perceived effectiveness of the performance of the teacher). The end-of-course questionnaire consisted of 13 items, three of which were related to the effectiveness of the course, five to the utility of the course, and five to the effectiveness of the performance of the teacher. The questions were based on a five-point Likert scale (from 1 = total disagreement, to 5 = total agreement). The hypotheses to be tested were examined by means of standard deviation and multiple regression.

The results of the research may be summarised as follows:

- the three perceived factors (the independent variables), although interrelated, are distinct in influencing the overall satisfaction of the participant (the dependent variable);
- the three perceived factors taken together have a significant impact on overall satisfaction;
- the utility of the course is the most useful predictor for overall satisfaction, followed by the effectiveness of the teacher and the organisation of the course;
- the performance of the teacher does not compensate for any shortcomings in terms of the content and organisation of the course; in the same way the quality of the contents and the organisation of the course do not offset any shortcomings in the performance of the teacher;
- the level of satisfaction recorded among the participants was on average higher for the courses with "soft"(relational) contents compared to those with "hard"(technical) contents.

The research outlined above, albeit within the statistical limits pointed out by the authors, provides material for discussion about the use and utility of assessment by the course participants and the need to ascertain whether it presents similarities to the assessment by students.

2.5 The Case of the University of Sassari

The case examined in the present study is based on the personal experience of the author in his capacity as President of the Assessment Unit of the University of Sassari. The case is of particular interest in that the Assessment Unit introduced the publication of the results of the assessment by the students not just at aggregate level for Faculty courses, but also at the level of individual courses for each lecturer. This is not the first time that results have been published in this way: the University of Venice was the first to take this step, but the experiment was immediately terminated due to the opposition of faculty members.

The evaluation of the courses by the students was carried out by means of the administration of a questionnaire, extensively used at national level, replicating the evaluation of teaching programmes adopted by the national assessment unit

(*Comitato nazionale di valutazione*) in 2002 (Document no. 09/2002) to safeguard the homogeneity and the comparability of data at national level.[18] In the academic year 2006/2007 1,360 university courses were subject to monitoring out of an estimated 1,659 courses activated. The rate of coverage was 82%. The objective to be achieved over the next two academic years is to bring this figure as close as possible to 100%. The questionnaires collected totalled 27,303, with 3.3 questionnaires collected for each active student. The result of the evaluation was that 90.71% of the university courses received a positive evaluation, whereas 9.29% were given a negative evaluation.

From the very beginning the results of the assessment were published and commented on at aggregate level for each Faculty and University, and reported to the faculty members responsible for the courses assessed, and to the Deans of the Faculties in the form of disaggregated data. In this connection the guidelines relating to the teaching responsibilities of faculty members state that *The Faculties and teaching structures involved shall publish the results of the teaching activity carried out by the faculty members, as shown by the findings from the Internal Assessment Unit of the University and by other forms of evaluation carried out by the individual Faculties and teaching structures*.

Experience has shown that the potential for the collection of questionnaire data has been developed in a limited manner. A survey carried out in recent years among Faculty Deans has shown that assessments by students have only a partial application. Further evidence in support of this claim is to be found in the repeated requests by student representatives in Faculty and university councils to provide more effective feedback in response to their observations.

The Assessment Unit, in response to the most recent requests put forward in a responsible manner by the student representative at the University Conference on Teaching Services, took the decision to make the assessments by the students for individual courses available in a transparent manner on an experimental basis. This decision was taken in order to make the exchange of information between faculty members and students more symmetrical, and to provide the University with reliable information for planning future teaching programmes, in order to develop the scientific community of faculty members and students in the various Faculties.

After informing the Rector and all the Faculty Deans, the Assessment Unit decided to go ahead with the publication of the results on the University website (showing the mean values recorded) in relation to the individual courses subject to assessment, starting from the academic year 2006–2007. Reflecting the experimental nature of the initiative, the Assessment Unit made provision, at least in this initial stage, for individual faculty members to be exempted from the publication of the results. On an experimental basis, access to the data relating to the evaluation of the students attending the courses is to be confined exclusively to students and

[18] In issuing Document 09/2002 the national committee adopted the proposal of a working party entrusted with the task of drafting a minimum set of questions to be adopted by all Universities. The working party consisted of M. Gola, B. Chiandotto, L. Fabbris, P. Massimi, N. Terzi, R. Viganò, C. Violani.

faculty members of each individual Faculty by means of a password issued to those entitled to access the data, in order to guarantee access to all the stakeholders in each Faculty, but not to external actors.

As a result, for faculty members who granted permission for their results to be distributed, it will be possible to examine (for each course subject to assessment by the students) the mean values obtained for each variable, for the academic year 2006/2007. With regard to faculty members withholding permission for their results to be distributed, the data to be made available (with the remaining data blanked out) will consist only of those variables not directly relating to the faculty member.

The decision of the Assessment Unit gave rise to contrasting reactions: alongside certain faculty members and faculties raising objections, there were others who gave their approval. At a practical level the *game* of assessment gave rise to responses that were perfectly comprehensible in a strategic behaviour perspective. It should be noted that the argument that nearly all the faculty members put forward to justify their refusal to distribute the data concerning the judgement of the students on their courses was the violation of privacy (of the faculty member). In this connection mention should be made of the rights and duties of university students as specified on the website of the Ministry of Higher Education and Research, Title III, Article 86, page 86, that states: *The publication of results deriving from the analysis of the assessment forms, for each course of study, shall be carried out for all the Degree Courses of the University by suitable means. The results of the assessment forms filled in by the students shall be evaluated by the Assessment Unit of the University, with regard to the overall functioning of the University, and by the Joint Committee on Teaching, with regard to the provisions concerning the Faculties*.

Table 2.5 Approval by faculty members for the publication of their course assessments

Faculty	Number of courses actived	Number of courses assessed	Courses assessed as percentage of courses actived	Number of courses assessed – results not published	Percentage of courses assessed – results not published
A	227	189	83.3	20	10.6
B	174	149	85.6	6	4.0
C	144	142	98.6	59	41.5
D	127	89	70.1	6	6.7
E	89	80	89.9	73	91.3
F	75	49	65.3	1	2.0
G	194	133	68.6	13	9.8
H	150	134	89.3	3	2.2
I	54	46	85.2	1	2.2
L	258	196	76.0	23	11.7
M	106	79	74.5	11	13.9
N	81	74	91.4	33	44.6
Total	1679	1360	81.0	249	18.3

Source: Assessment Unit, University of Sassari.

In concluding this overview of the case of the University of Sassari, the figures concerning the granting or withholding of approval by the faculty members for the publication of the data concerning their courses is shown in Table 2.5.

The case shows all the dimensions of the assessment of university teaching by students in a University described in the first part of the chapter: the content of the assessment, the technical tools, the power strategies of actors involved in the process. The main consideration of the case is the following one: a multidisciplinary approach which weighs the assessment as an organizational game is feasible to ensure an efficient assessment of university teaching by students.

References

1. Bezzi C (2006) Che cosa è la valutazione. Franco Angeli, Milan
2. Crozier M (1969) Le phénoméne bureaucratique. Editions du Seuil, Paris
3. Cucchi M, Roncalli P (1991), Il processo di formazione nella prospettiva della teoria dell'azione organizzativa. In: Maggi B (ed) La formazione:concezioni a confronto. ETASLIBRI, Milan
4. Fabbri T (2001) La valutazione della prestazione. Sviluppo e Organizzazione, January–February
5. Fabbri T (2003) L'apprendimento organizzativo. Carocci, Rome
6. Friedberg E (1977) L'acteur et le système. Editions du Seuil, Paris
7. Giangreco A, Sebastiano A, Peccei R (2009) Trainee's reactions to training: an analysis of the factors affecting overall satisfaction with training. Int J Hum Resour Man 20: 96–111
8. Golzio L (1976) Aspetti organizzativi della Direzione per Obiettivi in impresa. Antonino Giuffrè, Milan
9. Gronroos C (1990) Service management and marketing. A customer relationship. Lexington Books, Lexington
10. Hamblin AC (1974) Evaluation and control of training. McGraw-Hill, New York, NY
11. Kirpatrick DL (1960) How to evaluate training programs: an abstract. J Am Soc Train Dir XIV 13: 3–9
12. Landy FS, Farr JL (1980) Performance rating. Psychol Bull 87:72–107
13. Maggi B, Melocchi L (1984) Formazione: analisi dei bisogni e verifica dei risultati. Sviluppo e Organizzazione, n. 84
14. Neglia G (1999) La valutazione della qualità della formazione: esperienze a confronto. Fondazione Giuseppe Tagliercio, Lupetti Editori di Comunicazione, Milan
15. Normann R (1984) Service management: strategy and leadership in service business. Wiley, Chichester
16. Piattelli Palmarini M (1993) L'illusione di sapere. Arnoldo Mondadori Editore, Milan
17. Quagliano GP (1979) La valutazione dei risultati della formazione. Franco Angeli, Milan
18. Quagliano GP, Carrozzi GP (1981) Il processo di formazione. Franco Angeli, Milan
19. Salas E, Cannon-Bowers JA et al (1999) Training in organizations: myths, misconception and mistaken assumptions. In: Ferris G (ed) Research in personnel and human resources management, vol 17. Jai Press, Greenwich, CT
20. Schein EH, Hall DT (1967) The student image of the teacher. J Appl Behav Sci 3(3):305–337

Chapter 3
University League Tables

Methodological Options for Ranking Systems: Censis Approach and Alternatives

L. Bernardi, P. Bolzonello, and A. Tuzzi

3.1 Introduction

Since 2000, the Italian Censis research institute has compiled, on behalf of *La Repubblica* newspaper, the *Grande Guida all'Università*, a report which ranks Italian universities and faculties according to their quality. With the 2008 publication, devoted to students enrolled on degree courses in 2008–2009, the *Guida* has gone into its ninth edition.

For the administrators and practitioners of the Italian university system it may have been embarrassing to find themselves appraised and classified (even with unflattering rankings) in a competition in which, at least at the beginning, they did not know they were participating. All the more so if the league table was drawn up by a private organization assuming "civic responsibility" to inform the public about the work of the university system, and which was commissioned by a newspaper, which might therefore be more interested in sensationalism than in encouraging virtuous behaviour.

But how convincing and reliable are the general design, criteria, data sources, indicators and rules used to construct the league tables? And what are the possible reactions of the universities and faculties? Attack or defence, rejection or acknowledgement, acceptance or a decision to construct an alternative ranking system?

This study examined the contents and the methods of the Censis report and assessed possible alternatives. It explored the difficulties in achieving a reliable ranking system and sought ways to refine the Censis model. After a brief description of the Censis model, the discussion focuses on "evaluating" and "measuring", and

L. Bernardi (✉)
Dipartimento di Scienze Statistiche, Università di Padova, Padova, Italy
e-mail: bernardi@stat.unipd.it

In the Italian version the title "Arbitro, c'è rigore?" played upon words and with the meaning of the noun "arbitro" (referee and arbiter) and "rigore" (severity and penalty): "Arbiter, is there any penalty?"

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_3,

identified indicators as the means to produce objective, appropriate, and comparable measurement.

Data furnished by Censis on 27 Faculties of Political Science for 2006 were used to construct alternative ranking tables by employing a selection of current methods of normalization, aggregation and weighting. The Censis league table was then reconstructed on the basis of the *Note metodologiche* (methodological notes) attached to the *Guida* [10]. The results were compared and contrasted in terms of alternative rankings of the 27 Faculties. Although caution is obligatory when interpreting the results (either for the low number of statistical units or the context of the elaboration was different from that of Censis), this study finalized with comparative analyses of the rankings obtained using the various techniques, and with some proposals for alternative composite indicators.

3.2 The Censis Ranking System

Every year the *Grande Guida all'Università* proposes rankings of the Italian universities and faculties.[1] We decided to analyse only the Censis ranking system of the 27 Faculties of Political Science. The *Guida* can be evaluated from two viewpoints:

1. a vertical one, on which comparison is made among faculties as a whole and by "areas[2]" of indicators;
2. an horizontal one, on which the strengths and weaknesses of each faculty are assessed.

The *Guida* proposes a ranking of faculties to assist future freshmen and their families in making a more conscious choice. In order to translate this evaluative goal into quantities, Censis identified five areas:

- *productivity*, which measures a faculty's capacity to guarantee the regular fulfilment of examination requirements of degree courses;
- *educational sustainment*, which comprises a balanced student/academic staff ratio, the provision of adequate facilities, suitable course programmes, etc.;
- *research*, which evaluates the capacity of academic staff to plan their research, and the probability of a student to have lecturers with good research experience;
- *academic profile*, which identifies faculties that endeavour to rejuvenate their teaching staff and enhance international relations;
- *international relations*, which measures the openness of faculties to international study opportunities both for their students and their teaching staff [11];

[1] Censis evaluates universities along four dimensions: services, study grants, facilities, and website. Faculties are assessed by means of composite indicators of five areas. It should be borne in mind that the university league table does not depend on the results obtained by faculties, and *vice versa*.

[2] Censis calls "family" each "area" of the university system.

- *attractiveness* of each faculty, in regard to other universities and faculties of the same field.

Three methodological considerations seem to be necessary before the analysis:

1. The five areas have changed over time. The *attractiveness* was included in the *Guida* only in the first 2 years [8, 9], and the *academic profile* was introduced in subsequent years (but is no longer present in the [12] report). Moreover, the set of simple indicators has changed from year to year.
2. Data were available for the year 2006, i.e. when the reform of the Italian university system (according to D.M. 509/99) was just consolidated. Consequently, indicators would not be affected by institutional changes.
3. The Censis approach implies an underlying – and non-explicit – compensatory logic whereby good performances on a particular aspect off-set negative results on a different one (as often happens when attempts are made to synthesis the diverse features of a complex concept into a single measure).

3.3 Indicators for Evaluation and Measurement

The main task of the statistician is to translate the characteristic features of a phenomenon into numbers by means of a sensible definition of a pertinent concept. In social research, the quantitative and qualitative evaluation of a given phenomenon (hereafter designated by a concept) consists in a procedure whereby the particular feature possessed by a particular statistical unit is determined [11] by a number (quantitative information) or a category (qualitative information). Quantitative information lies at a higher level than that of qualitative information, even if the qualitative information is its basis. The question *how much*, in fact, often implies implicitly the other question *what*, whilst the reverse does not often occur. The process of measurement enables the feature measured to be represented and quantified by numbers, and it states the empirical relationships of interest in algebraic relationships among the numerical values assigned [4].

If the concepts to be evaluated are not directly measurable, it is necessary to use *indicators*. Indicators must be simple and are specific tools which can be translated into terms tied to general concepts by a linkage of semantic representation [13].

Given the copious output of statistical information in social research, there is some enthusiasm for the construction and production of social indicators. As a result, a number of questions arise concerning the sensibleness of the choices taken and the methods used when constructing indicators.

One of the main issues is what indicators should be used, and for what purpose [7]. Indicators constitute the linkage between observations and the complex concept to be measured. On the assumption that the aim of the research determines the indicators, which assume the meaning of meta data: they help shed light on the concept to measure, and they perform the dual task of specifying and measuring the concept.

Indicators are items of information which synthesise the characteristics of a concept or highlight what is occurring within it. They often result from a compromise reached at reasonable cost between scientific accuracy and availability of information. A "composite" indicator[3] is not just the result of a thorough process of evaluation: it may also be the starting-point for political discussion of the phenomenon under study.

It is difficult to identify the best way to measure a complex, multidimensional, and abstract concept, both from the point of view of the sense of the measurement and the field of application. But what is the use of comparing specific components if the aim is to compare systems, and not a set of specific components of the system? If the objective is to receive warning signals, attention should focus on measuring the components, keeping the information disaggregated into "simple" indicators. When the purpose of the analysis is to compare systems or situations, synthesis with a "composite" indicator is necessary [3].

Indicators are classified according to various criteria. An important distinction is drawn between simple and composite indicators. *Simple* (or elementary) indicators refer to a simple unidimensional concept, or to one of the immediately quantifiable dimensions of a complex multidimensional concept. The aggregation and possible weighting of several simple indicators give rise to what is called a *composite* indicator. From the computational point of view, there may be three "key steps" which lead to the determination of a composite indicator: *normalization*, *aggregation*, and *weighting* of the simple indicators.

According to Land [18], an indicator is meaningful only when it possesses informational value within a theoretical model, however it may be defined – mathematically, operationally, logically, orally, etc. – for the analysis and interpretation of social phenomena. In recent decades, the history of indicators seems to have a further principle to this definition: an indicator is usually the outcome of the decomposition of a complex concept into its elementary components. It is a process of reassembling through procedures which normalize, aggregate and weight the simple indicators. This process obviously come from qualitative and quantitative information, subjective and objective observations, descriptions, analysis, and interpretation of existing sources or ad hoc surveys. The indicator therefore exists within a model and it is also produced by the model itself. According to this principle, the indicator often increases the content and meaning of the complex concept being examined within the model.

An indicator is a tool to convert the measurement of complex concepts into a systematic array of interpretative conjectures and relations incorporated into a functional model. But some distance persists between the heuristic intent and the operational feasibility. There exists, in other words, a gap between the (convinced, essential, sometimes normative/legislative) intention to assess a complex concept and its realistic measurement (broadly determined by the system of operational conditions actually adopted or adoptable, even when accompanied by careful and explicit reflection on the methodological rigour of the entire process).

[3] A definition of simple and composite indicators follows.

3.4 The Censis Data

Censis gave us the raw data used to compile the 2006 rankings, and made available the data for the 27 Faculties of Political Science of the Italian public universities.

The techniques of normalization and aggregation adopted by Censis, the list of simple indicators and the preliminary analyses are explained in the following paragraphs.

3.4.1 Normalization and Aggregation

The normalization technique used by Censis is a max-min standardization which converts the values into indicators, dividing by the range:

$$I = \frac{X - \min(X)}{\max(X) - \min(X)} \times 1{,}000 \tag{1}$$

where X is the value of the raw indicator, whilst $\min(X)$ and $\max(X)$ are respectively the minimum and maximum value that the indicator assumes in the set of homogeneous faculties considered.[4] The transformed values will therefore vary from a minimum of 0 to a maximum of 1,000, and they will be comparable within each cluster of faculties: in fact, it is not possible to compare different faculties by means of the same indicator.

Censis rescales the values of the indicators in the interval 66–110, which represents the range of grades awarded for degrees in Italy. Because the formula for this transformation was not reported in the methodological notes, and since the results did not change because it is a linear transformation, this rescaling was not necessary and was not performed in our calculations.

The average final grade M attributed to each faculty was calculated as the arithmetic mean of the normalized scores of the five areas considered:

$$M_f = \frac{std(P_f) + std(D_f) + std(R_f) + std(PD_f) + std(RI_f)}{5} \tag{2}$$

where f denotes each faculty (from 1 to 27) and P is the score for the productivity area, D the score for educational delivery, R for research, PD for the academic staff profile, and RI for international relations.

[4] An interesting alternative would be the use of the theoretical maximum and minimum with the simple indicators for which such values are determinable: this technique would make it possible to reduce the distances among units observed in terms of residuals among normalized values of the simple indicators.

3.4.2 The Simple Indicators Used by Censis

The simple indicators are reported in Tables 3.1, 3.2, 3.3, 3.4 and 3.5. To highlight that simple indicators are calculated for each faculty, we use the subscript f (from 1 to 27). Each simple indicator is normalized according to the formula 1. This transformation is indicated in Tables 3.1, 3.2, 3.3, 3.4 and 3.5 as $std(\cdot)$.

The thresholds for the k values were stated by Censis.

3.4.3 Preliminary Analysis

Since exploratory analysis of the data and study of relations among the variables is an important phase, we started with this operational step in order to verify the existence of relations among the simple indicators considered.

Table 3.1 Simple indicators for "productivity" area (Censis, 2006)

	Productivity
P_{1f}	Rate of persistence between 1st and 2nd year: *(students enrolled in the 2004–2005 academic year who were freshmen in the previous year)/(freshmen in 2003–2004)*
P_{2f}	Regularity index of students: $60 \times$ *(credits acquired in 2004 by students enrolled on the first level 3-year degree or on the "single-cycle" 5-year degree courses)/(students enrolled on the first level 3-year degree or on the "single-cycle" 5-year degree courses in the 2003–2004 academic year)*
P_{3f}	Rate of students enrolled "in corso"[5]: *(total students enrolled – freshmen – students enrolled "fuori corso")/(total students enrolled – freshmen)*
P_{4f}	Rate of 3-year graduates: *(graduates in 2004 from 3-year degree courses who were enrolled in the 2001–2002 academic year)/(freshmen on 3-year degree courses in the 2001–2002 academic year)*
P_{5f}	Rate of graduates "in corso": *(graduates "in corso" in 2004 from 3-year single-cycle degree courses and from previous 4-year degree programmes)/(total graduates from the courses stated)*

Aggregation formula[6]

$$P_f = \frac{std(P_{1f}) + std(P_{2f}) + std(P_{3f}) + std\left(\frac{std(P_{4f})n_1 + std(P_{5f})n_2}{n_1 + n_2}\right)}{4}$$

where
$k = 1$ if $D_{8f} < 75$
$k = 1.05$ if $D_{8f} \geq 75$

[5] Students "in corso" have fulfilled their examination requirements within the scheduled deadlines, whilst students "fuori corso" are still attending university beyond the duration of their courses because they have not yet completed their examination requirements.

[6] Because $P5$ was furnished by Censis as a rate, it was not possible to derive the value of n_2. The score for P was obtained as the simple arithmetic mean of the simple indicators.

Table 3.2 Simple indicators for "educational delivery" area (Censis, 2006)

	Educational delivery
D_{1f}	*Number of degree courses on the faculty programme in 2004–2005*
D_{2f}	*Number of subjects-courses on the faculty programme in 2003–2004*
D_{3f}	*(Tenured academic staff)/(number of subjects-courses in 2004 and 2004–2005)*
D_{4f}	*(Tenured academic staff on 31.12.2004)/students enrolled in 2004–2005)*
D_{5f}	*(Lecture room places NUCLEI 2004)/(students enrolled in 2002–2003)*
D_{6f}	*(Lecture room places NUCLEI 2005)/(students enrolled in 2003–2004)*
D_{7f}	*Student work experience placements (stage) in 2003–2004*
D_{8f}	*Monitoring and evaluation of courses in 2003–2004*

Aggregation formula[7]:

$$D_f = \frac{std\left(\frac{std(D_{1f})+std(D_{2f})+std(D_{3f})}{3}\right) + std(D_{4f}) + std\left(\frac{std(D_{5f})+std(P_{6f})}{2}\right) + 0.5std(D_{7f})}{4} \times k$$

where:
$k = 1$ if $D_{8f} < 75$
$k = 1.05$ if $D_{8f} \geq 75$

Table 3.3 Simple indicators for "research" area (Censis, 2006)

	Research
R_{1f}	*(Number of research units funded by the COFIN and FIRB programmes in 2003)/(tenured staff on 31.12.2002)*
R_{2f}	*(Number of research units funded by the COFIN and FIRB programmes in 2004)/(tenured staff on 31.12.2003)*
R_{3f}	*(Number of research units funded by the COFIN and FIRB programmes in 2005)/(tenured staff on 31.12.2004)*
R_{4f}	Average COFIN and FIRB funding:*(total funding obtained by research units from the COFIN and FIRB programmes in 2003)/(number of units funded)*
R_{5f}	Average COFIN funding:*(total funding obtained by research units from the COFIN programme in 2004)/(number of units funded)*
R_{6f}	Average COFIN and FIRB funding:*(total funding obtained by research units from the COFIN and FIRB programmes in 2005)/(number of units funded)*
R_{7f}	*Number of research projects funded by the EC V and VI Framework Programme and Tempus Programme*

Aggregation formula:

$$D_f = \frac{std\left(\frac{std(R_{1f})+std(R_{2f})+std(R_{3f})}{3}\right) + std\left(\frac{std(R_{4f})+std(R_{5f})+std(R_{6f})}{3}\right)}{2} \times k$$

where:
$k = 1$ if $R_{7f} = 0$
$k = 1.05$ if $R_{7f} > 0$

[7] In the case of the educational delivery, the weighting used by Censis raises obvious questions concerning the weights (it is not stated whether specific choices were made) because the denominator of the formula should be 3.5 instead of 4. For the sake of consistency, we decided to keep the formula applied by Censis.

Table 3.4 Simple indicators for "academic staff profile" area (Censis, 2006)

	Academic staff profile
PD_{1f}	*Average age of tenured academic staff in 2005*
PD_{2f}	Ageing: *(average age of tenured academic staff in 2005) – (average age of tenured academic staff in 2001)*
PD_{3f}	Outgoing Erasmus students per member of academic staff: *(students with Erasmus grants in 2004–2005)/(tenured academic staff on 31.12.2004)*
PD_{4f}	*(Courses taught by untenured "extra-academic" lecturers)/(total courses taught in 2003–2004)*
PD_{5f}	"Rientro dei cervelli" programme: *number of lecturers participating in the international mobility programme for Italian and foreign scholars in thet three-year period 2004–2006*

Aggregation formula[8]:

$$PD_f = \frac{std\left(\frac{std(PD_{1f})+std(PD_{2f})}{2}\right) + 0,5std(PD_{3f}) + 0,5std(PD_{4f})}{3} \times k$$

where:
$k = 1$ if $PD_{5f} = 0$
$k = 1.05$ if $PD_{5f} > 0$

Table 3.5 Simple indicators for "international relations" area (Censis, 2006)

	International relations
RI_{1f}	Outgoing Erasmus grant-holders per student: *(outgoing students with Erasmus grants in 2004–2005)/(students enrolled net of matriculants in 2004–2005)*
RI_{2f}	Incoming Erasmus grant-holders per student: *(average number of foreign students who obtained an Erasmus grant at the faculty in 2003–2004 and 2004–2005)/(students enrolled in 2004–2005)*
RI_{3f}	Host universities per lecturer: *(number of foreign universities which hosted Erasmus students in 2004–2005)/(tenured lecturers on 31.12.2004)*
RI_{4f}	International opportunities: *(number of contributions obtained by the faculty for international cooperation schemes in 2003–2006: lecturer exchanges financed by Miur in 2004; Programma Vigoni 2003–2004; Programma Italia-Germania 2003–2004; Azioni Italia-Spagna 2004–2005; Programma Italia-Germania 2004–2005; Programma Galileo Italia-Francia 2004–2005; Cooperazione Internazionale finanziata dal Ministero degli Esteri – Accordi Bilaterali 2002–2006)*

Aggregation formula:

$$RI_f = \frac{std(RI_{1f}) + std(RI_{2f}) + std(RI_{3f})}{3} \times k$$

where:
$k = 1$ if $RI_{4f} = 0$
$k = 1.05$ if $RI_{4f} > 0$

[8] Likewise the case of the educational delivery, in the academic staff profile the denominator should be 2 and not 3. For the sake of consistency, we decided to keep the formula applied by Censis.

The independence between pairs of variables was measured by means of the Bravais-Pearson coefficient of correlation. The statistical significance was tested with a null hypothesis equal to zero.

We first examined the correlation among the simple indicators belonging to the same area: evidence of correlations among the various indicators would indicate that some aspects had been measured – and therefore considered – several times within the same area. This would not have complied with the parsimony criterion which should guide the construction of composite indicators.

Of course, the results were affected by the small number of faculties available: 27 units, in fact, did not represent a number of observations sufficient to produce stable and convincing results. Moreover results were not extendable to the universe of the Italian faculties. Table 3.6 lists the correlations higher than ±0.4 within the areas (we consider only the values of the correlations because the analysis refers to all the faculties).

We analyzed the correlations among all the indicators. The resulting matrixes showed correlations among indicators belonging to different areas, which suggested the existence of a hypothetical – and not unrealistic – effect of the same measures on different dimensions by means of indicators belonging to different areas.

Table 3.6 Pairwise correlations among simple indicators (values higher than ±0.4)

Area	Value	Indicators
Productivity	0.648	P2 vs. P4
Educational delivery	0.701	D_1 vs. D_2
	0.617	D_6 vs. D_7
	0.587	D_5 vs. D_6
	0.546	D_5 vs. D_7
	–0.435[9]	D_2 vs. D_3
Research	0.773	R_1 vs. R_3
	0.527	R_1 vs. R_6
	0.491	R_4 vs. R_6
Academic staff profile	0.519	PD_1 vs. PD_2
	0.486	PD_3 vs. PD_5
	–0.445[10]	PD_1 vs. PD_4
International relations	0.894	RI_1 vs. RI_2
	0.645	RI_1 vs. RI_3
	0.528	RI_2 vs. RI_3

[9] The correlation between D_2 and D_3 is negative because in the Italian university system who offer a higher number of courses usually has a minor tenured academic staff.

[10] The correlation between PD_1 and PD_4 is negative because a higher average age of tenured academic staff implies that the same staff taught the majority of the courses (untenured "extra-academic" staff usually taught a minor number of courses).

Table 3.7 Matrix of correlations among areas (data obtained using the Censis procedure)

	P	D	R	PD	RI
P	1	0.454	0.423	0.443	0.297
D		1	0.419	0.093	0.111
R			1	0.145	0.087
PD				1	0.200
RI					1

We finally calculated the matrix of correlations among the overall scores of the areas using the scores of the areas constructed by means of the Censis methodology (in some cases obtaining values slightly different from those published) as shown in Table 3.7.

High correlations among simple indicators belonging to the same area were highlighted: this may indicate that two indicators cover areas that overlap each other. This affects the validity property of the measurement process. The Censis aggregation method was used to synthesise the simple indicators of the same area (without changing the weights).

The aim of this correlation analysis was to point out the redundancy among the indicators used by Censis and to notice that this redundancy did not exist among areas. This analysis would be done by Censis considering the complete dataset relative to all the Italian faculty: in this way it could have stable results.

In order to complete a preliminary analysis of the data, we wanted to devote a specific section to multivariate analysis [20] intended to evaluate the number of latent statistical dimensions derivable from the simple indicators. Given the small size of the dataset available, it was not possible to obtain information useful to help us in constructing a different configuration of the areas.

3.5 Alternative Ways to Analyse the Data

Before adopting our strategy of analysis, we considered a list of techniques of normalization, aggregation and weighting [2, 5, 14, 16, 20–22].

- Normalization comprises all the operations performed to transform the simple indicators so that they are comparable with each other in terms of direction, unit of measurement, and order of magnitude. It can be performed by means of:
  - *linear transformations* ($Y = \alpha + \beta X$ where the response variable Y is a linear function of the explanatory variable X [1]) as dividing by the range, as transformation into index numbers, standardization, comparison with the unit leader or a control group, distance from the median;
  - *non-linear transformations* (where the relationship f between Y and X, $Y = f(X)$, is nonlinear); the most used non linear function essentially to convert the data into ordinal values (ranks).

- Aggregation is the choice of merging through an appropriate function which combines different dimensions of the concept under study. It can be performed by means of:
  - *ordinal approach*, which synthesises the indicators transformed into ranks with their mean or sum;
  - *additive cardinal approaches*, which involve calculation of the mean of the transformed values;
  - *non-compensatory multi-criteria approach*, which solves the compensation problem via comparisons among couples of units;
  - *geometric aggregation*, an intermediate solution in terms of compensation between additive aggregations and the multi-criteria approach;
  - *multivariate aggregation techniques*, based on principal components analysis or factor analysis, which draw the latent dimensions that the data describe.
- Weighting is the phase of the process when weights are assigned to the indicators and/or to the dimensions of the concept. The weights may be:
  - *equal* for every variable: this is not a "non-choice" but it grants equal status to all the indicators;
  - based on multivariate models (the most common are *regression* and *factor analysis*);
  - derived from the application of *participatory methods*;
  - calculated by applying the *hierarchical analytical process* which breaks a problem down into a hierarchy and systematically collects opinions on the indicators through pairwise comparisons;
  - derived by the distance from a defined *efficiency frontier*;
  - estimated using an *unobserved components model*.

Among all the normalization, aggregation and weighting techniques listed above, we decided to use those that the literature indicates as the most robust and convincing. Some methods of analysis were discarded due to the small amount of data available. We wanted to adopt techniques which were mutually compatible but based on different approaches and selected two normalization methods: linear and non-linear. We consequently decided to use two different aggregation methods applicable to any normalization. Finally we also adopted two systems of weights: equal for every area (as in the Censis procedure), and the other one based on the participatory method.

It is worth to mention that we first applied the Censis aggregation and weighting techniques to our data, in order to obtain the same results published in the *Guida*. The starting point for our procedure was the set of simple indicators that we had constructed from the variables furnished by Censis.

- The simple indicators were therefore normalized by means of three different techniques:
  1. dividing by the range (as in the Censis procedure);
  2. standardization with z scores;
  3. rank transformation.

- Indicators were aggregated in two distinct steps: first the simple indicators of the same area were aggregated; then the scores of the five areas were aggregated to produce the final league table. We performed only the second aggregation, keeping unchanged the one made by Censis to calculate the final values of each area. The two methods selected for the aggregation were the following:

  1. arithmetic mean;
  2. geometric aggregation.

  Rather than the non-compensatory multi-criteria approach, we opted for geometric aggregation for several reasons: because it is a simply-to-use technique; it is easy to understood; it is better suited to a small dataset; it is a good compromise in terms of compensation between the multi-criteria approach (which excludes compensation) and linear approaches (which do not concern compensation). Moreover the geometric aggregation enabled us to compare our results with those published keeping our assumption close to those adopted by Censis.

- Two methods were selected for the weighting:

  1. equal weights for each area;
  2. participatory method with the "allocation of a budget" by experts.

  The weighting based on the expert judgments was done by us: we "arrogated" this role to ourselves by assigning a weight equal to 0.25 to *educational delivery*, *research* and *academic staff profile* areas, and a weight equal to 0.125 to *productivity* and *international relations* areas. A lower weight was assigned to *productivity* because it was too closely tied to the composition of the student component, and because of the ambivalence of the indicator's information content (good rates of graduates and students "in corso" do not necessary mean a good performance in terms of *productivity*). A lower weight was given also to *international relations* because these substantially only concerned the Erasmus Programme, whilst other activities were omitted. It would be interesting to use the participatory approach with experts on the university system to obtain a shared system of weights. This could also be done by Censis using the results of the surveys conducted with the faculty deans.

Hence 12 ranking tables were obtained by applying the three different normalization methods, the two aggregation techniques, and the two systems of weights. They are summarized in Table 3.8.

In the following analysis we did not considered two methods out of 12. There were marked differences for the C2 and D2 methods due to the computational problems in the geometric aggregation of the standardized z scores.

Censis prefers simple mathematical processes instead of complex statistical models because the readers of the *Guida* are future freshmen and their families which could not appreciate complex statistical methods. For this reason we decided to work in the same perspective.

Table 3.8 Combination of the normalization, aggregation, and weighting techniques in constructing the 12 ranking tables

	Range	Z scores	Ranks
Mean	A1 equal weights B1 weights by experts	A2 equal weights B2 weights by experts	A3 equal weights B3 weights by experts
Geometric Aggregation	C1 equal weights D1 weights by experts	C2 equal weights D2 weights by experts	C3 equal weights D3 weights by experts

3.6 Results

For each method we obtained a list of 27 values and a position for each faculty in a ordered list (ranking). We compared and contrasted the 10 ranking tables of the combination of the normalization, aggregation, and weighting techniques and the league table published by Censis. Finally, we synthesized them into a combined ranking table (the best estimation of the "true" league table of the faculties).

The results of the 10 rankings are reported in Table 3.9, where the cells show the position of each faculty according to each method. The last column of the table reports the position of the faculties in the league table published by Censis.

Table 3.9 Rankings of the faculties (10 ranking methods and the league table of Censis)

Faculty	A1	A2	A3	B1	B2	B3	C1	C3	D1	D3	Censis
Bari	26	26	27	27	27	27	21	27	21	27	26
Bologna	2	3	1	2	2	3	1	3	1	4	1
Cagliari	15	16	14	14	16	13	14	15	13	13	12
Calabria – Cosenza	7	6	11	3	4	9	8	8	6	5	8
Catania	25	24	25	25	24	25	20	25	20	25	25
Firenze	3	2	4	4	3	5	3	2	3	3	5
Genova	10	13	16	15	15	18	12	14	16	16	9
Macerata	23	23	23	24	25	23	23	23	23	23	21
Messina	21	21	21	23	22	22	22	22	22	22	23
Milano 1	8	8	3	7	7	1	6	4	4	2	7
Napoli Orientale	17	19	16	19	19	17	23	17	23	18	17
Napoli 1 – Federico II	27	27	26	26	26	26	23	26	23	26	27
Padova	12	12	11	13	14	14	11	12	12	15	15
Palermo	24	25	20	22	23	20	23	21	23	21	24
Pavia	5	5	4	6	5	5	4	6	5	6	2
Perugia	9	9	7	11	12	7	7	7	10	11	10
Piemonte Orientale	1	1	4	1	1	2	17	1	9	1	6
Pisa	22	20	21	20	20	21	23	20	23	20	20
Roma 1	20	22	24	21	21	24	19	24	19	24	22
Roma 3	6	7	9	8	8	9	5	11	7	12	4
Salerno	18	17	19	17	18	19	16	19	17	19	18
Sassari	16	15	15	16	13	14	15	16	15	14	16
Siena	11	10	8	10	10	8	9	10	11	8	11
Teramo	13	14	13	9	9	12	10	13	8	10	13
Torino	14	11	9	12	11	9	13	9	14	9	14
Trieste	4	4	1	5	6	4	2	5	2	7	3
Urbino – Pesaro	19	18	18	18	17	16	18	18	18	17	19

In order to compare and contrast the positions of the faculties in the ten ranking tables with respect to the position in the league table of Censis we reported the frequencies of the absolute differences in Table 3.10. We noted a general concordance of the results, with the exception of some faculties for which the distances from the Censis values seemed rather wide: Genova, Piemonte Orientale, Roma 3, and to a lesser extent, Padova and Torino.

We found a good concordance between our results and the league table published by Censis. A brief inspection of the tables immediately showed that the positions of the faculties were not particularly variable among ranking methods and with respect to the ranking table published by Censis. The highest cograduation was between A1 and Censis. The same normalization, aggregation and weighting techniques were used to construct the two ranking tables; the only difference consisted in the initial set of simple indicators, for which, however, there was no evidence of a close correspondence between our indicators and those elaborated by Censis. Results showed a generally high consistency between the league table of Censis and our methods that use data normalized in ranks (in order, A3, C3, B3 and D3). The attribution of different weights to the areas became important (D-type methods): despite the presence of few data, the weighting had an important role in defining the positions in the ranking tables.

The 10 ranking tables obtained with different normalization, aggregation and weighting methods showed a high level of concordance. To obtain a measure of this concordance we used Kendall's coefficient W([17]: 95) and its chance-corrected version W_1[6, 15]. For our ten ranking tables we obtained $W = 0.91$ and $W_1 = 0.92$. Since the coefficients are close to 1 we could estimate a combined "best" ranking table of the faculties. According to Kendall [17] the best ranking table could be obtained by means of the sum of the ranks, i.e. the position of a faculty is determined by the sum of its positions in the ten ranking tables. Although this approach would increase the computational complexity, it would ensure greater robustness and reliability of the final results [6, 15].

The result of this combined ranking table is reported in Table 3.11. The second and third columns report the position of the faculty in the league table published by Censis and the absolute difference between the positions. We noted again a general concordance of the results, with the exception of some faculties for which the distances between the position in the combined ranking table and the position proposed by Censis were wider: Genova, Pavia, Roma 3, and to a lesser extent, Piemonte Orientale, Firenze and Macerata.

3.7 Conclusions

The ranking of university institutions always causes controversies, expectations, and criticisms in the actors (areas, actual and potential university students, and academic "actors"). In this study we have addressed the core of the problem by focusing on the ranking method proposed by Censis in its *Grande Guida all'Università* and analysing its structure, our purpose being to understand what measurement

Table 3.10 Distribution of absolute differences (10 ranking methods vs. the league table of Censis)

Differences	A1	%	A2	%	A3	%	B1	%	B2	%	B3	%	C1	%	C3	%	D1	%	D3	%
0	11	40.7	4	14.8	3	11.1	5	18.5	3	11.1	3	11.1	2	7.4	6	22.2	4	14.8	3	11.1
1	6	22.2	10	37.0	8	29.6	9	33.3	9	33.3	10	37.0	8	29.6	6	22.2	7	25.9	7	25.9
2	6	22.2	5	18.5	7	25.9	6	22.2	4	14.8	4	14.8	6	22.2	4	14.8	3	11.1	5	18.5
3	3	11.1	5	18.5	3	11.1	1	3.7	4	14.8	4	14.8	5	18.5	6	22.2	7	25.9	5	18.5
4	0	0.0	2	7.4	3	11.1	3	11.1	5	18.5	2	7.4	2	7.4	1	3.7	1	3.7	2	7.4
5	1	3.7	1	3.7	2	7.4	2	7.4	1	3.7	2	7.4	2	7.4	3	11.1	3	11.1	3	11.1
6+	0	0.0	0	0.0	1	3.7	1	3.7	1	3.7	2	7.4	2	7.4	1	3.7	2	7.4	2	7.4
Tot.	27	100	27	100	27	100	27	100	27	100	27	100	27	100	27	100	27	100	27	100

Table 3.11 Rankings of the faculties in the combined ranking table

Faculty	Comb. Class.	Censis	diff.
Bari	26	26	0
Bologna	1	1	0
Cagliari	14	12	2
Calabria – Cosenza	7	8	1
Catania	25	25	0
Firenze	2	5	3
Genova	15	9	6
Macerata	24	21	3
Messina	21	23	2
Milano 1	5	7	2
Napoli 1 – Federico II	27	27	0
Napoli Orientale	19	17	2
Padova	13	15	2
Palermo	23	24	1
Pavia	6	2	4
Perugia	9	10	1
Piemonte Orientale	3	6	3
Pisa	20	20	0
Roma 1	22	22	0
Roma 3	8	4	4
Salerno	18	18	0
Sassari	16	16	0
Siena	10	11	1
Teramo	11	13	2
Torino	12	14	2
Trieste	4	3	1
Urbino – Pesaro	17	19	2

instruments can be used, how to combine them, and how to obtain robust final results. Our intention has not been to criticise the Censis ranking system a priori, but rather to analyse how it can be adjusted and/or improved, as well as to suggest possible alternatives to it. However, we wish to make a proposal: we regard it as both necessary and desirable for Censis to clearly state how it has selected and/or determined the "areas" used to evaluate the university system when its university league table is published.

We have considered indicators as basic tools to operate, and we have argued that the synthesis of indicators is crucial for evaluation processes. When discussing the complex process of constructing a composite indicator, we highlighted the normalization, aggregation and weighting phases, and we illustrated a set of techniques based on different theories and suited to different purposes. With a view to comparison among several situations, as well to give warning signals on individual aspects, analysis must synthesise the information. This, therefore, is what we have sought to do: apply different operational techniques to the data in order to obtain results that enable comparison among university faculties on the basis of a synthesis of a wide range of alternative applications. Geometric aggregation becomes preferable to the simple linear aggregation which calculates the average of the items; weighting

assumes significant importance in synthesis of the information; simple normalization techniques (e.g. the transformation of the simple indicators into ranks) and more complex standardizations conduct to similar results. In general, we would suggest the use of normalization, aggregation and weighting techniques that are not overly complex with respect to the assumptions and objectives of the analysis. This will foster better understanding of the methodology employed by the *Guida* for the readers, and especially its target audience of future university students and their families. Moreover, standardization by the range proved not to be a good normalization technique, because large distances between maximum and minimum values were amplified. The use of ranks was a good alternative method of data normalization instead of the method based on the range. Geometric aggregation (except in the case based on z scores) was a good aggregation technique based on a logic of non-complete compensability among the indicators (for the role assumed by weights in different aggregation methods see [19]).

The data used in our calculations have been collected and furnished by Censis. It was, therefore, essential to regard them as "quality" data and attribute to them – a posteriori for obvious reasons – the properties of accuracy, validity and consistency. This, however, prompts a necessary consideration: it is essential to verify the quality of data also by making careful selection of the information deemed useful and necessary, without giving in to the temptation to "cherry-pick" information from the sources available.

The small amount of data available for our calculations has restricted the range of possible applications. In particular, it has precluded analysis of the structure underlying the data using multivariate analysis methods. Factor analysis of the entire set of simple indicators might yield areas different – in number and significance – from those (pre-)determined by Censis, considering the complexity and delicacy of establishing them a priori. Ex post cluster analysis might instead be useful for verifying the existence of geographical areas or types of faculties which are problematic or virtuous according to the aspects analyzed. Also preliminary analyses based on correlations, if performed on a larger dataset, could highlight redundancies among the indicators belonging to the same area or overlaps among areas.

Given the overall structure of the league table of Censis, an important observation concerns its lack of measures of variation in position within the ranking table and of year-by-year changes in the scores for the areas for each faculty. Of course, any evaluation in this sense must consider the changes that take place every year in the structure of the indicators and areas, changes which entail that annual rankings are not entirely comparable.

In this study we have asked whether Censis is a reliable "referee". A series of choices made by Censis produce results quite similar to those yielded by the alternative strategies used here. From this point of view, we may say that the technical-methodological aspects of constructing composite indicators seemingly do not give rise to significant differences in the results. There appear to be two main discriminating factors: the nature, articulation and quality of the database used to represent the sub-dimensions of the concept considered; and the strategy used for their weighting within the areas (or "family" as Censis calls it), and among areas. In its *Guida* and

on its website, Censis allows examination and evaluation of the general procedure that it adopts. We might call this an ex-post "search for transparency": we believe that there should be a joint effort by the actors involved, and more generally by the stakeholders. All that in order to establish the objects, the rules of the game could be a way, laborious but necessary, to improve the process and to achieve an outcome which creates less wrangling, less discontent, less indifference, less ill-feeling. A participatory process involving all the stakeholders is less agile and efficient than appointment of an actor external to the system. Nevertheless, the issue is a highly sensitive one, and it warrants higher-level discussion if the results of a ranking method will be more believable and have a real effect on the university system. The literature on evaluation devotes ample space to the issue of the quality of the interaction among actors, especially in complex, dynamic and turbulent contexts. The correct management of relations among actors when a ranking system is adopted is necessary so that there is a co-responsibility (collective assumption of responsibility) for processes, greater recognition of the value of the results achieved and, therefore, also greater future use of the indications obtained.

References

1. Agresti A, Finlay B (1997) Statistical methods for the social sciences. Prentice-Hall Inc, Englewood Cliffs, NJ
2. Aiello F, Attanasio M (2004) How to transform a batch of simple indicators to make up a unique one? Atti del Convegno SIS giugno 2004, Bari – Sessioni Specializzate
3. Aiello F, Attanasio M (2006) Some issues in constructing composite indicators. VIII international meeting on quantitative methods for applied sciences, Certosa di Pontignano, 11–13 settembre 2006
4. Aiello F, Librizzi L (2006) Gli indicatori nelle scienze sociali: dal qualitative al quantitativo. In: Diamond I, Jefferies J (eds) Introduzione alla statistica per le scienze sociali. Mc Graw-Hill, Milano
5. Allegra FS, La Rocca A (2004) Sintetizzare misure elementari: una sperimentazione di alcuni criteri per la defnizione di un indice composto. Istat, Roma
6. Attanasio M, Capursi V (1997) Graduatorie sulla qualità della vita: prime analisi di sensibilità delle tecniche adottate. Atti XXXV Riunione Scientifica SIEDS, Alghero
7. Bernardi L, Capursi V, Librizzi L (2004) Measurement awareness: the use of indicators between expectations and opportunities, SIS: Sezione Specializzata. Atti della XLIII Riunione Scientifica, Bari
8. Censis (2000) Grande guida all'università, La Repubblica, Milano
9. Censis (2001) Grande guida all'università, La Repubblica, Milano
10. Censis (2006) Grande guida all'università, La Repubblica, Milano
11. Censis (2007) Grande guida all'università, La Repubblica, Milano
12. Censis (2008) Grande guida all'università, La Repubblica, Milano
13. Corbetta PG (1999) Metodologia e tecniche della ricerca sociale. Il Mulino, Bologna
14. Delvecchio F (1995) Scale di misura e indicatori sociali. Cacucci Editore, Bari
15. Fagot RF (1994) An ordinal coefficient of relational agreement for multiple judges. Psychometrika 59(2):241–251
16. Freudenberg M (2003) Composite Indicators of Country Performance: A Critical Assessment, OECD Science, Technology and Industry Working Papers 2003/16, OECD, Directorate for Science, Technology and Industry

17. Kendall MG (1962) Rank correlation methods. C. Griffin & Co. Ltd, London
18. Land KC (1971) On the definition of social indicators. Am Sociol 6:322
19. Munda G, Nardo M (2005) Constructing consistent composite indicators: the issue of weights. Institute for the Protection and Security of the Citizen, Luxemburg
20. Nardo M, Saisana M, Saltelli A, Tarantola S (2005a) Handbook on constructing composite indicators: methodology and user guide. OECD statistics working papers, 2005/3. OECD Publishing, Paris
21. Nardo M, Saisana M, Saltelli A, Tarantola S (2005b) Tools for composite indicators building. European Commission-Joint Research Centre, Ispra, Italy
22. Saisana M, Tarantola S (2002) State-of-the-art report on current methodologies and practices for composite indicator development. European Commission-Joint Research Centre, Ispra, Italy

Part II
The Evaluation in the Italian Universities: Student Teaching Evaluation

Chapter 4
Structural Equation Models and Student Evaluation of Teaching: A PLS Path Modeling Study

Simona Balzano and Laura Trinchera

4.1 Introduction

In Italian universities, teaching evaluation is in part based on students judgments concerning aspects related to courses and considered of preeminent interest for university management. A questionnaire is generally used to collect such data. The students judgments are expressed as a score on an ordinal scale.

Even if a synthetic measure of quality is required, there is no single methodological solution for aggregating individual scores. Until now several approaches have been proposed in order to define a synthetic measure of teaching quality by using student evaluations, see among others [1, 5, 18].

A possible solution is to use Structural Equation Models (SEM) [3, 14] that are used for describing and estimating conceptual structures where some *latent variables*, linked by linear relationships, are measured by sets of *manifest variables*. A double level of relationships characterizes each SEM: the first involves relationships among the latent variables (*structural model*), while the other considers the links between each latent variable and its own block of manifest variables (*measurement model*).

Given that both the quality of teaching and student satisfaction cannot be observed directly but can be measured through several real indicators, they can be treated as latent variables.

SEM applications in both evaluation and teaching quality measurement have been widely used [6, 11, 12, 15, 16].

Several techniques can be used to estimate model parameters in SEMs, which can be grouped under two different approaches. The first is the so-called *covariance-based* approach, based on the search for the best parameters in reconstructing the observed covariance matrix of manifest variables. A number of estimation techniques are used to estimate model parameters, including the maximum likelihood

S. Balzano (✉)
Dipartimento di Scienze Economiche, Università degli Studi di Cassino, 03043 Cassino, Italy
e-mail: s.balzano@unicas.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_4,

approach, which has long been the point of reference for SEM estimation. However, especially in quality evaluation studies, some limits impair its application: using maximum likelihood estimation for covariance-based SEM requires that the manifest variables follow a multinormal distribution and may lead to non-unique solutions (i.e. the model is not identifiable). Especially in social research, this distributional hypothesis is very hard to verify. Indeed, since the manifest variables are often judgments expressed on ordinal scales, they cannot properly be considered continuous variables and they are unlikely to meet the multinormal distribution hypothesis. Other estimation techniques that do not require a multinormal assumption can be used to estimate SEM parameters in a covariance-based approach, such as the Unweighted Least Squares. Nevertheless, all these techniques are based on the covariance matrix and do not allow individual behaviour to be directly taken into account.

A different approach is the *component-based one*. Following this approach, model estimation is basically geared to determining the latent variable scores, i.e. values of the latent variable for each individual in the sample. The main aim is to identify a latent variable explaining at the same time both its own block of indicators and the relationships between blocks. Among the component-based techniques, the most widely used method is the PLS Path Modeling algorithm (PLS-PM), also called the PLS approach to SEM [20, 24]. PLS-PM does not rely on a specific distributional hypothesis. Moreover, according to Tenenhaus [19] it provides systematic convergence of the algorithm; it allows data to be managed with a small number of individuals and a large number of variables; it provides a practical interpretation of the latent variable estimates; and it represents a general framework for multi-block analysis.

For these reasons we propose to use PLS-PM for SEM estimation in teaching evaluation, also because, since we are interested in describing students opinions, the explorative approach (typical of *component-based* methods) is much more coherent than the strong confirmatory one (typical of *covariance-based* methods).

We note that this is our contribution since in the literature of SEM application to students evaluation of teaching PLS-PM has never been used before.

4.2 PLS Approach to Structural Equation Models

The PLS approach to Structural Equation Models uses an iterative algorithm to obtain latent variable estimates through a system of multiple and simple regressions. The iterative algorithm works by alternating inner and outer estimates of the latent variables. In more formal terms, given the generic latent variable ($\boldsymbol{\xi}_q$), the outer estimation of the latent variable ($\mathbf{v}_q$) is obtained as a linear combination of its own manifest variables $\mathbf{x}_{pq}$:

$$\mathbf{v}_{q\propto} \pm \sum_{p=1}^{P_q} w_{pq}\mathbf{x}_{pq} \tag{1}$$

where P_q equals the number of manifest variables associated to the q-th latent variable and w_{pq} represents the outer weight, i.e. the weight associated to each manifest variable to obtain the latent variable estimate.

In the second step (inner estimation), each latent variable is computed by considering its relations with the other latent variables. In other words, for a given outer estimate of the latent variables obtained in the previous step, the inner estimate $\mathbf{z}_q$ of each latent $\boldsymbol{\xi}_q$ is obtained as:

$$\mathbf{z}_q \propto \sum_{q'} e_{qq'} \mathbf{v}_{q'} \tag{2}$$

where $\mathbf{v}_{q'}$ is a generic latent variable connected to the q-th latent variable and $e_{qq'}$ is an inner weight, usually obtained as the sign of the correlation between the outer estimates of the q-th latent variable and the q'-th latent variable (*centroid scheme*). The symbol $\propto$ means that each estimate of the latent variable has to be standardized, both in the outer and inner estimates.

The iterative procedure goes on to compute the outer weights (w_{pq}). Each of these weights is then used in the following outer estimate of the latent variable (equation (1)). Two different schemes are available to compute the outer weights according to the nature of the latent variables. If the latent variable is obtained as a reflective construct (*mode A*), i.e. if the observed variables are assumed to be the reflection of a latent concept, then the latent variable is considered a predictor of the manifest variable. Thus, each relation in the block is a simple linear regression model and may be expressed as follows:

$$\mathbf{x}_{pq} = \lambda_{pq} \boldsymbol{\xi}_q + \boldsymbol{\epsilon}_{pq} \tag{3}$$

where λ_{pq} is the generic loading (i.e. the correlation coefficient, if the manifest variables are scaled to unit variance) associated to the p-th manifest variable linked to the q-th latent variable, and $\boldsymbol{\epsilon}_{pq}$ is a residual term.

Indeed, in a reflective block each manifest variable is considered to be the reflection in the real world of an underlying concept, that is the latent variable. As a consequence, the generic outer weight w_{pq} used in the outer estimate of the latent variable is the regression coefficient of the simple linear regression of each manifest variable on the inner estimate of the corresponding latent variable. The inner estimates of the latent variables being standardized, each outer weight (for a *reflective block*) is the covariance between each manifest variable and the corresponding latent variable as follows:

$$w_{pq} = Cov(\mathbf{x}_{pq}, \mathbf{z}_q) \tag{4}$$

In a formative scheme (*Mode B*), instead, each latent variable is formed by its own manifest variables. In other words, the latent variable is a function of its own indicators. In this case, a multiple linear regression model defines the relation between the latent and manifest variables:

$$\boldsymbol{\xi}_q = \mathbf{X}_q \boldsymbol{\omega}_q + \boldsymbol{\delta}_q \tag{5}$$

where $\mathbf{X}_q$ is the matrix of the manifest variable linked to the q-th latent variable, $\boldsymbol{\omega}_q$ is the vector of the weights associated to the q-th latent variable, and $\boldsymbol{\delta}_q$ the residual term.

Hence, in a *formative scheme* the outer weights in the iterative procedure are the regression coefficients of a multiple regression model of the inner estimate of each latent variable on its own manifest variable. For each block, the vector containing the P_q outer weights is:

$$\mathbf{w}_q = \left(\mathbf{X}_q' \mathbf{X}_q\right)^{-1} \mathbf{X}_q' \mathbf{z}_q. \tag{6}$$

After updating the outer weights, they are used to obtain a new outer estimate of the latent variables.

These steps are repeated until convergence between inner and outer estimates is reached. The final estimate of the generic latent variable (i.e. the latent variable score, $\hat{\boldsymbol{\xi}}_q$) are then computed. Then, the structural relations among the endogenous latent variable scores ($\hat{\boldsymbol{\xi}}_j$) and the exogenous one ($\hat{\boldsymbol{\xi}}_m$) are estimated by using standard multiple/simple linear regression models.

For a generic endogenous latent variable $\boldsymbol{\xi}_j$ in the model, the structural model can be written as:

$$\boldsymbol{\xi}_j = \sum_{m=1}^{M} b_{jm} \boldsymbol{\xi}_m + \boldsymbol{\zeta}_j \tag{7}$$

where $\boldsymbol{\xi}_m$ is the generic exogenous latent variable impacting on $\boldsymbol{\xi}_j$, b_{jm} is the OLS regression coefficient (*path-coefficient*) linking the m-th exogenous latent variable to the j-th endogenous latent variable, $\boldsymbol{\zeta}_j$ is a residual term, and M is the total number of exogenous latent variables impacting on the j-th endogenous latent variable.

As already stated, the PLS-PM is considered a *soft modelling* approach since no hard distributional hypotheses have to be made either with regard to the manifest variables or to the latent variable scores.

Unlike other estimation techniques used in the SEM framework, the PLS-PM is more prediction-oriented. Thus the quality of the model has to be evaluated in terms of prediction capability. Since two sub-models comprise each SEM, four different indexes have to be used to assess the prediction capability of the model (one measuring the performance of the measurement model, one considering the structural model and the last measuring the goodness of fit of the whole model):

- the average communality index (measurement model goodness of fit index);
- the redundancy indices and the R^2 values of each structural relation in the model (structural model goodness of fit index);
- the goodness of fit index (GoF, goodness of fit index for the model as a whole).

For each block the measurement model quality is assessed by using the average communality index. This index is computed as the average of the squared correlations between each manifest variable in the q-th block and the q-th latent variable score:

$$Com_q = \frac{1}{P_q} \sum_{p=1}^{P_q} cor^2 \left(\mathbf{x}_{pq}, \hat{\boldsymbol{\xi}}_q \right). \tag{8}$$

The average communality index is a measure of the capability of each latent variable score in explaining the variances in the manifest variables.

The quality of each relation in the structural model is measured by using the R^2 value. Moreover, for each endogenous block the redundancy index may be computed as:

$$Red_j = Com_j \times R^2 \left(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{m:\boldsymbol{\xi}_m \to \boldsymbol{\xi}_j} \right). \tag{9}$$

This index provides information on the part of the variability of the manifest variables linked to the j-th endogenous latent variable explained by the M exogenous latent variables impacting on it.

The global model quality is measured by means of the goodness of fit index (GoF) proposed by Amato et al. [2]. This index was constructed to provide a measure of model quality by considering model performance in both the measurement and structural models. Indeed, the GoF index comprises two parts:

$$GoF = \sqrt{\frac{\sum_{q=1}^{Q} \sum_{p=1}^{P_q} Cor^2 \left(\mathbf{x}_{pq}, \hat{\boldsymbol{\xi}}_q \right)}{P} \times \frac{\sum_{j=1}^{J} R^2 \left(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{m:\boldsymbol{\xi}_m \to \boldsymbol{\xi}_j} \right)}{J}}. \tag{10}$$

The first term refers to the quality of the measurement model, while the second takes into account the performance of the structural model. J is the total number of endogenous latent variables in the model and P is the total number of manifest variables in the model, with $P = \sum_{q=1}^{Q} P_q$.

4.3 Applying PLS-PM to Students Evaluation of Teaching

4.3.1 The Data and Model Specification

We show an example of teaching quality evaluation using a Structural Equation Model estimated by a PLS-PM algorithm. The analyzed data are the judgments expressed by 7,369 students attending courses at the Faculty of Humanities at a

university in southern Italy. Judgments were collected through questionnaires distributed to the students during usual daily teaching activities in the academic year 2004/2005.

Each questionnaire is a statistical unit.

Observations do not cover the totality of enrolled students, nor are they a random sample: they were selected not by a sampling procedure, but they are the students present at one lesson of *all* courses (on different days). This means, for example, that each student could have filled in the questionnaire even more than once.

The structure of the questionnaire is based on a standard set of questions, as stated by the National University Evaluation Committee (CNVSU) [8] to ensure universities to have a common database recording students opinion on teaching (so that comparisons among universities, faculties, courses, etc. may be made).

The CNVSU questionnaire is organized in 5 sections. We believe that each of these sections can be considered a latent variable, such that the 15 questions can be treated like manifest variables for each of them, in an SEM sense (see Table 4.1).

In particular, we consider that the latent variable *Interest and satisfaction* is the only endogenous latent variable in the model. In other words, we suppose that *Interest and satisfaction* can be explained by all other aspects, so that its estimated score can be interpreted as a measure of students' evaluation of teaching effectiveness.

In the measurement model, manifest variables are connected to the corresponding latent variables according to a *reflective scheme*: responses are supposed to be a logical consequence (the "reflection") of the latent factor they are connected with.

A preliminary study [4] showed that in such a model the manifest variables describing the block *Teaching and study activities* are correlated with at least two different dimensions while each block should express one latent concept (see composite reliability analysis in Table 4.3). In order to avoid this inconsistency and based on the analysis of the covariance matrices among the manifest variables and

Table 4.1 The logical structure of the CNVSU questionnaire

Latent variables	Manifest variables
Programme organization	v2. Study load v3. Overall organization (course timetable, exams, etc.)
Course organization	v4. Clarity on exam procedure v5. Adherence to course timetable v6. Lecturer's availability for explanations
Teaching and study activities	v7. Understanding of lecture given student's preliminary knowledge v8. Lecturer's ability to stimulate student's interest v9. Lecturer's clarity v10. Proportion between study load and number of credits v11. Suitability of study materials v12. Usefulness of supplementary lessons (practicals, workshops, seminars, etc.)
Facilities	v13. Lecture hall v14. Rooms and equipment for supplementary lessons
Interest and satisfaction	v15. Interest in course subjects v16. Overall satisfaction

Table 4.2 Measurement model definition

Latent variables	Manifest variables
Course organization	v4. Clarity on exam procedure v5. Adherence to course timetable v6. Lecturer's availability for explanations v10. Proportion between study load and number of credits
Teaching	v8. Lecturer's ability to stimulate student's interest v9. Lecturer's clarity v11. Suitability of study materials v12. Usefulness of supplementary lessons (practicals, workshops, seminars, etc.)
Facilities	v13. Lecture hall v14. Rooms and equipment for supplementary lessons
Interest and satisfaction	v15. Interest in course subjects v16. Overall satisfaction

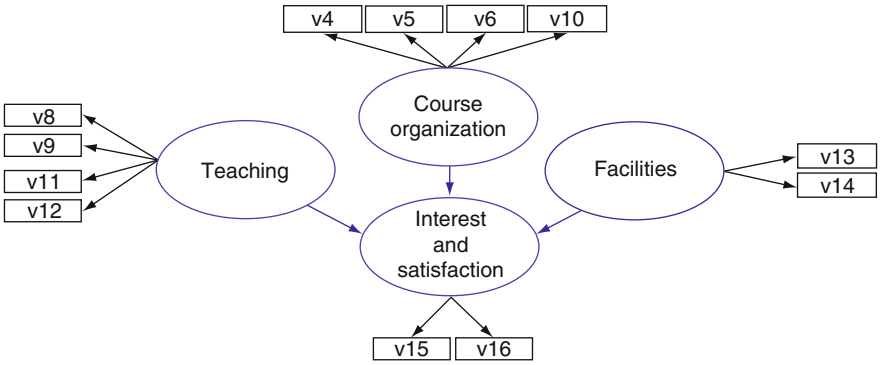


Fig. 4.1 An SEM model for students evaluation of teaching

among manifest and latent variables, we specified a different model, whose structure is shown in Table 4.2.

In the new model, the variable v7 (*Understanding of lecture given student's preliminary knowledge*) and the block *Programme organization* were dropped and variable v10 (*Proportion between studying load and number of credits*) was moved from *Teaching and study activities* to *Course organization* block. Finally, according to the redefinition of the model, the block *Teaching and study activities* has been renamed *Teaching*. The final model is shown in Table 4.2 and in Fig. 4.1.

4.3.2 The Results

XLSTAT software by Addinsoft [25] was used to perform PLS-PM analysis involving only reflective indicators and the centroid scheme for the inner estimation. Since each reflective block represents only one latent construct, it needs to be unidimensional. This is why a preliminary exploratory analysis for verifying the composite

reliability of blocks is required. Two different measures are available to test block unidimensionality in PLS-PM framework: Dillon-Goldstein's rho and Cronbach's alpha. According to Chin [7], Dillon-Goldstein's rho is considered a better indicator than Cronbach's alpha as it is based on the results from the model (i.e. the loadings) rather than on the correlations observed between the manifest variables in the dataset. A block is considered homogeneous if this index is greater than 0.7 [23].

As shown in Table 4.3, all five blocks of manifest variables can be considered unidimensional. Indeed, the Dillon-Goldstein Rho index is always greater than 0.7.

Once the composite reliability is verified, we may look at the relationships between each manifest variable and its own latent variable. Table 4.4 shows the weights of the relationships between each manifest variable and its own latent variable, together with the average communality index, i.e. the ability of each latent variable to explain its own manifest variables. Since this index is always higher than 0.5, we can conclude that globally all the latent variables are powerful at explaining their own manifest variables.

Table 4.3 Composite reliability

Latent variables	Cronbach alpha	D.G. Rho (PCA)	Critical value	Eigenvalues
Course organization	0.677	0.809	0.621	1.323 0.537 0.344 0.279
Teaching	0.729	0.833	0.724	1.623 0.608 0.407 0.258
Facilities	0.297	0.744	0.820	1.055 0.585
Interest and satisfaction	0.668	0.858	0.627	0.942 0.312

Table 4.4 Normalized outer weights and average communalities

Latent variables	MV	Normalized outer weights	Average communality
Course organization	V4	0.324	
	V5	0.214	
	V6	0.273	0.501
	V10	0.189	
Teaching	V8	0.313	
	V9	0.308	
	V11	0.205	0.556
	V12	0.174	
Facilities	V13	0.652	
	V14	0.348	0.585
Interest and satisfaction	V15	0.386	
	V16	0.614	0.741

The normalized weight measure the impact of the corresponding manifest variable in computing the latent variable score as an index. It is evident, for example, that the manifest variable v13 (*Lecture hall*) is the most important driver in computing the latent variable *Facilities*. The same occurs for manifest variable v16 (*Overall satisfaction*) with respect to the latent variable *Interest and satisfaction*, and for latent variable *Teaching* with the two manifest variables directly tied to lecturer's quality and ability (v8 and v9).

As the distribution of PLS estimates is unknown, conventional significance testing is impossible. However, testing may be accomplished by Bootstrap methods [9]. The results of the bootstrap estimation of the standardized loadings of manifest variables are shown in Table 4.5.

Tables 4.6 and 4.7 and Fig. 4.2 show the results of the structural model estimates. Table 4.6 shows the correlation and regression coefficients linking each exogenous latent variable to the endogenous *Interest and satisfaction*. We can conclude that all path coefficient estimates of the structural model are significant.

According to the results in Table 4.6 the structural equation may also be written as follows:

$$\textit{Interest and satisfaction} = 0.605 \times \textit{Teaching} + 0.105 \times \textit{Facilities} + 0.077 \times \textit{Course organization}$$

Table 4.5 Measurement model estimates: loadings

Latent variables	MV	Standardized loadings	Standardized loadings (Bootstrap)	Lower bound (95%)	Upper bound (95%)
Course organization	V4	0.801	0.800	0.783	0.817
	V5	0.693	0.694	0.668	0.718
	V6	0.753	0.753	0.732	0.771
	V10	0.561	0.561	0.533	0.588
Teaching	V8	0.851	0.851	0.838	0.863
	V9	0.859	0.860	0.849	0.870
	V11	0.673	0.673	0.649	0.694
	V12	0.557	0.560	0.530	0.587
Facilities	V13	0.861	0.859	0.827	0.886
	V14	0.654	0.657	0.605	0.700
Interest and satisfaction	V15	0.790	0.791	0.768	0.811
	V16	0.927	0.927	0.921	0.934

Table 4.6 Impact and contribution of exogenous latent variables on the endogenous *Interest and satisfaction*

	Teaching	Course organization	Facilities
Correlation	0.699	0.466	0.429
Path coefficient	0.605	0.077	0.105
p-value	0.000	0.000	0.000
Contribution to R^2 (%)	83.538	6.885	8.774

Table 4.7 Goodness of fit index for the structural model

R^2	R^2 (Bootstrap)	Standard deviation	Lower bound (95%)	Upper bound (95%)
0.504	0.504	0.010	0.482	0.523

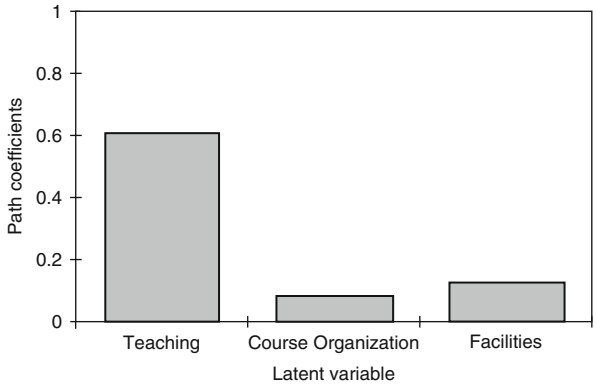


Fig. 4.2 Impact of exogenous latent variables on *Interest and satisfaction*

Looking at the path coefficients (see Fig. 4.2 and Table 4.6), we note that students interest and satisfaction mainly depend on teaching quality (path coefficient = 0.642 and contribution to R^2 higher than 80%) while the quality of facilities and course organization have lower effects (path coefficients: 0.108 and 0.091). This is probably due to how data were collected. Since the questionnaires were distributed during the course, students attached more importance to characteristics intrinsic to that course than to general matters: aspects related to lectures prevailed very largely.

The goodness of fit indices for both the structural and measurement models are very satisfactory with an absolute GoF value of 0.537 and an equal contribution of measurement and structural models in constructing it (see Tables 4.7 and 4.8).

Finally, in Table 4.9 some descriptive statistics for latent variables scores (computed on a 0–100 scale) are shown. Recalling that the individual score of latent variables can be interpreted as the quality level perceived by a student, we can conclude that for both the latent variables *Teaching* and *Course organization* the students are fairly satisfied. Instead, the latent variable *Facilities* does not reach a very satisfactory level.

Table 4.8 Goodness of fit index for the whole model

	GoF	*GoF* (Bootstrap)	Standard deviation	Lower bound (95%)	Upper bound (95%)
Absolute	0.537	0.538	0.006	0.525	0.551
Relative	0.962	0.961	0.003	0.954	0.967
Outer model	0.993	0.993	0.001	0.991	0.994
Inner model	0.968	0.967	0.003	0.960	0.973

Table 4.9 Goodness of fit index for the structural model

Latent variable	Mean	Standard deviation	1st Quartile	Median	3rd Quartile	Variation coefficient
Course organization	48.970	12.665	40.914	50.912	57.924	0.259
Teaching	51.838	15.478	43.219	52.186	63.704	0.299
Facilities	25.894	8.440	20.652	25.000	33.152	0.326
Interest and satisfaction	36.852	11.441	32.938	39.289	49.407	0.310

4.4 Concluding Remarks

In this chapter we used an SEM estimated by a PLS-PM algorithm to define and compute an index for measuring student's evaluation of teaching effectiveness in universities. The proposed approach provides individual values of the index: for each student we compute a score that represents the measure of his/her perception of teaching quality. Moreover, a major advantage of using the PLS-PM approach is that it is possible to derive the weighting system for observed indicators by a data-driven procedure, once the structural and measurement models have been specified.

This issue can be set in a *composite indicator* framework [17]. In this perspective, the PLS-PM estimation provides a double-level weighting system [22]. Indeed, the results can be interpreted as follows: *path coefficients* represent the impact of the exogenous latent variables on the composite indicator (*Interest* and *satisfaction*), while the *normalized weights* are the weights for simple indicators (manifest variables). Together they define the coefficients of the final linear combinations (*aggregation functions*) for computing the composite indicator (*latent variable score*) at individual level.

Finally, we note that in order to consider more homogeneous contexts and compare results it would be interesting to perform PLS-PM analysis for separate groups of students according to a priori information (for example by using external variables such as the programme attended) or by running a so-called *response-based* clustering algorithm such as REBUS-PLS [10, 21] or FIMIX-PLS [13]. This latter issue may be an interesting topic for further work.

References

1. Aiello F, Attanasio M (2004) How to transform a batch of simple indicators to make up a unique one. In: Atti della XLIII Riunione Scientifica della SIS, Padova, pp 327–338
2. Amato S, Esposito Vinzi V, Tenenhaus M (2005) A global goodness-of-fit index for PLS structural equation modeling. Technical report HEC School of Management, France
3. Bollen KA (1989) Structural equations with latent variables. Wiley, New York, NY
4. Balzano S, Trinchera L (2008) Structural equation models and student evaluation of teaching: a PLS approach. In: Atti del convegno DIVAGO, Palermo
5. Capursi V, Porcu M (2001) La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni di giudizi. In: Atti Convegno Intermedio della Società Italiana di Statistica "Processi e Metodi Statistici di Valuatzione", Rome

6. Chiandotto B, Bini M, Bertaccini B (2006) Evaluating the quality of the university educational process: an application of the ECSI model. In: Fabbris L (ed) Effectiveness of university education in Italy: employability, competences, human capital. Springer, Heidelberg
7. Chin WW (1998) The partial least squares approach to structural equation modeling. In: Marcoulides GA (ed) Modern methods for business research. Lawrence Erlbaum Associates, Mahwah, NJ, pp 295–336
8. CNVSU – Comitato Nazionale per la Valutazione del Sistema Universitario (2007) Note tecniche su dati ed informazioni per la Rilevazione Nuclei 2007, DOC 3/07
9. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, New York, NY
10. Esposito Vinzi V, Trinchera L, Squillacciotti S, Tenenhaus M (2008) REBUSPLS: A response-based procedure for detecting unit segments in PLS path modeling. Appl Stochastic Models Bus Ind (ASMBI) 24:439–458
11. Grilli L, Rampichini C (2007) Multilevel factor models for ordinal variables. Struct Equ Modeling 14(1):1–25
12. Guolla M (1999) Assessing the teaching quality to student satisfaction relationship: applied customer satisfaction research in the classroom. J Mark Theory Pract 7(3):87–98
13. Hahn C, Johnson M, Herrmann A, Huber F (2002). Capturing customer heterogeneity using a finite mixture PLS approach. Schmalenbach Bus Rev 54:243–269
14. Jöreskog KG, Sörbom D (1979) Advances in factor analysis and structural equation models. Abstract Books, Cambridge, MA
15. Lovaglio PG (2002) La stima di variabili latenti da variabili osservate miste. Statistica LXII 2:203–213
16. Martensen A, Gronholdt L, Eskildsen JK, Kristensen K (2000) Measuring student oriented quality in higher education: application of the ECSI methodology. Sinergie Rapporti di Ricerca 9:372–383
17. Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffman A, Giovannini E (2005) Handbook on constructing composite indicators: methodology and user guide. OECD statistics working paper
18. Ramipichini C, Grilli L, Petrucci A (2004) Analysis of university course evaluations: from descriptive measures to multilevel models. Stat Methods Appt 13(3):357–373
19. Tenenhaus M (2008) Component-based structural equation modelling. Total Qual Manage Bus Excel 19(7):871–886
20. Tenenhaus M, Esposito Vinzi V, Chatelin YM, Lauro NC (2005) PLS path modeling. Comput Stat Data Anal 48:159–205
21. Trinchera L (2007) Unobserved heterogeneity in structural equation models: a new approach in latent class detection in PLS path modeling. PhD thesis, DMS, University of Naples
22. Trinchera L, Russolillo G (2009) Role and treatment of categorical variables in PLS path models for composite indicators. In Esposito Vinzi V, Tenenhaus M, Guan R (eds) Proceedings of the 6th international conference on partial least squares and related methods, pp 23–27, PHEI, ISBN: 978-7-121-09342-5
23. Werts CE, Linn RL, Jöreskog KG (1974) Intraclass reliability estimates: testing structural assumptions. Educ Psychol Meas 34(1):25–33
24. Wold H (1982) Soft modeling: the basic design and some extensions. In: Jöreskog KG, Wold H (eds) Systems under indirect observation, Part 2. North-Holland, Amsterdam, pp 1–54
25. XLSTAT (2009) Addinsoft, Paris, France (www.xlstat.com)

Chapter 5
A Study on University Students' Opinions about Teaching Quality: a Model Based Approach for Clustering Ordinal Data

Marcella Corduas

5.1 Introduction

In complex surveys aimed at measuring the satisfaction level of final users of a given product or service, several items are generally investigated. Also, respondents often belong to different categories since they are stratified according to relevant features, such as geographic location or gender. In such situations, the comparison among the distributions of ratings given to a selection of items by interviewees or to a single item by different groups of interviewees can provide a meaningful summary of observed data.

Sometimes, decision makers, who are interested in information arising from statistical analysis, are not very familiar with statistical theory. For this reason, graphical techniques or simple statistical indices (such as the average or mode) have been widely used in empirical analysis. However, this approach wastes relevant information about the distribution of ratings since other aspects concerning shape are not considered. In this respect, for instance, [7] introduced a statistical index based on a distance measure between ordinal data distributions; [2] discussed the fundamental principles for constructing composite indicators and examined some specific measures for assessing university teaching quality. In this chapter, a mixture distribution is used for modeling ratings and a procedure for detecting significant similarities and differences in the distribution of judgements expressed by raters is proposed. Specifically, the case study refers to the yearly survey done at the University of Naples Federico II in order to assess teaching quality.

The chapter is organized as follows. In Sect. 5.2, some results concerning a mixture distribution for ordinal data are briefly illustrated. Then, in Sects. 5.3 and 5.4, a testing procedure based on Kullback-Liebler (KL) divergence and a clustering technique are discussed, and finally, in Sect. 5.5, a case study is presented.

M. Corduas (✉)
Dipartimento di Scienze Statistiche, Università di Napoli Federico II, 80138 Napoli, Italy
e-mail: corduas@unina.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_5,

5.2 A Mixture Distribution for Ordinal Data

A statistical model for ordinal data has been recently proposed by [12]. The model provides the probability distribution of the random variable generating ordinal data describing judgments or evaluations that respondents express on a given item.

Specifically, the preference or score is represented by the random variable R such that:

$$P(R = r) = \pi \binom{m-1}{r-1} (1-\xi)^{r-1} \xi^{m-r} + (1-\pi)\frac{1}{m}, \quad r = 1, 2, \ldots, m \quad (1)$$

where $\pi \in (0, 1]$, $\xi \in [0, 1]$, and m is the number of grades for evaluating an item.[1] For $m > 3$, (1) is a mixture of a *Uniform* and a (shifted) *Binomial* distribution. The parameter π determines the role of *uncertainty* in the final judgment: the lower the weight $(1-\pi)$ the smaller the contribution of the Uniform distribution in the mixture and then the smaller is the subject uncertainty in selecting the final rating. Moreover, the parameter ξ characterizes the shifted Binomial distribution and, therefore, it is related to the strength of the intimate belief of respondents concerning the object of evaluation. In other words, $(1 - \xi)$ is the strength of the positive *feeling* expressed by raters about the item (see, [10, 21] for a discussion). Then, the closer ξ is to 1 the less the item has been rated positively.

The model is very flexible and is capable of describing distributions having very different shapes. The formulation of the asymmetry and kurtosis coefficients of R as function of the π and ξ parameters have been derived by [20]. Specifically, it can be shown that $Asim(\pi, \xi) = 0$ for $\xi = 0.5$ and, in addition, $Asim(\pi, \xi) = -Asim(\pi, 1 - \xi)$, for a given $\pi \in (0, 1]$. When $\xi < 0.5$, the distribution of R is skewed negatively and the probability that raters express positive opinions about the given item increases as ξ moves towards 0. The opposite consideration applies when $\xi > 0.5$, the distribution of R is skewed positively and the probability that raters express negative opinions increases as ξ moves towards 1. Also, for a given $\pi \in (0, 1]$, the kurtosis increases as ξ approaches the borders of the parameter space, and $Kurt(\pi, \xi) = Kurt(\pi, 1 - \xi)$.

The influence of external factors in the final judgement may be introduced by adding two relations which connect the model parameters to significant *covariates* by means of a logistic link function (see [22]). This fact originated the acronym CUB which the authors used to identify the model. Note that, in the rest of this chapter, that acronym will simply denote the model (1).

Finally, given the observed ratings $\mathbf{r} = (r_1, r_2, \ldots, r_N)'$ expressed by N judges towards a certain item, the log-likelihood function for the model (1) is:

$$logL(\boldsymbol{\theta}) = \sum_{r=1}^{m} n_r \; log(p_r(\boldsymbol{\theta})), \quad (2)$$

[1] In order to facilitate the reading, according to the special case study which will be examined in the final section, in the following discussion, we assume that 1 refers to the worst judgement and m to the best one.

being $\boldsymbol{\theta} = (\pi, \xi)'$ the parameter vector, n_r the observed frequency of $R = r$, $(r = 1, \ldots, m)$ and $p_r(\boldsymbol{\theta}) = P(R = r|\boldsymbol{\theta})$. Maximum likelihood estimation of $\boldsymbol{\theta}$ can be performed by E-M algorithm; an efficient procedure is discussed by [21].

5.3 The Kullback-Liebler Divergence

CUB models provide a meaningful and parsimonious parametric representation of rating distributions which can be used for clustering purposes. In this respect, a measure of dissimilarity is needed. Specifically, we introduce the KL divergence measure and, in this section, we briefly recall some useful results.

In general, KL divergence measures the dissimilarity between two probability distributions $f_1(x, \theta_1)$ and $f_2(x, \theta_2)$ characterizing a random variable X under two different hypotheses, respectively [16].

Specifically, the KL divergence is defined as:

$$J(f_1, f_2) = I(f_1, f_2) + I(f_2, f_1), \tag{3}$$

where, assuming the case of a continuous random variable:

$$I(f_1, f_2) = \int_{-\infty}^{\infty} f_1(x, \theta_1) \ln \frac{f_1(x, \theta_1)}{f_2(x, \theta_2)} dx = E_1 \left(\ln \frac{f_1(x, \theta_1)}{f_2(x, \theta_2)} \right) \tag{4}$$

is the mean information, with respect to f_1, for discrimination in favor of the first hypothesis against the second one. The other term in (3), $I(f_2, f_1)$ is similarly defined. Of course, the case of a discrete random variable can be easily introduced by extending (4) accordingly.

Note that the KL divergence is not a metric. As a matter of fact, it satisfies the following relationships: $J(f_1, f_2) \geq 0$, the equality holds if and only if f_1 and f_2 coincides; $J(f_1, f_2) = J(f_2, f_1)$; but it doesn't satisfy the triangular inequality.

However, due to its statistical properties, it represents a very interesting tool for establishing the comparison of CUB models as a problem of hypotheses testing.

For this aim, we illustrate a general result derived from [17]. Consider two discrete populations each characterized by a probability distribution function having the same functional form $p(x, \boldsymbol{\theta}_i)$ with unspecified parameters $\boldsymbol{\theta}_i$, $i = 1, 2$. Also assume that, for all points in the random variable support, $p(x, \boldsymbol{\theta}_i) > 0$. Suppose that two samples of N_1 and N_2 observations have been randomly drawn from the specified i-th population and we wish to decide if they were in fact generated from the same population. In order to test the hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ against $H_1 : \theta_{1j} \neq \theta_{2j}$ (for at least one of the distribution parameters), the KL divergence statistic is defined by:

$$\hat{J} = \frac{N_1 N_2}{N_1 + N_2} \left[\sum_x (p(x, \boldsymbol{\theta}_1) - p(x, \boldsymbol{\theta}_2)) \ln \frac{p(x, \boldsymbol{\theta}_1)}{p(x, \boldsymbol{\theta}_2)} \right]_{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2}, \tag{5}$$

where the vector parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ have been replaced by the maximum likelihood estimators. It can be shown that $\hat{J}$ is asymptotically distributed as a χ^2_g random variable when the null hypothesis is true, being g the dimension of the vector parameter [16]. In the case under investigation, the objects of comparison are CUB distributions, each characterized by $g = 2$ parameters, then the $100\alpha\%$ critical region for hypotheses testing is simply given by: $\hat{J} > \chi^2_{(2,\alpha)}$.

5.4 Clustering

Although CUB models only describe univariate distributions of judgements, they may help to give further insights into data originated by complex surveys.

In literature several approaches have been proposed for clustering ordinal data. The problems related to the choice of an adequate measure of dissimilarity between ordinal data and the necessary techniques for producing clustering have been investigated (see, for instance, [23] for a review).

Moreover, model-based approaches which use estimated membership probabilities to classify cases into the appropriate cluster have been introduced and widely studied (see [5, 6, 14] for a general discussion). In this respect, a well established technique is the mixture-model clustering, where each latent class represents a hidden cluster ([19, 24] and references therein).

The approach that we discuss in the present chapter moves from a different point of view since the elements which are object of comparison are the distributions of ratings. In other words, the focus is not on the judgements that each individual expresses, but on the shape of the overall rating distribution that their judgements originate for a given item. Moreover, it is worth noting that we are not postulating that the population is clustered in two groups behaving according to one of the two unobserved components of the mixture distribution (1). The latter is only a probability distribution which results being flexible enough to represent observed ratings, but any other distribution which ensures a good fit for the data may be used.

Multivariate approaches, which account for the dependence among judgments expressed by subjects, exploit data information more efficiently than CUB models which, at this stage of the work, simply represent univariate distributions. However, CUB models have proved to be effective in numerous real applications arising in various fields such as social analysis [15], medicine [11], marketing [22], linguistics [1] and others (see [10] for a discussion) and, for this reason, they deserve further attention.

Coming to the clustering problem, the strategy that we propose relies on the following steps:

- firstly, the mixture distribution (1) is fitted to the observed rating distributions concerning the items object of evaluation;
- secondly, each estimated CUB model is compared with the others by means of the KL divergence and a symmetric square matrix of KL divergences between all models is evaluated;

– finally, a hierarchical clustering technique (complete or simple linkage method) is applied to the matrix of divergences.

Regarding the last step, we suggest that the $100\alpha\%$ critical value, derived for hypotheses testing, be used for sectioning the dendrogram and for the subsequent identification of groups.

As is well known, both clustering methods impose a fixed hierarchical rule in order to decide when a new group has to be created: simple linkage method allows for elongated clusters whereas complete linkage method tends to recognize compact clusters. This fact reduces the flexibility of the approach with respect to other approaches, such as the BEA algorithm which has been studied by [9, 18].

Notwithstanding this limit, hierarchical clustering still represents an effective device to provide a very simple graphical display of data.

Moreover, having used the $100\alpha\%$ critical value derived for hypotheses testing as threshold makes the interpretation of resulting clusters more meaningful. As a matter of fact, with reference to complete linkage method, the suggested criterion for dendrogram's sectioning ensures that all ratings distributions belonging to the same cluster have been generated by the same population. Instead, as regards simple linkage method, the criterion ensures that, given a certain group, there exists a single link path along clustered elements which joins ratings distributions which are similar according to the KL divergence test.

5.5 The Analysis of Students' Opinions

In this section, by means of the analysis of an empirical data set, we will illustrate how the proposed technique can be used in practice for clustering rating distributions. In particular, the study refers to a real data set from the yearly survey on students' opinions about teaching quality at the University of Naples Federico II.[2]

5.5.1 The Data Set

According to the CNVSU guidelines [8], the questionnaire aims at assessing the students' opinions about various elements which characterize teaching activity: (1) quality of lecture halls and teaching equipments; (2) several features of the specific course the interviewees are attending; (3) instructor's abilities: clear explanations, ability to inspire and motivate students' interest in the course content, instructor's availability for consultation outside of class, time-table respect, instructor's concern for students' learning problems and adequacy of textbooks and other material.

[2] The problems related to the evaluation of university teaching and services has been widely discussed (see, for instance, [3, 4, 13]). Further references can be found at the following websites: http://dssm.unipa.it/divago/ for the *DIVAGO* project, http://www2.stat.unibo.it/prin2006/ for the PRIN2006 project on "Metodi e modelli statistici per la valutazione dei processi formativi", http://valmon.ds.unifi.it for the *VALMON* project.

Table 5.1 Students' profiles and participation

Year	(%)	HS qualification	(%)	Attendance	(%)	N.courses	(%)
1	35.9	Classical studies	19.8	<20%	0.7	1	5.3
2	29.6	Scientific education	53.3	20–50%	2.9	2	10.8
3	21.7	Technical education	17.1	50–80%	20.1	3	22.0
4	4.0	Professional education	9.8	>80%	76.3	4	30.2
5	1.8					5	21.3
6	0.5					>5	10.5
>6	6.5						

The data set consists of 34,507 valid records. It was gathered at the end of 2005–2006 academic year from the 13 Faculties belonging to the University Federico II (Medicine, Veterinary medicine, Pharmacy, Agricultural Science, Biotechnology, Engineering, Architecture, Mathematics and Natural Science, Classics and Modern Studies, Law, Economics, Political Sciences, Sociology). The students' ratings are expressed using a 7 point Likert scale where 7 relates to the highest positive judgement.[3]

Table 5.1 illustrates the students profile and participation. More than half of the interviewees have a scientific education at senior high school level; only 20% have specialized in classical studies and the remaining 27% have a technical or professional education.

The attendance of courses is generally high; only 3.6% of the students show low attendance of courses. Since the questionnaires are submitted in the last weeks of the term, this result confirms that most students have a solid knowledge of the course they are requested to evaluate and repute it important for their curricula. In this respect, it is worth noting that students usually are not requested to attend courses in order to be admitted to final exams. Then, it is more likely that those who attend lectures regularly are satisfied about teaching. Moreover, about 84% of students attend more than two courses; in other words, interviewees prefer to attend most courses offered in each term within their curricula. This is in part justified by the fact that most students (87%) are enrolled in the first 3 years of their university degree course.

5.5.2 The Results

The rest of this study refers to the assessment of 6 instructor's abilities. Specifically, CUB models have been fitted to the rating distributions of each items observed for the 13 Faculties. The parameter estimates were significant in all the examined cases.

Figure 5.1 displays the estimated parameters in the (π, ξ) parameter space. Since, $\widehat{\pi} \in (0.57, 0.99)$ and $\widehat{\xi} \in (0.18, 0.37)$ only a sub-region of the unit square is shown.

[3] The study presented in the chapter refers to Faculties and no investigation at a lower level of aggregation (such as curricula) has been carried out. As a matter of fact, the use of the database was restricted by the University of Naples Federico II which also requested that no identifier establishing the identity of the Faculties could be included in any publication.

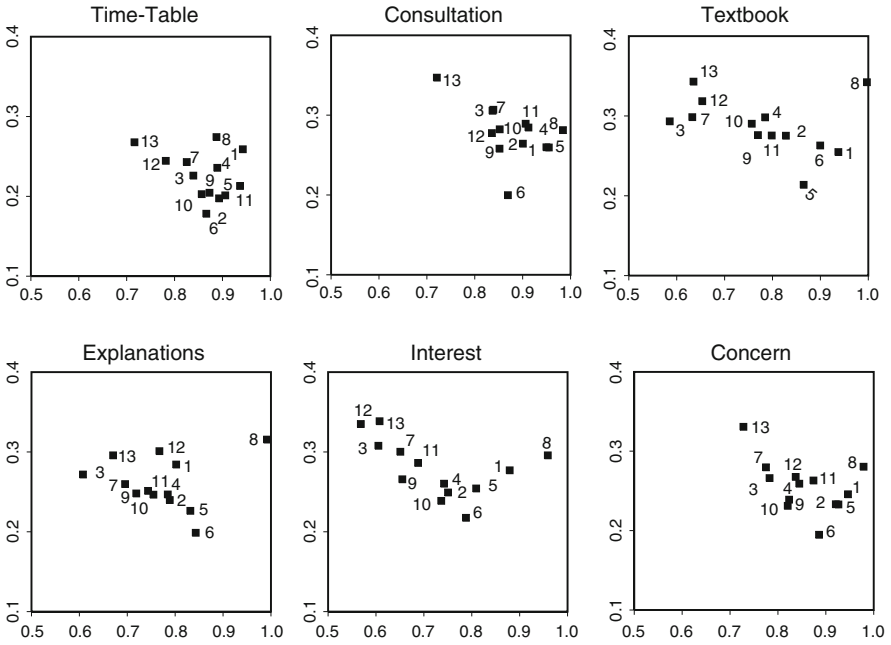


Fig. 5.1 CUB models of the rating distributions of instructor's abilities

In general, students express a rather positive evaluation of the teacher's abilities as confirmed by the high values of $(1 - \widehat{\xi})$. But, at least at this stage of the analysis, this consideration does not allow a clear discrimination among Faculties since the estimated values, $\widehat{\xi}$, appear very close. Interestingly, the *uncertainty*, measured by $(1 - \widehat{\pi})$, is generally low but it seems to vary more widely among Faculties, and this behaviour is common to all the items. This implies that the students of some Faculties (for instance (8)), have a stronger attitude towards the assessment with respect to others and they express a more convinced opinion.

The representation of CUB models in the parameter space can not be used for clustering purposes, since, as discussed in [9], the Euclidean distance among points, established by visual inspection of the graph, does not reflect the true dissimilarity between the shape of the underlying distribution. Even a small change in position of a point in reference to the orizontal and vertical axis implies very different consequences in terms of the shape of the related rating distribution. For this reason we use the KL divergence to compare CUB models and proceed according to the strategy previously described.

First, we examine in some detail the instructor's ability to raise student interest.

The dendrogram derived by complete linkage method (Fig. 5.2) helps the identification of clusters of Faculties which students assign similar ratings to the considered item. The threshold used for ending the aggregation process is the percentile $\chi^2_{(2,0.01)}$. The dotted line in the graph corresponds to a very extreme divergence.

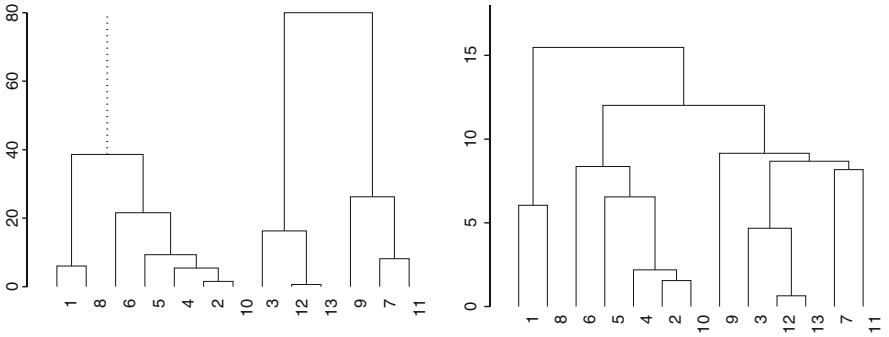


Fig. 5.2 Ability to raise student interest: complete linkage (*lhs*), single linkage method (*rhs*)

From the inspection of the dendrogram, the following elements are immediately connected: (1, 8) (5, 4, 2, 10) (12, 13) (7, 11).

If elongated clusters are allowed (see the single linkage dendrogram in Fig. 5.2), other elements merge so that only three clusters are classified: G1 = (3, 12, 13, 7, 11, 9) which refer to the Faculties strongly related to professional skills and vocational education, G2 = (2, 4, 5, 10, 6) which includes the Faculties related to humanities and arts, and G3 = (1, 8) which includes two Faculties with a marked specialization. Figure 5.3 shows the estimated rating distributions belonging to each group.

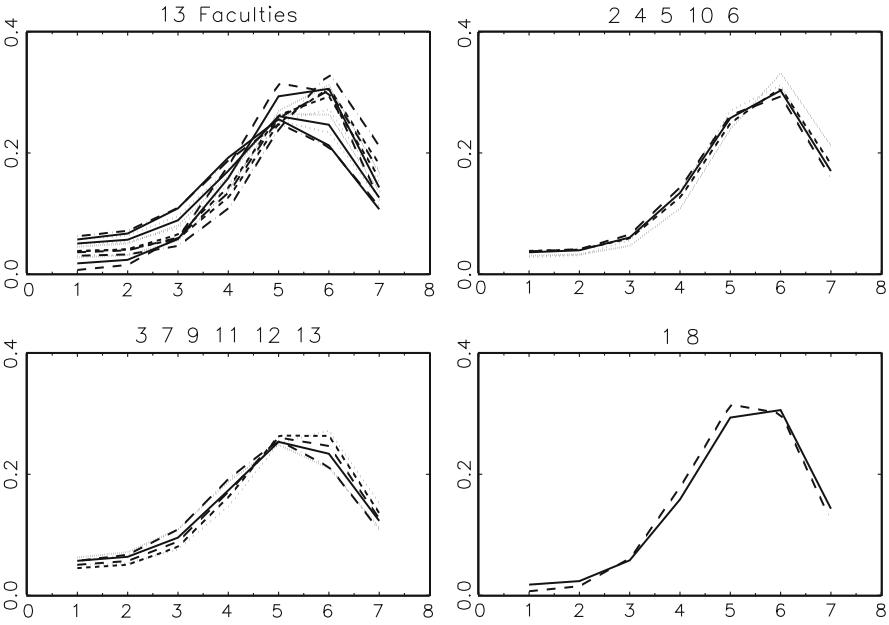


Fig. 5.3 Clustered estimated CUB distributions

It is evident that KL-divergence is able to discriminate between distributions which apparently are characterized by similar overall pattern.

The representation of the CUB models in the parameter space (Fig. 5.1) shows that the resulting clusters (from G_1 to G_3) are ordered according to the π estimates, that is in terms of decreasing uncertainty. Furthermore, due to the specific shape of the rating distributions, this fact reflects the ordering of the groups in terms of increasing probability of positive judgements (which are 0.62, 0.73 and 0.74, respectively).

It is not surprising that students attending Faculties in G_1 find the judgement of this item is more problematic with respect to the other two groups. As a matter of facts, students of those Faculties are generally more demanding. In addition, on the one hand the related disciplines are very technical and less fascinating, on the other a number of instructors are not specifically trained for classroom teaching practice since they are involved in professional activities.

In Figure 5.4, the dendrograms of the remaining items are illustrated. In order to facilitate the reading, due to the presence of extreme divergences, the joining level of extreme points has been rescaled. The horizontal line helps to detect the groups identified by cutting the unscaled dendrogram according to the criterion described in the previous section.

Some behaviours are worth of comments. Firstly, the clustering of the rating distributions concerning instructors' ability to explain concepts clearly reveals the presence of several isolated Faculties and few small groups. This may be due to the fact that "clarity" is a very personal trait of instructors, and in addition it may depend to some extent by the nature of disciplines. It is, therefore, unlikely that large clusters can be recognized at the considered level of aggregation.

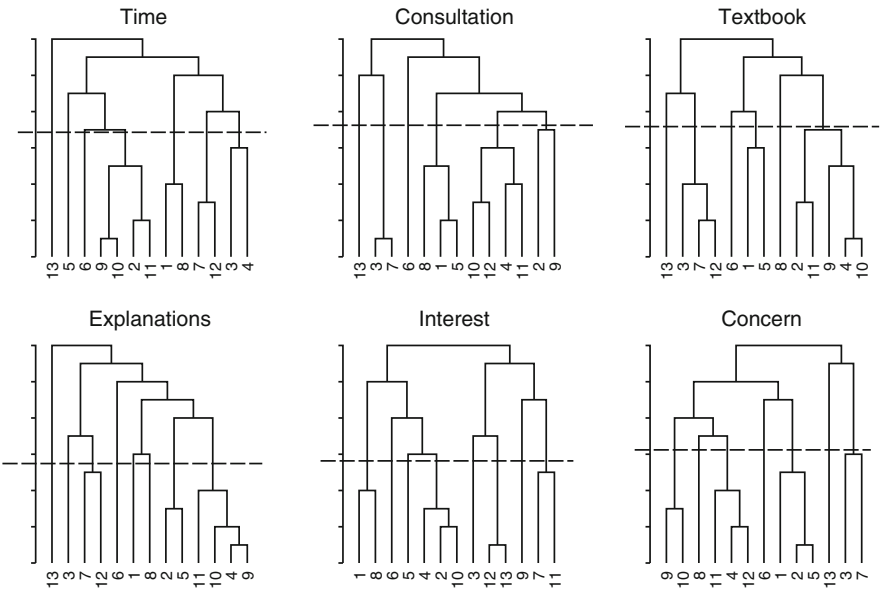


Fig. 5.4 Dendrogram by complete linkage method

From the comparison, other distinctive characteristics emerge. Faculties (6) and (13) are generally isolated. The former achieves the highest probability of positive judgements with respect to the others whereas the latter attain the lowest one. Moreover, (3) is generally close to (7) for most items confirming the partial overlapping of the two Faculties with respect to the scientific areas which the instructors are related to. Furthermore, (4), (9), (10) and (11) show many similarities with respect to various items. The Faculties (8) and (1) seem to share a common behaviour of the students rating distributions on two organizational aspects, "time table respect" and "availability outside of class". Finally, (2) and (10) receive similar evaluation on "clarity", "time table respect" and "adequacy of textbooks", but they belong to different groups with respect to the other two items which instead concern the interaction with students.

5.6 Final Remarks

In this chapter an approach to ordinal data clustering based on KL divergence has been presented. The proposed technique helps the identification of similarities in the behaviour of groups of judges when they are asked to express their ratings on a set of items. Specifically, the technique is able to discriminate the scores distributions even when the latter are apparently characterized by similar overall patterns. Moreover, combining the hypotheses testing based on KL divergence with the clustering technique adds further strength to the identification of groups.

Acknowledgements This research is part of PRIN 2006–2008 project on: "Metodi e modelli statistici per la valutazione dei processi formativi" and has benefited from support of MiPAAF ex CFEPSR (Portici). The author thanks the University of Naples Federico II, and especially the Nucleo di Valutazione di Ateneo and UPSV for kindly providing the data set which has been analyzed in this chapter.

References

1. Balirano G, Corduas M (2008) Detecting semiotically expressed humor in diasporic TV productions. Humor: Int J Humor Res 3:227–251
2. Bernardi L, Capursi V, Librizzi L (2004) Measurement awareness: the use of indicators between expectations and opportunities. In: Atti della XLII Riunione Scientifica SIS. Cleup, Padova, pp 315–326
3. Biggeri L (2000) Valutazione: idee, esperienze, problemi. Una sfida per gli statistici. In: Atti della XL Riunione Scientifica SIS. CS2p, Firenze, pp 31–48
4. Biggeri L, Bini M (2001) Evaluation at University and State level in Italy: need for a system of evaluation and indicators. Tertiary Educ Manage 7:149–162
5. Bock HH (1996) Probabilistic models in cluster analysis. Comput Stat Data Anal 23:5–28
6. Bock HH (1998) Probabilistic aspects in classification. In: Hayashi C, Yajima K, Bock HH, Oshumi N, Tanaka Y, Baba Y (eds) Data science, classification and related methods. Springer, New-York, NY, pp 3–21
7. Capursi V, Porcu M (2001) La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni di giudizi. In: Atti della Riunione Intermedia SIS on "Processi e Metodi Statistici di Valutazione", Roma

8. CNVSU (2002) Proposta di un insieme minimo di domande per la valutazione della didattica da parte degli studenti frequentanti. Doc 9/02, http://www.cnsvu.it
9. Corduas M (2008) A testing procedure for clustering ordinal data by CUB models. Proceedings of SFC-CLADAG 2008 Meeting. ESI, Napoli, pp 245–248
10. Corduas M, Iannario M, Piccolo D (2009) A class of statistical models for evaluating services and performances. In: Bini M, Monari P, Piccolo D, Salmaso S (eds) Statistical methods for the evaluation of educational services and quality of products. Contribution to Statistics. Physica-Verlag, Heidelberg, pp 99–117
11. D'Elia A (2007) A statistical modelling approach for the analysis of TMD chronic pain data. Stat Methods Med Res 16:1–15
12. D'Elia A, Piccolo D (2005) A mixture model for preference data analysis. Comput Stat Data Anal 49:917–934
13. Fabbris L (ed) (2006) Effectiveness of university education in Italy: employability. Competences, human capital. Springer, Heidelberg
14. Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. JASA 97:611–631
15. Iannario M (2007) A statistical approach for modelling Urban Audit Perception Surveys. Quaderni di Statistica 9:149–172
16. Kullback S (1959) Information theory and statistics. Dover Publ., New York, NY
17. Kupperman M (1957) Further applications of information theory to multivariate analysis and statistical inference. George Washington University, Washington, DC
18. McCormick WT, Schweitzer PJ, White TW (1972) Problem decomposition and data reorganization by a clustering technique. Oper Res 20:993–1009
19. McLachlan GJ, Basford KE (1988) Mixture models: inference and application to clustering. Marcel Dekker, New York, NY
20. Piccolo D (2003) On the moments of a mixture of Uniform and shifted Binomial random variables. Quaderni di Statistica 5:85–104
21. Piccolo D (2006) Observed information matrix for MUB models. Quaderni di Statistica 8:33–78
22. Piccolo D, D'Elia A (2008) A new approach for modelling consumers' preferences. Food Qual Prefer 19:247–259
23. Podani J (2005) Multivariate exploratory analysis of ordinal data in ecology: pitfalls, problems and solutions. J Veg Sci 16:497–510
24. Vermunt JK, Magidson J (2002) Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL (eds) Advances in latent class analysis. Cambridge University Press, Cambridge

Chapter 6
The Impact of Teaching Evaluation: Factors that Favour Positive Views from Student Representatives

Simone Gerzeli

6.1 Introduction

The surveying of student opinions on teaching activities was introduced as an evaluation tool in Italian Universities in 1999. The aim of the considerable legislative changes put into place by the Ministry of Universities and Scientific and Technological Research, and the procedural efforts carried out by the National Committee for the Evaluation of University Systems (CNVSU), were to set up a procedure for teaching evaluation, where student opinions were to become one of the key parameters used within universities for "decision-making" purposes.

These evaluation activities have reached an advanced stage both in terms of their unanimous acceptance by the universities and the development of common instruments and methodologies used. The compulsory nature of teaching evaluation has been an important stimulus and, in different periods and different ways, all universities have adopted the guidelines provided by CNVSU. Nevertheless, until now, a systematic study aimed at assessing the results of such practices in Italy had not been carried out and the impact that such evaluations may have had upon teaching remained to be determined.

This chapter presents the results of a meta-evaluative study that was carried out in response to the need for a critical evaluation of the problems connected to the construction and application of measures to evaluate the university systems.

Two different statistical research projects were undertaken in four universities – Padua, Palermo, Pavia and Siena. They surveyed the faculty deans and student representatives to find out if and how the various faculties could obtain suggestions and policy criteria from the results of the evaluation [4]. The results from the survey of deans, discussed by Gerzeli et al. [2], showed that the impact of the teaching evaluation was different between the four universities considered: the way in which the opinions were gathered seemed to play an important role as did how the results were read and shared among the students, teachers etc. Moreover, there was a large

S. Gerzeli (✉)
Dipartimento di Statistica Applicata ed Economia "Libero Lenti", Università di Pavia, Pavia, Italy
e-mail: simone.gerzeli@unipv.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_6,

amount of variability between the views of the deans, independent of the evaluative experience of the individual universities. This leads us to hypothesize that a certain degree of "personalization" exists between faculties, also within the same university, regarding how the evaluation procedures are carried out, and that individual faculty traditions may also play an important role.

The objective of this study was to identify and measure what student representatives' perceive to be the effects of teaching evaluation upon the university system in order to understand the factors that can favour a positive impact resulting from this practice.

6.2 Methods

6.2.1 Study Design

Since these evaluation practices are focused on student opinions about the courses they attend, their points of view regarding the effects of the evaluation are also of particular importance. It is also reasonable to expect that they will express themselves with more sincerity and in more detail if they believe that their opinions and ideas will be actually used (analyzed and discussed) and then taken into consideration when determining new measures for improvement of teaching.

In order to gather information about student views, a survey was carried out on the student representatives (considered to be privileged observers) as they are, theoretically, more interested and informed about university issues and more perceptive and willing to collaborate in such a project. The information obtained completes the picture presented by the results from the survey of the deans, whose views represent those of the faculty and thus the "institutional" stance on the university's organization of teaching activities.

The survey of the student representatives was done using a web-based questionnaire from November 2006 to February 2007. To maximize participation within each university, different strategies were used to present the survey, even involving groups and individuals who hold vested interests in the outcome of the study; including the evaluation committee itself, the academic senate and other key people who form part of the universities' managements. Due to the differing contexts, each university chose to adopt specific strategies. In Siena, the representatives were contacted by e-mail; when e-mail addresses were not available (for example, the faculty of medicine) a letter was sent to the representative for the corresponding faculty. In Padua, e-mails were sent to all representatives holding valid e-mail addresses. In Pavia, the invitation to participate was sent by both e-mail and normal post to all representatives with valid addresses. In Palermo, only the economics, engineering, and medicine faculties participated and only the representatives with a valid e-mail address were contacted.

A questionnaire was prepared, validated by a group of evaluation experts and tested on a group of students. The questionnaire asked the respondent to first clarify their personal characteristics, before going on to ask for opinions on the impact

of teaching evaluation. The addition of these opinions to the data collected by the survey of deans permitted the comparison of the students' perceptions with those of the deans. The questions regarding the impact of the teaching evaluation were divided into: (i) accessibility to the evaluation results; (ii) changes produced by the evaluation; (iii) the perceived usefulness and adequateness of the evaluation instruments.

6.2.2 Statistical Analysis

The initial presentation of the results was mainly descriptive and aimed at underscoring the consequences of the teaching evaluation survey, looking at how the survey related to individual and contextual factors.

To identify better those factors that favour a positive impact of the teaching evaluation, a hierarchical (or "multilevel") regression model was used [3, 5, 6]. Given the results from the survey of deans [2] and the initial descriptive analysis, we were particularly interested in verifying if, and to what extent, the individual faculty contexts influenced student opinions. Given the structure of the data and the limited number of interviews from each faculty, the multilevel approach seemed to be the most appropriate for this purpose.

These models are able to measure the net effect of possible determinants (explanatory factors), taking into account the hierarchical structure of the given objects of the study. A multilevel model is advisable whenever the first-level units – in our case the student representatives – are "naturally" aggregated into different groups (second-level units) – i.e. the faculties.

In fact, the universities could even be taken into consideration as third level units; however, the addition of this level did not improve the information power of the model, so the analyses were performed for only the first and second units. Perhaps the faculties' opinions towards teaching evaluation are more important than those of the universities themselves. It should also be noted that second level units also take the specific university into account, because faculties with the same name, belonging to different universities, were considered as separate second level units.

In this situation, we can say that the variability of the studied phenomenon depends not only upon the individual explanatory variables (first level) but also upon the fact that each student belongs to a particular faculty from a specific university with its own unique characteristics that renders the faculty in question distinct from others. Specific to our study, we have already emphasized that the single faculties, even at the same university, contain a certain degree of personalization in their teaching "evaluation practices".

The applied multilevel random intercept model is defined using the indices i and j to indicate the ith student representative and the jth faculty as follows:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta' \mathbf{x}_{ij} + \gamma' \mathbf{w}_j + u_j + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim N(0; \sigma^2) \\ u_j &\sim N(0; \tau^2) \end{aligned} \tag{1}$$

where y_{ij} is the outcome variable, x_{ij} is a vector of observed explanatory variables, measured at the student i of faculty j level to which the vector of fixed coefficients β is associated, β_0 is a fixed quantity that refers to all the representatives. The term w_j is a vector of observed explanatory variables at faculty j level with the associated vector of parameters γ. The term u_j is a random variable with mean 0 and variance τ^2 indicating the random effect of faculty j. Finally ε_{ij} is the random error that represents the unexplained variation referring to the representative i in faculty j. The intercept for faculty j is thus given by a fixed component, β_0, plus a random component, u_j. We have not considered the hypothesis that the regression coefficients included into the β vector can vary between the faculties (random slope).

The degree of dependence between the observations is measured by the intraclass correlation coefficient (ρ), defined as:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{2}$$

This reflects the degree of "nesting" within the hierarchical data structure.

From the various models studied, we will only report the most significant outcomes.

Two multilevel regressions were applied, taking as outcome variables changes in teaching organization and perceptions regarding the usefulness of teaching evaluation. These variables were selected because they seemed to better differentiate the opinions of the representatives.

The factors that can favour a positive impact were identified based on an understanding of the phenomenon being investigated and on the results of the descriptive analyses. Thus, the first-level explanatory variables (which refer to the individual representatives) included in the estimates of the model were:

1. whether or not the individual was a representative on the teaching committee (TC) – Commissione Didattica paritetica – as opposed to other bodies (faculty boards – Consiglio di Facoltà – or degree teaching committees – Consiglio di Corso di Laurea). The teaching committee appears to function as a kind of proxy that promotes the contemplation of the results from the teaching evaluation (yes vs. no);
2. whether or not the individual served at least one term as a student representative: this experience may better enable the individual to perceive the changes and usefulness of the teaching evaluation (second term or ex-representative vs. first term);
3. the accessibility of the evaluation results to the students: this factor represents an indication of the impact of the evaluation, but also a precondition for the perception of the changes and use of the evaluation procedure (yes vs. no);
4. agreement between the representatives and the faculty deans on the discussion of the evaluation results; discussions are carried out by specific assessment bodies set up for that exact purpose: this factor can favour a positive impact (yes vs. no).

The second-level units considered were the 28 faculties in which at least 5 representatives responded. The explanatory variable for second-level units considered was to what degree the various courses were covered, considering the students that responded to the teaching evaluation survey, assuming that higher levels of coverage would correspond to a greater impact of the data gathered.

6.3 Results

6.3.1 Respondents

One clear result from the survey was that contacting student representatives to ask them to participate in a project was difficult. Of the 1,310 representatives from the four universities, valid e-mail addresses could be identified for only 970 students; thus this is the number of students that were requested by email to complete the questionnaire. A total of 410 representatives responded, equal to 42.3% of those potentially contacted. There was a noticeable difference between universities (Table 6.1) and between faculties within each university. It must be emphasized that this response rate is undoubtedly an underestimate of the number of representatives actually contacted to participate in the survey; but, it was not possible to quantify the number of students who had received and read the request.

The particularly low response rate implies that the results should be interpreted with caution; more attention should be paid to the size and the trends of the phenomena observed than to their precise estimation.

Most of the differences between the universities seem to be attributable to the differences in the communication channels used to contact the students and not by differences in student levels of interest in the project. In some cases, assistance from the university bodies enabled us to obtain valid addresses for almost all of the students; however, in other cases we were not able to do so for a considerable number of representatives.

In order to interpret the results of the analysis for each university correctly, it is important to keep in mind that student representative participation was restricted and highly non-homogeneous between faculties. The respondents probably represent a select group of the most "stable" individuals (holding valid and accessible e-mail addresses), who are more likely to be "motivated" to express their opinions.

Table 6.1 Participation in the survey

	Padua	Palermo	Pavia	Siena	Total
No. of representatives	444	256	263	347	1310
No. of representatives potentially contacted	323	164	213	270	970
Gross response rate[a] (%)	40.9	26.2	58.2	41.1	42.3

[a] The ratio between the number of respondents and the number of student representatives potentially contacted.

Table 6.2 Distribution by faculty and university of the number of respondents

Faculty	Padua	Palermo	Pavia	Siena
Agriculture	3	0	0	0
Economics	4	6	13	21
Pharmacy	5	0	8	8
Law	11	0	6	9
Engineering	18	16	12	7
Arts and Philosophy	12	0	14	25
Medicine	16	21	22	9
Veterinary Science	8	0	0	0
Music	0	0	7	0
Psychology	11	0	0	0
Science of Education	3	0	0	0
Science	27	0	23	27
Political Science	0	0	13	3
Statistical Sciences	13	0	0	0
Inter-faculty	1	0	6	2
Total	132	43	124	111

Nevertheless, the evidence shown below is enough to sound important "alarm bells", even when considering all the possible limitations of coverage and responses.

Of the 410 respondents, 37.8% are second-term or ex-representatives and 58.5% are represented on the teaching committee (TC).

Table 6.2 shows the distribution by faculty of the number of respondents. We can clearly see a strong variability not only due to the size and type of faculty but also to the differing number of individuals who were actually contacted. In some faculties participation was less than in others since it was not possible to have a valid address for every representative.

The multilevel analysis considers the 28 faculties with at least 5 respondents, for which the response rate rises to over 50%.

6.3.2 The Availability and Discussion of the Teaching Evaluation Results

The first issue that clearly emerges from the survey relates to the degree of accessibility of the evaluation results. It appears that the opportunity to examine the results was reserved for only some of the representatives of certain faculties, and opportunities to openly discuss the results were rare. Only 27.6% of respondents stated that the results of the teaching evaluation were available to the students and/or their representatives; despite the fact that the respondents had been chosen to be those apparently more informed about the topic and the decisions of the faculties and universities, 37.3% of the respondents said that they did not know whether the evaluation results were available (Table 6.3).

Among the factors that appear to favour accessibility to the teaching evaluation results, is whether or not the individual has completed at least one term as

Table 6.3 Availability of information

Are the results of the evaluation available to the students and/or their representatives?	Total %	Experience as repr.		Committee body	
		First term %	Second term or ex-repr. %	Teaching committee %	Other %
Yes	27.6	24.7	32.3	39.4	19.2
No	35.1	34.9	35.5	30.6	38.3
Don't know	37.3	40.4	32.3	30.0	42.5
Total respondents	410	255	155	240	170

representative (second experience or ex-representative) and if he/she is a representative on the teaching committee (TC), which, at least in some universities, is the faculty body charged with reviewing the evaluation results (Table 6.3).

Accessibility to the teaching evaluation results differed between the universities. In some cases the representatives seem to be aware of at least some aspects that emerged from the opinions of attending students, while in others it appears they did not form part of the group to whom the information was made available to. Siena and Padua stand out in this sense; the percentage of representatives who said that the results were available was 43 and 17% respectively, while for Palermo and Pavia it was 21 and 28%. The factors that determine the level of accessibility depend upon choices made (or not made) at the university level and the decisions taken (or not taken) at the faculty level. Communication strategies were set up in some cases (in particular at Siena University) to enable at least a proportion of the students to be reached; in other cases, information channels have not been activated yet or are, at the very best, ineffective (e.g. at Padua University, the results are available on the university website; however, this availability is not known to the majority of students). The availability of the evaluation results only on the website is not an effective strategy to inform the students.

The differences that emerged between the four universities are largely attributable to differences between faculties. At Padua University, the extreme cases are represented by the Law and Veterinary faculties on the one hand, and by the Science faculty on the other, where 0 and 44% of representatives replied respectively, stating that the evaluation results were available. However, at Palermo University, the highest and lowest percentage of students stating that results were available were obtained from the faculties of Economics and Medicine respectively. At Pavia University, the Economics faculty represents the lower extreme, where very few students responded positively (only 1 out of 13), and the Science faculty the higher extreme, with about half the representatives stating that the results were available (12 out of 23). At Siena University, there was greater homogeneity with the extreme cases being the faculty of Science (7 positive answers out of 27) and the faculty of Economics (11 positive responses out of 21).

A more homogeneous behavior emerges with regards to the discussion of the evaluation results at faculty meetings. In almost all cases analyzed, the faculty board and the degree teaching committee ignored the evaluation results, the discussion of

which, according to the statements of almost half of the participating students, was reserved to the teaching committee meetings. It appears, therefore, that the teaching committee was the body assigned to the task of analyzing and sharing the teaching evaluation results in all of the universities surveyed. In fact, considering only the responses from representatives carrying out their second terms as representatives, we see that the results were rarely discussed in the faculty board meetings (7.3% of representatives) and in the degree teaching committee meetings (12.5% of representatives), but much more often in the teaching committee meetings (56%).

It is also interesting to note the level of communication between faculties and student bodies. Nineteen percent of representatives confirmed the deans' statements that a discussion took place at some committees. In this case, the students were more informed and motivated and probably more sensitive to perceiving the effects of the teaching evaluation within their faculties.

6.3.3 Changes Induced by the Results of the Teaching Evaluation

Most respondents claimed that specific measures and changes resulting from attending student opinions were almost non-existent. According to statements from the representatives, the evaluation process that was set up was limited to the gathering of "information" which was then not considered any further by the various groups whose responsibility is to make decisions on teaching activity.

There is a widespread conviction that any changes and measures undertaken that followed teaching evaluation were minimal. According to the majority of the representatives, the teaching evaluation has not led to any changes either at the faculty or the teacher level to date: the mean score given to changes (on a scale of 1–7, where 1 = non-existent and 7 = relevant) regarding structures and equipment, course organization, and the behavior of individual teachers was 2.2 (Table 6.4). It is important to note that the use of numerically scaled data (ordinal data in a restrictive interpretation) as quantitative data has been widely discussed in literature. It is not our intention in this chapter to enter into that debate; nevertheless, it should be emphasized that: (i) the use of means as indicators of synthesis are appropriate for descriptive purposes (as in our case); (ii) the use of models should be considered on a case by case basis with respects to the objectives and (iii) suitable controls should be carried out to confirm any assumptions made by the model (e.g. normality, etc.).

It is interesting to note how the conclusion that the impact of the evaluation was minimal was not common among representatives with more experience and, most probably, more motivation (i.e. representatives who had previously served at least one term). On the other hand, the impact appears to be greater among those participating on teaching committees (Table 6.4). As we have seen in Sect. 6.3.2, these students said that they had participated more than others in the discussion of the results. This would indicate that the analysis of the results based on the opinions of attending students could provide an effective support to the decision-making process.

Table 6.4 Availability of information

		Experience as repr.		Committee body	
	Total	First term	Second term or ex-repr.	Teaching committee	Other
Changes to[a]:	Mean (C_V)[b]	Mean (C_V)[b]	Mean (C_V)[b]	Mean (C_V)[b]	Mean (C_V)[b]
Structures and equipment	2.14 *(0.62)*	2.25 *(0.62)*	1.95 *(0.62)*	2.31 *(0.62)*	2.02 *(0.61)*
Course organization	2.22 *(0.62)*	2.36 *(0.60)*	1.98 *(0.64)*	2.35 *(0.65)*	2.12 *(0.59)*
Behavior of individual teachers	2.19 *(0.64)*	2.30 *(0.65)*	2.01 *(0.60)*	2.32 *(0.64)*	2.10 *(0.63)*
Total respondents	410	255	155	240	170

[a] (1 = non-existent 7 = relevant).
[b] C_V = coefficient of variation.

Limited amounts of change to the specific measures clearly characterize the four universities surveyed. However, the impact upon course organization (schedule, rooms, etc.) by faculty, while limited, does show some variability among faculties: the mean score goes from 1.50 (in the faculty of Pharmacy at Pavia and the faculty of Psychology at Padua) to 3 and above in the faculty of Pharmacy at Siena, the faculty of Political Science at Pavia, and the faculty of Statistical Science at Padua.

6.3.4 The Usefulness of the Teaching Evaluation as Perceived by the Student Representatives

Most of the student representatives appear to have a fairly positive image of the teaching evaluation carried out through student surveys. The opinions gathered on the use of the findings differ, although for the most part they tend to be fairly favorable. In some cases the opinion improves at the end of the interview after the respondent has had a chance to reflect on the various aspects of the evaluation, including negative aspects such as the scant impact and limited availability of the results.

More than half of the students thought that until now the evaluation had not had any important effect upon teaching; however, at the same time they do think that it would be useful to invest resources in gathering student opinions with the view of future potential benefits (Fig. 6.1[1]).

[1] At the start of the interview: "Do you think that gathering opinions from at-tending students is useful?"
At the end of the interview: "Does the gathering of opinions from attending students involve a considerable economic and organizational effort? Do you think that it is useful to use university resources for this activity?"

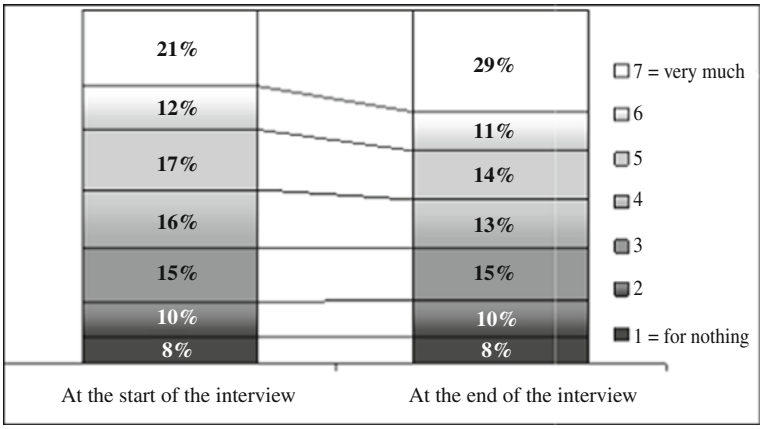


Fig. 6.1 The usefulness of the teaching evaluation, as perceived by the student representatives

Table 6.5 The usefulness of the teaching evaluation as perceived by student representatives

		Experience as repr.		Committee body	
	Total	First term	Second term or ex-repr.	Teaching committee	Other
	Mean (C_V)	Mean (C_V)	Mean (C_V)	Mean (C_V)	Mean (C_V)
Do you think it is useful to use university resources for teaching evaluation?[a]	4.64 *(0.43)*	4.49 *(0.45)*	4.90 *(0.39)*	4.66 *(0.43)*	4.63 *(0.44)*
Total respondents	410	255	155	240	170

[a] (1= not at all 7= very much).
C_V = coefficient of variation

The views from the teaching committee representatives and other university bodies on the use of teaching evaluations are very similar.

The only characteristic that seems to explain some of the variability in opinions is the respondent's experience: representatives in their second term or ex-representatives have substantially more positive opinions than those in their first term (Table 6.5). This is particularly interesting if we consider that, as mentioned above, these representatives have more experience and are also the most critical of the evaluation's impact.

Opinions on the perceived use show a certain variability related to the faculty that the representatives belong to. In 9 faculties, views on the usefulness of the evaluations, as perceived by student representatives, had a mean score above 5; while in 5 faculties the mean score was below 4. The range of variation of the mean score on use varied from 2.69 at Pavia University to 1.46 at Siena University; thus suggesting the determining roles played by the individual faculties in the evaluation procedure.

6.3.5 The Multilevel Regression Model

We estimated two hierarchical linear models with HLM 6.0.

The two models provide an estimate of the effects of the first- and second-level explanatory variables on the "perceived actual" and "potential" usefulness of the teaching evaluation, measured, respectively, by the impact on course organization and the usefulness of gathering student opinions regarding teaching activities[2] (Table 6.6).

All the variables that may have affected teaching-evaluation in the initial assumptions were deliberately retained in order to show that some of these variables do not result as being statistically significant when a multilevel model is used.

Above all, we see a statistically significant effect of "faculty" upon both the "perceived actual" and "potential" usefulness, as is shown by the estimate of the random effects. The two values of the intra-class correlation coefficient (ρ), equal to 3.8 and 4.8%, measure the part of variability that is due to the grouping effect.

The degree of teaching coverage in the survey of student opinions does not appear to explain the variability of the second-level units (the faculties). Even though it does

Table 6.6 Effects upon course organization and the perception of the usefulness of teaching evaluation: estimates of the coefficients of the variables used in the model

	Impact upon course organization	Perception of usefulness term
	Coefficient (S.E.)	Coefficient (S.E.)
Fixed effects		
Intercept	**** 2.22 (0.08)	**** 3.99 (0.29)
"Faculty" Explanatory variable		
Degree pf coverage of the teaching evaluation survey (%)	0.63 (0.38)	−0.68 (0.78)
Explanatory variables regarding the student representatives		
Student representatives on the TC (yes vs. no)	0.01 (0.24)	−0.05 (0.26)
Term (at least one vs. first)	**** −0.41 (0.10)	* 0.41 (0.16)
Availability of the evaluation results (yes vs. no)	−0.08 (0.19)	−0.19 (0.33)
Dean/student representatives agreement about discussion (yes vs. no)	0.32 (0.25)	** 0.77 (0.28)
Random effects		
Variance between faculties τ^2	* 0.071	** 0.194
Variance within faculties σ^2	1.801	3.857
Intra-class correlation coefficient ρ	3.8%	4.8%

* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; **** = $p < 0.0005$

[2] Kolmogorov-Smirnov test has shown that the two dependent variables distribution does not differ significantly from the normal distribution. Also the residual analysis was applied; it has highlighted no violation of regression model assumptions.

not reach a significant level, the coefficient has the opposite sign in the two models: in the case of "perceived actual" usefulness, a rise in the coverage rate corresponds to a rise in the impact upon the organization of courses ($p = 0.10$).

Of the 4 first-level explanatory variables, having served at least one term is the only variable that is statistically significant in both models. It is interesting to note, however, that "potential" usefulness increases by 0.41 points whilst "perceived actual" usefulness falls by 0.41 points when we move from representatives in their first term to those with at least one completed term. There are no statistically significant effects for the other three first-level explanatory variables for "perceived actual" usefulness of teaching evaluation.

Concerning "potential" usefulness, on the other hand, we should note that the variable "agreement" is also statistically significant: when we move from those representatives who do not agree with the dean of the faculty that there has been a discussion on the teaching evaluation results to those who do agree, the score increases by 0.77 points.

6.4 Concluding Remarks

We have tried to evaluate the effects of the teaching evaluation from a student perspective; that is, from the perspective of those who stand to gain the most from an effective evaluation procedure.

The participation of the student representatives in the survey was not entirely satisfactory. On one hand, given the limited resources made available for the study and the difficulties met in contacting the student representatives, the results could be considered as good; on the other hand, assumptions about the population of student representatives must be made cautiously.

Nevertheless, the results are strong enough to support their valid interpretation and the importance of the meta-evaluation undertaken.

In fact, the results of the study bring out several suggestions that could contribute to the evolution of evaluation process within university systems. According to the student representatives, the results of the teaching evaluation presented some weaknesses as well as some strong points. Little impact has emerged regarding the availability of the results, the changes within facilities, the organization of courses and the behavior of teachers. On the other hand, there was a considerable number of favorable opinions about the evaluation tool, in terms of a general consensus on its usefulness for teaching evaluation.

However, it should be noted that in some faculties the availability of teaching evaluation results is closely linked to the representatives' awareness of an objective situation: in this case the different degree of availability between the faculties can in part also be attributable to the various types of behavior and communication strategies adopted by the faculties.

On the other hand, little perception about the changes may be due to various factors: (i) the lack of substantial changes actually being introduced within faculties;

(ii) changes occurring but not noticed (or "perceived") by the representatives; (iii) it is not being realized that the perceived changes were actually introduced due to the results of teaching evaluation.

Considering these three interpretation hypotheses, we see that the low level of impact perceived by the representatives does not exclude the possibility of a greater impact from the teaching evaluation. Once again, in some faculties the apparent lack of student perception could actually reflect a deficit in communication with the representatives, independent of the body that they belong to.

This situation, that in certain respects appears discouraging, nevertheless reveals several factors that can favour a positive impact of the teaching evaluation. The multilevel regression model reveals that the context in which the evaluation procedure is carried out is important: we observe a "faculty" effect upon the impact of teaching evaluation. This supports the hypothesis that students perceive when a faculty is working towards realizing the potential of and returning the results of the teaching evaluation.

Thus, in some cases, the commitment of resources and energy to the survey of student opinions about teaching activities leads to the student representatives' perception of usefulness (theoretical usefulness). Even the impact upon changes to the organization of courses (perceived actual usefulness), although modest, is perceived differently across the faculties. Thus the experiences and evaluative traditions of the individual faculties, their degree of personalization and their commitment to continue with the evaluation procedure appear to play a key role.

Moreover, the length of experience as a representative provides students with the chance to see and perceive the "fruits", although sporadic, of the evaluation procedure. Student representatives in their first term observe more substantial effects upon course organization but perceive less usefulness of the teaching evaluation, while those that have already spent at least one term as representative recognize the usefulness of the evaluation process but have difficulty in perceiving its effects.

The clearest, and perhaps most obvious, indications concern the communication of the evaluation results. When the deans and student representatives agree that the evaluation results have been discussed in the appropriate bodies, then the students appear to recognize the usefulness of the evaluation. The availability of the teaching evaluation results alone does not appear to be a factor that contributes to increasing the level of "theoretical" usefulness.

Once the limits of teaching evaluation have been recognized, the opinions of the representatives reveal a substantial acceptance of the evaluation instrument, while at the same time they highlight the lack of use of this tool. Only if we understand what the purpose of this type of evaluation is and place it, even formally, within a precise organizational and decision-making context, can we benefit from this important university asset. This would entail putting more importance upon the evaluation results and providing feedback opportunities to the students who have produced these results. Now that the evaluation of our universities has been institutionalized, it is time to begin a new journey: towards the use of this instrument [1].

References

1. Campostrini S, Bernardi L, Slanzi D (2008) Le determinanti della didattica attraverso il parere degli studenti. Franco Angeli, Milano
2. Gerzeli S, Parise N, Campostrini S, Magni C, Bernardi L (2008) L'impatto della valutazione della didattica sull'organizzazione universitaria: il parere dei presidi. In: Capursi V, Ghellini G (eds) Dottor DIVAGO. Discernere, Valutare e Governare la nuova Università. Franco Angeli, Milano
3. Goldstein H (2003) Multilevel statistical models, 3rd edn. Edward Arnold, London
4. Lang J (2001) Improving structural policy under conditions of hybrid governance: multi-actor constellations, integrated feedback instruments and the use of evaluation results. Evaluation 7:7–24
5. Snijders T, Bosker R (1999) An introduction to basic and advanced multilevel modelling. Sage, London
6. Yang M, Goldstein H, Browne W, Woodhouse G (2002) Multivariate multilevel analyses of examination results. J R Stat Soc Ser A Stat Soc 165(1):137–153

Chapter 7
University Teaching and Students' Perception: Models of the Evaluation Process

Maria Iannario and Domenico Piccolo

7.1 Introduction

The diffusion of a culture of *evaluation* in the Italian Universities has changed the logic and the development of several activities/procedures. As a consequence, Universities perform periodic surveys in order to assess the students' satisfaction with respect to the main conditions of teaching and the environment where teaching takes place. In addition, several projects and groups have been involved with statistical analyses of University evaluation.

In compliance with procedures established by law, responses are collected among students that attend lectures in a period close to the ending date of courses. This circumstance influences the responses because the students involved with the survey have attended lectures for more than half. Thus, they are aware of problems and tend to give substantially positive answers to the items of the questionnaire. In this context, we think that this large mass of data should be used in effective ways in order to discover useful information with regard to the evaluation process.

This work is organized as follows: in Sect. 7.2 we discuss the concept of students' perception of teaching quality; in Sect. 7.3 we emphasize how the transformation from perception to rating is a complex decision, and thus it calls for adequate statistical approaches. These considerations are deepened in Sect. 7.4 by considering the role of latent variables in the evaluation process and the main logical framework where questionnaires are examined (Item Response Theory). In Sect. 7.5 we introduce a different model for data evaluation that explicitly aims at interpreting the probability distribution of ordinal choices, for each item. The mixture random variable we will discuss about and related generalizations of *CUB* models are briefly illustrated with special reference to students' perception and teaching evaluation. Then, in Sect. 7.6, empirical evidences related to the survey conducted at University of Naples Federico II are presented. Some concluding remarks end the chapter.

M. Iannario (✉)
Dipartimento di Scienze Statistiche, Università di Napoli Federico II, Napoli, Italy
e-mail: maria.iannario@unina.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_7,

7.2 Measurement of Students' Perception About Teaching Quality

Objective of the survey submitted to students is the measurement of the satisfaction of University teaching and related aspects: timing, structures, courses consistency, and so on. These measures are not physical characteristics of some objects but psychological constructs related to respondents. This condition affects both the planning of the experiment and the analysis of data.

In this context, the plan of the experiment consists in a list of several questions (items) submitted to students with regard to relevant issues of the University satisfaction. Most of the items are derived from the standard guidelines [24], and each University specifies/qualifies them on the basis of local requirements.

Since satisfaction is a continuum latent variable, responses to items are based on some ordinal scale (generally, high values are related to high satisfaction). In this way, for each item, respondents are asked to select one of the first m integers related to a Likert scale points. In Italy, several questionnaires are based on a 4-points scale; however, we are strongly convinced that wider ranges for scale ratings are more effective and convenient, even for dynamic comparisons and selective discussions of the results (in any case, we suggest an odd number of alternatives).

In the statistical literature, data analysis of expressed ratings is usually performed by means of several exploratory and inferentially based methods. However, in periodical reports of the "Nuclei di Valutazione" and Councils meetings ("Consigli di Facoltà", "Corsi di Laurea"), simple indicators are presented as common benchmarks for discussion and decisions. As a rule, they are related to the frequency of positive answers and/or the average of quantified responses (sometimes, along with dispersion measures). Most of the critical issues on the evaluation process is currently based on these measures.

We think that indicators without models may cause misunderstanding. Surely, numerical syntheses simplify patterns and complex considerations and allow a large audience to interact with results and assess a final judgement. However, the reduction of large mass of data to just one or two indicators without reference to a generating process may be often misleading, even if some indicators seems illuminating. Everybody knows that average is a correct location measure for a well balanced and unimodal distribution with no extreme data; however, considerable attention must be paid when averages are applied to mere quantifications of qualitative variables without any consideration of the stochastic nature of human decisions.

In this regard, we show in Fig. 7.1 two hypothetical distributions of preferences/rating to a given item expressed on a 9-point scale by two sets of respondents (we refer them as Model 1 and 2, respectively); the average is 6 in both cases but distributions are completely different.[1] For instance, the preferred options (=modal values) are 9 and 6, respectively; moreover, $Pr\,(5 \leq R \leq 7)$ is equal to 0.728 and

[1] Notice that we are joining discrete values of probabilities just for enhancing the different shape of the distributions.

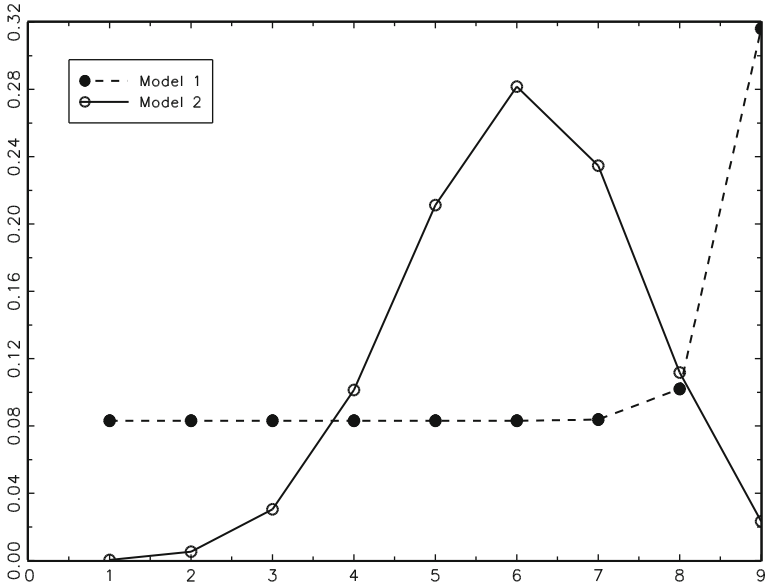


Fig. 7.1 Two hypothetical preference/rating distributions

0.249, respectively. Thus, although the distributions produce the same mean value any comparative decision should be substantially different. This point should not be underestimated since, in these cases, any function based on expectation may hide important information.

Thus, we adhere to the conclusion that "the indicator exists in a model and that the indicator itself is the product of a model" [12] and support the search for adequate measures [20].

In fact, what is really important in studying a complex phenomenon as students' satisfaction is the modelling of the evaluation process that transforms a personal perception into an ordinal answer to a specific item. Thus, a model includes in a consistent way the role and the weight of the real uncertainty that is always pervasive in any decision process. In addition, modelling allows for statistical tests and confidence intervals for any indicator of interest by means of exact, asymptotic or simulated distributions.

7.3 Perception and Rating as Complex Decisions

The perception of an object/service/item is a psychological process by which a subject synthesizes sensory data in forms that are meaningful for his/her conscience. In fact, when we ask a student to answer a specific question on a questionnaire concerning the quality of teaching we are looking for his/her perception of the problem. Then, we are asking to summarize this perception into a well defined category (included in a set of ordinal finite values).

Thus, the expressed evaluation is the final act of complex causes and the answer we collect is affected by the real consideration of the problem and some inherent uncertainty that accompanies human decisions. As a consequence, any expressed perception becomes the realization of a stochastic phenomenon and it should be analyzed with statistical methods that rely on the possibility to investigate the generating data process.

Actually, psychological processes when faced with discrete choices manifest themselves by two main factors that explain the final decision:

- a *primary* component, generated by the sound impression of the respondent, related to awareness and full understanding of problems, personal or previous experience, group partnership, and so on;
- a *secondary* component, generated by the intrinsic uncertainty of the final choice. This may be due to the amount of time devoted to the answer, the use of limited set of information, partial understanding of the item, laziness, apathy, and so on.

From the point of view of the *interviewer*, the first component is hopefully the most important in determining the answer in order to gain information on the real motivations that generated the observed result. Instead, from the point of view of the *interviewee*, the second component may become considerable if he/she is not really involved/interested to give a meditated answer.

Moreover, by constraining the choice process into an ordinal finite set of alternatives, we produce a hierarchical procedure since respondents first orient themselves in a coarse evaluation (negative, indifferent, positive) and then refine their final judgement.

Actually, empirical evidence shows that extreme choices are assessed in a sharper way. On the contrary, when the number of alternatives increases people tend to be not so extreme even if they are really satisfied with item.

Finally, it is important to realize that specific circumstances may increase the observed evaluations in some classes. This happens, for instance, when some category is expressed in a way that induces to simplify more elaborate decisions (we call them *shelter choices* [45]).

7.4 Latent Variables and Item Response Theory

Since satisfaction and perceived quality are not observable, some remarks are necessary in order to define their role in the modelling approaches we will speak about. Surely, few latent traits (constructs, variables, factors) are common features that drive the general pattern of responses to a questionnaire aimed at evaluating a service [9–11, 36]. Empirical evidence confirms that similarities, differences, contrasts among the responses are quite common. Thus, although a huge amount of hypothetical patterns could be conjectured, only a limited subgroup of them are observed in a significant frequency.

Latent variables force the evaluation process and statistical researches should focus on substantive models for explaining observed patterns within a consistent rational framework. In the literature, several independent approaches bring to similar modelling. Among them, we quote those originated by psychophysical and sensorial studies: [17, 51, 52, 78]. Similarly, starting from [57], econometricians refer to "Random Utility Models" (RUM): [2, 28, 39–41, 79]. Then, in the vein of "Unobservable Variable approach" (UVA), several statisticians have been involved with the introduction of useful models for evaluation data: [18, 23, 48, 60–62, 65]. For an updated survey, see: [16].

A common feature of these approaches is that the answers to items are supposed to be generated by a latent variable that explains the dependency and manifests the most important characteristics (features, constructs, traits) of the survey.

Models related to *Item Response Theory* (IRT) are diffuse in psychological and medical studies, marketing and political researches, and different motivations, usages and notations may obscure a common framework. In fact, some papers are aimed at defining a recognized taxonomy: [75, 77]. Main distinctions are based on: dichotomous or politomous responses, number of parameters, ordinal nature of the responses, availability of covariates, number of latent traits, and so on.

From an historical point of view, IRT has been generated as a critical reaction to classical test theory [49], where people assume that responses to several items are numerous enough to apply standard methods to the total score. Main critical issues are the non independence of the items and the existence of common patterns in the responses. In addition, when a set of items in educational contexts are submitted, responses are function both of *ability of respondents* and *difficulty of items*. This is a problem that classical test theory does not tackle in a simple and effective way.

Thus, IRT is based on several assumptions, and the more important are the following:

- *Unidimensionality*. Theory assumes that questionnaires are measuring a continuous latent variable defined on the real line.
- *Local Independence*. Theory assumes that any relationship among items is fully explained by few common latent variables; thus, for a given trait level, item responses are independent.
- *Normality*. Often, latent variables are assumed to be Normally distributed.

In this approach, the starting point for new directions has been the introduction of Rasch model [71], with just one parameter, later generalized by Birnbaum – in a series of reports from 1957 onwards, reported in [51] – and Lord [50] who considered two and three parameters model, respectively. Rasch originally proposed a model for dichotomous responses. Instead, Bock [15] considered politomous values where the probability of a category is proportional to the sum of all others. Specifically, the non-negativity of probabilities suggests the ratio of exponentials; further additional constraints on the parameters are required to ensure identifiability.

To establish notation, we assume that R_{ik} is the response random variable of the i-th respondent to the k-th item, for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$. Then,

sample information is contained in the following $(n \times K)$ matrix, consisting of the observed answers of n respondents to K items:

$$\begin{bmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,j} & \dots & r_{1,K} \\ r_{2,1} & r_{2,2} & \dots & r_{2,j} & \dots & r_{2,K} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{n,1} & r_{n,2} & \dots & r_{n,j} & \dots & r_{n,K} \end{bmatrix}.$$

Observed values are the expressed ratings of an evaluation process. They could be $0, 1$ for dichotomous situations (as it happens for tests) or included in $\{1, 2, \dots, m\}$, $m > 2$ for politomous cases (as it happens in evaluation and preference surveys[2]).

Each row is the response pattern of a given subject and each column represents the observed evaluation to a given item expressed by different subjects. It seems evident that information deduced by rows should be related to *subjects' ability* while information derived by columns should be related to *items' difficulty*. This interpretation has historically generated the whole family of Rasch models, as shown by [38]. More specifically, George Rasch searched for an item response function such that it implied a complete separation among the *person's ability* and the *items' difficulty* parameters. This requirement has been called *specific objectivity* and it related to the joint sufficiency property of the parameters estimators [73].

For a single item that admits a dichotomous solution (correct: $R = 1$, or wrong: $R = 0$), the standard formulation[3] of the original Rasch model for the j-th item:

$$Pr\left(R_j = 1 \mid \theta\right) = \frac{1}{1 + e^{-a\,(\theta - \delta_j)}},$$

expresses the probability of a correct answer for a given person's ability θ with respect to the item difficulty parameter (δ_j) and the discrimination parameter (a).

A generalization of this formulation leads to a three parameter (logistic) Rasch model defined, for each item $j = 1, 2, \dots, K$, by:

$$Pr\left(R_j = 1 \mid \theta\right) = c_j + \frac{1 - c_j}{1 + e^{-a_j\,(\theta - \delta_j)}}.$$

Here, we are assuming that there is a probability c_j to guess the correct response to the j-th item even if respondents do not know it; in addition, we allow *discrimination parameters* a_j to vary among the items.

[2] We are simplifying the analysis to the case where each items is supposed to have a constant number of answers. For a more general discussion, see: [16].

[3] This formulation maps a non-negative function into the range [0, 1] and introduces the need for a logistic function in a simple manner. In several papers, the negative exponent is set to -1.7 (instead of -1) since this adjustment solves in a better approximation of logistic to Normal density.

When responses are ordinal (as it is common in the evaluation context), Samejima [74] proposed a *graded model* which expresses the probability that a response will be observed in a specific category or above. Similarly, *partial credit* model proposed by Masters [54] assumes that the discrimination parameter does not vary among items and that the probability of scoring a given category over the previous one is a function of parameters. Finally, a *rating scale model* proposed by Andrich [4] introduces a further parameter with respect to the partial credit model to locate the item position on the underlying construct (but it is constrained to the same number of categories among items). These models are related to proportional odd models, adjacent categories logit models and continuation ratio models for ordinal data, proposed in the Generalized Linear Models framework, as derived by McCullagh [55] and discussed by [1, 35].

In the context of students' evaluation [3], the ability assessment has been transformed into a quality assessment. Then, the ability (subjects) and difficulty (items) parameters are now transformed into *satisfaction* (subjects) and *quality* (items), respectively. This approach has been fully discussed in several papers by [29–31] with regard to evaluation and customer satisfaction data; they prefer the *extended logistic model*, that generalizes the rating scale model, as proposed by [5, 6].

We defer to the vast literature[4] for further considerations about these and related problems (for instance, multilevel, hierarchical, multidimensional, mixture and nonparametric IRT models: [47, 72, 76, 80]). We only quote here that the inclusion of subjects' characteristics as explanatory variables of latent traits are examined by IRT researchers by means of "Differential Item Functioning" (DIF). This representation is useful for showing evidence of significant covariates in subgroups; often, the presence of clusters is considered as a bias in the responses expressed by a limited number of subjects as in [64].

7.5 An Alternative Model for the Evaluation Process

We introduce a different paradigm in order to explain ordinal choices that people routinely perform when faced with the evaluation process. The model that we will introduce is parsimonious and flexible with respect to alternative distributions [66]. In this case, the reference to latent traits is again valid but the probability of ordinal values is explicitly estimated and checked by data. In addition, it is immediate to add subjects' covariates (even of continuous nature) for taking the behaviour of the respondents into account. Finally, clustering evaluation data by means of estimated models turns out to be efficient and selective, as shown by Corduas [25–27].

[4] Un updated account of several methods, models and procedures in the IRT framework is contained in [70].

7.5.1 Rationale for CUB Models

In our model, rating is interpreted as the final outcome of a psychological process, where the investigated trait is intrinsically continuous but it is expressed in a discrete way on a given scale. Then, it is possible to quantify the impact of individual covariates on the perception of the main aspects of University teaching, and to study how perception changes with students' profiles.

The rationale for *CUB* models[5] stems from the interpretation of final choices of respondents as weighted combinations of a personal *agreement* (*feeling*) and some intrinsic *uncertainty* (*fuzziness*).

The first component is parameterized by a shifted Binomial random variable which is able to map a continuous latent variable (with unimodal distribution: Normal, Student t, logistic, etc.) into a discrete set of values $\{1, 2, \ldots, m\}$. Its shape depends on the cutpoints we assume for the latent variable.

The second component is a discrete Uniform random variable and describes the inherent uncertainty of an evaluation process constrained to be expressed by discrete choices. Actually, it is a building block for modelling the *propensity* of a respondent towards the extreme solution of a totally indifferent choice.

Although a mixture distribution may be interpreted as a two steps stochastic choice between two discrete distributions, we are not saying that population is composed of two subgroups (respondents whose choice is without and with uncertainty, respectively). Instead, we are assuming that each subject acts *as if* his/her final choice would be generated with *propensities* (π) and $(1-\pi)$ to belong to one of the two distributions, respectively. In this regard, we observe that $(1-\xi)$ is a measure of agreement/feeling towards the item and $(1-\pi)$ is a measure of the uncertainty that accompanies the choice.

7.5.2 CUB Models

On a more formal basis, for a given $m > 3$, we consider the expressed rating r as a realization of a random variable R, with probability distribution given by:

$$Pr(R = r) = \pi \binom{m-1}{r-1} \xi^{m-r} (1-\xi)^{r-1} + (1-\pi)\frac{1}{m}, \quad r = 1, 2, \ldots, m.$$

The model, firstly introduced by [33, 66], is fully specified by the parameters $\pi \in (0, 1]$ and $\xi \in [0, 1]$ that are inversely related to the weight of uncertainty and feeling, respectively. Its identifiability has been proved by Iannario [43].

Later, Piccolo [67] generalized this mixture random variable by introducing logistic links between the model parameters and the subjects' and objects' covari-

[5] The acronym *CUB* derives from the circumstance that in these models we introduce Covariates in a mixture of Uniform and shifted Binomial random variables.

ates (as applied in [68]). This class has been called $CUB(p, q)$ models, depending on the numbers of $p \geq 0$ and/or $q \geq 0$ parameters related to covariates for π and ξ parameters, respectively. Of course, a $CUB(0, 0)$ is just a probability mixture distribution for the ratings, that is a model without covariates.

In this way, the class of $CUB(p, q)$ models is generated by two components:

- a *stochastic component*:

$$Pr\left(R = r \mid \boldsymbol{y}_i;\ \boldsymbol{w}_i\right) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} + (1-\pi_i)\left(\frac{1}{m}\right);$$

  for $r = 1, 2, \ldots, m$, where the parameters π_i and ξ_i, for any i-th subject, $i = 1, 2, \ldots, n$, are defined by:
- a *systematic component*:

$$\begin{cases} \pi_i = \dfrac{1}{1 + e^{-\boldsymbol{y}_i \boldsymbol{\beta}}}; \\ \xi_i = \dfrac{1}{1 + e^{-\boldsymbol{w}_i \boldsymbol{\gamma}}}. \end{cases}$$

Here, $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ are the i-th subjects' covariates, for explaining π_i and ξ_i, respectively.

Finally, a $CUB(p, q)$ model with a logistic link is defined, for any $i = 1, 2, \ldots, n$, by:

$$Pr\left(R = r_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \frac{1}{1 + e^{-\boldsymbol{y}_i \boldsymbol{\beta}}} \left[\binom{m-1}{r_i - 1} \frac{\left(e^{-\boldsymbol{w}_i \boldsymbol{\gamma}}\right)^{r_i - 1}}{\left(1 + e^{-\boldsymbol{w}_i \boldsymbol{\gamma}}\right)^{m-1}} - \frac{1}{m} \right] + \frac{1}{m}.$$

A random sample consists of the joint set of expressed evaluations and covariates $(r_i,\ \boldsymbol{y}_i,\ \boldsymbol{w}_i)'$, for $i = 1, 2, \ldots, n$, and this information (for moderate and large size) is sufficient to generate sensible inference on the parameters $(\pi,\ \xi)'$ via the log-likelihood function and related asymptotic results.[6]

Notice that *CUB* models adhere to the logic of the Generalized Linear Models (GLM), advocated by [56, 63], since they introduce linear functions of covariates for improving inference on observed data. However, they do not belong to GLM class since the chosen mixture distribution is not in the exponential family and a link among expectations and parameters is not required. In fact, our models are included in a more general framework [53].

[6] For more technical discussions about statistical issues arising from the inference on $CUB(p, q)$ models, see: [67, 69]. Successful applications of *CUB* models are now available in several different fields: [7, 8, 19, 21, 42, 44, 46, 68].

Furthermore, we quote the *extended CUB models* proposed by [45] who are able to take into account the possible presence of atypical frequency distributions[7] generated by subgroups that select a specific category (*shelter choice*). This kind of problem is relevant also in educational context: for instance, with reference to the evaluation survey to be discussed in Sect. 7.7, we found that the adjective *satisfied*, positioned just after an indifference option, caused everywhere a sensible *shelter effect* in the responses given by students, with a high impact on the frequency distribution ranging from 10 to 30%. Then, by using extended *CUB* model, this effect has been explained with a substantial improvement of the fitting measures, from 2 up to 10 times.

7.6 Empirical Evidences for University Teaching Evaluation

Several statistical methods and empirical evidences have been derived from the evaluation of University teaching in Italy. They are based on principles and foundations [13, 14] as well as on methods and applications [22, 23, 37, 58, 59, 65]. Moreover, the approach proposed in the previous section has been pursued with several evaluation data set [32, 34].

In this chapter, we present some results related to the evaluation of students' satisfaction at University of Naples Federico II, based on data collected during the academic year 2005/2006 (the sample concerns $n = 34{,}507$ validated questionnaires), and we limit ourselves to discuss only few features. Unfortunately, more specific considerations cannot be derived since data base was explicitly delivered by University offices with the constraint of non-identifiability of Faculties (as a consequence of privacy rules).

In this survey, the perceived feeling/satisfaction to different items (quality of lecture halls, objectives and adequacy of courses, instructors' ability and availability, time-table respect, and so on) has been rated from 1 = *completely unsatisfied* to 7 = *completely satisfied*. Thus, in the first subsection we examine $CUB(0, 0)$ models for these ratings with respect to some elements of stratifications: Faculty, gender and attendance. Moreover, in the second subsection, we will estimate *CUB* models with covariates in order to show how the global satisfaction rating is related to significant subjects' covariates.

7.6.1 CUB Models Without Covariates

In Fig. 7.2, we present the estimated *CUB* models with reference to responses given to the global evaluation item. All Faculties are characterized by a low uncertainty

[7] Actually, this extended structure generalizes the class of *CUB* models since it allows the (extreme) possibility to fit the (degenerate) situation of all data collapsing at an intermediate category $R \neq 1, m$.

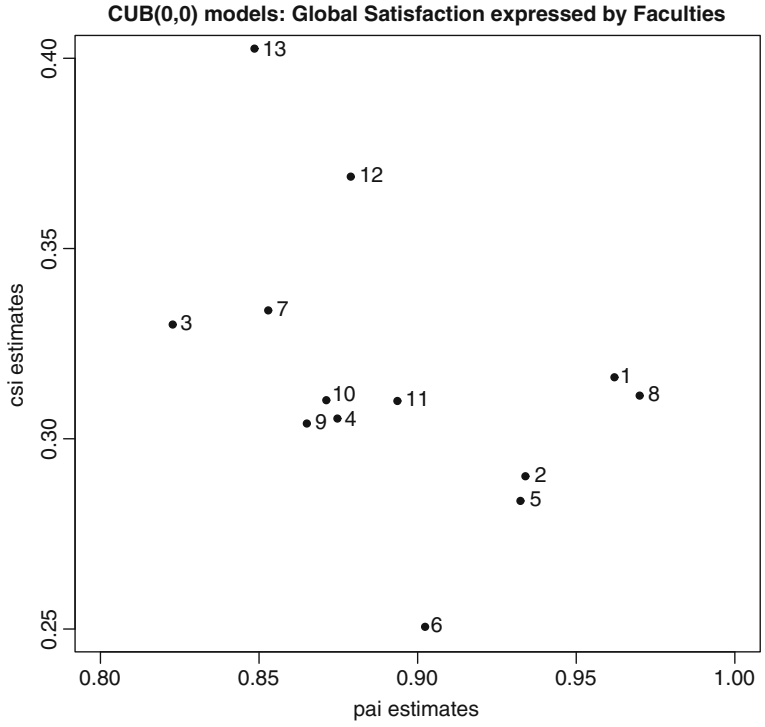


Fig. 7.2 *CUB* models for global satisfaction of 13 Faculties

(estimated π imply uncertainty shares always less than 3%) but the level of positive evaluation is more unstable (since estimated ξ vary from 0.25 to 0.40).

For a better understanding of this global assessment, Fig. 7.3 presents the location of estimated models in the parametric space for the items concerning the evaluation of lecture halls, quality of teaching and global satisfaction. It seems evident how the last issue is related to (and almost confused with) the expressed judgement towards teaching. Anyway, responses related to global and teaching evaluations are less uncertain and manifest a more positive feeling with respect to lecture halls evaluations.

Respondents are mostly women (55%) and different profiles arise when we consider the estimated models for various items with respect to genders, as confirmed by Fig. 7.4. For both genders we observe a common patterns of models on the parameter space: there is a difference among items related to organization and structure of courses and items related to personal relationship with the instructors. We register better judgments of instructors expressed with low uncertainty, whereas we see lower and more definite judgments towards structural components. However, women are more resolute about their evaluations in a sensible measure.

Expressed results are clearly related to the typology of respondents since the sample consists of students with a generally high attendance: more than 76% declared to attend lectures for more than 80% of the term and only 0.7% of them for less

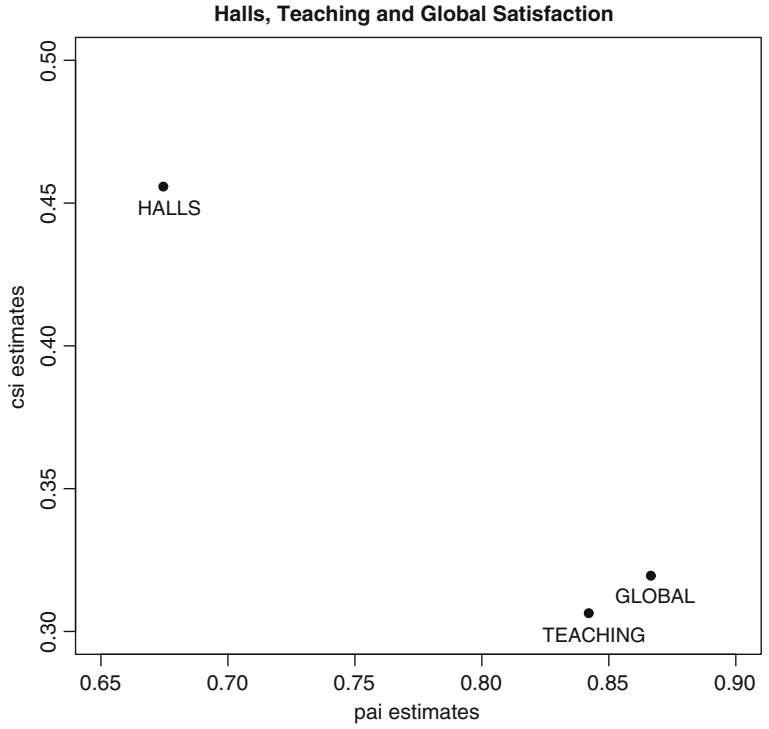


Fig. 7.3 *CUB* models and halls, teaching and global evaluations

than 20%. In fact, judgments are quite similar if *CUB* models are estimated for subgroups of students characterized by a given attendance rate; thus, remarkable differences of estimated models are found only for students with occasional attendance (Fig. 7.5). As a matter of fact, low attendance reduces positive evaluation and increases uncertainty in the responses.[8]

7.6.2 CUB Models with Covariates

Taking into account the results of previous subsection, we look for models which explains the expressed evaluation as a function of selected covariates. We limit ourselves to a large Faculty for which a considerable number ($n = 10{,}572$) of validated questionnaires is available and we study the behaviour of respondents with reference to the global evaluation item.

We found that positive evaluation is significantly related to *Attendance* (expressed by four ordinal classes), *Age* (in years) and the number of passed *Exams*

[8] We observe that active participation to the University life and, specifically, attendance to courses are often related to a possible job for a student. However, we were not able to discriminate subgroups of respondents in correspondence with the nature of their job.

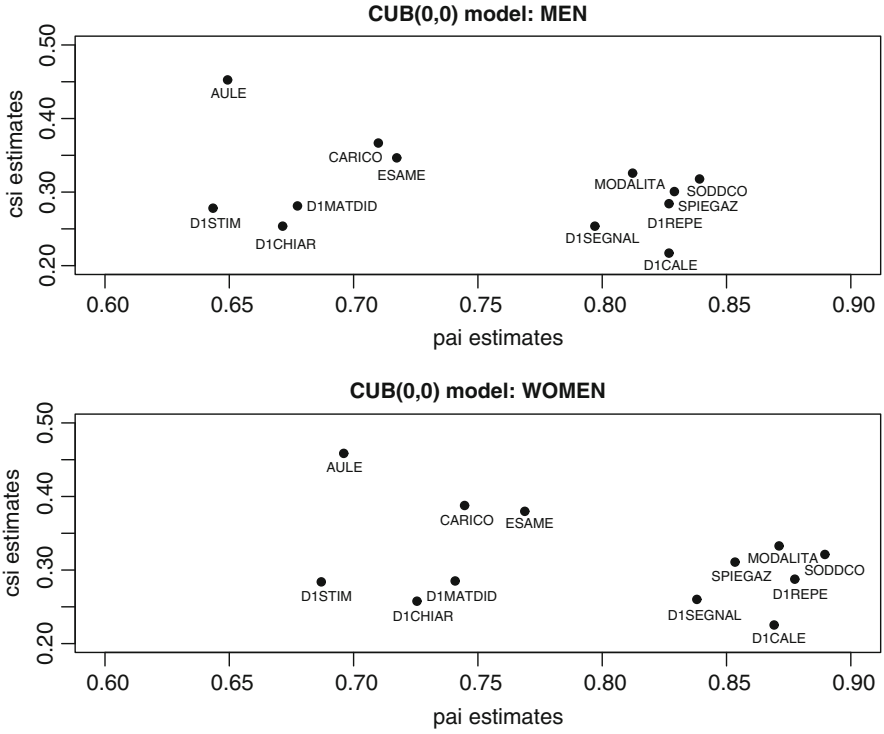


Fig. 7.4 *CUB* models and gender of respondents

of respondents. Table 7.1 presents estimated *CUB* models of increasing complexity with respect to the numbers of significant covariates (which enter in the model in decreasing order of significance). Notice that all parameters are significant, although the effect of the last covariate (=*Exams*) in the last estimated model seems feeble.

Table 7.2 summarizes these results by showing the asymptotic tests for the previous models. Observed tests should be compared with the critical values of a χ^2 random variable, which for a level $\alpha = 0.05$ and degrees of freedom $g = 1, 2, 3$, are given by: $\chi^2_{(1)} = 3.841$; $\chi^2_{(2)} = 5.991$; $\chi^2_{(3)} = 7.815$, respectively.

Specifically, given covariates $\boldsymbol{w}_i = (Attendance_i,\ Age_i,\ Exams_i)'$ for the i-th respondent and $m = 7$, the estimated $CUB(0, 3)$ models implies that:

$$Pr(R = r \mid \boldsymbol{w}_i) = 0.024 + 0.831 \binom{6}{r-1} (1-\xi_i)^{r-1} \xi_i^{7-r}, \quad r = 1, 2, \ldots, 7,$$

where the $\xi_i = (\xi_i \mid \boldsymbol{w}_i)$ parameters, for $(i = 1, 2, \ldots, n)$ are specified by:

$$\xi_i = \frac{1}{1 + \exp\{-2.028 - 0.253\, Attendance_i + 0.973 \log(Age_i) + 0.005\, Exams_i\}}$$
$$= \frac{1}{1 + 0.132\, Age_i^{0.973}\ 0.776^{Attendance_i}\ 1.005^{Exams_i}}.$$

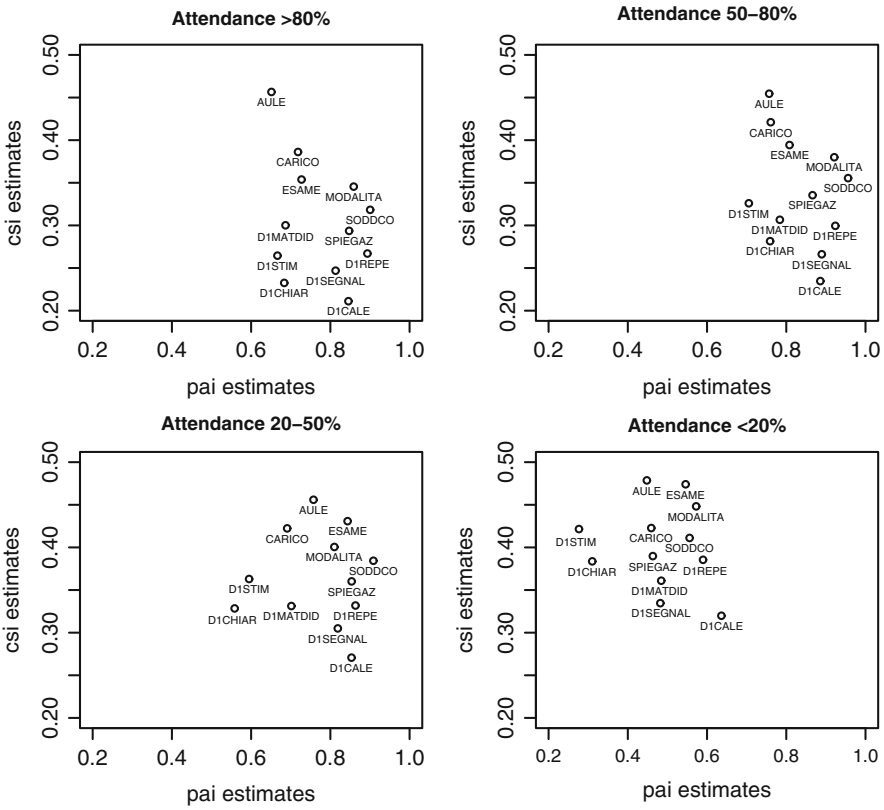


Fig. 7.5 *CUB* models and lectures attendance of respondents

Table 7.1 Estimated $CUB(p, q)$ models for expressed global evaluation

Models	$\hat{\pi}$	$\hat{\xi}(\boldsymbol{w})$	Log-likelihood
▶ *CUB(0,0)*	$\hat{\pi} = 0.821$ *(0.008)*	$\hat{\xi} = 0.331$ *(0.002)*	$\ell_{00} = -17689$
▶ *CUB(0,1)*	$\hat{\pi} = 0.826$ *(0.008)*	$\hat{\gamma}_0 = -0.985$ *(0.031)*	$\ell_{01} = -17639$
Attendance		$\hat{\gamma}_1 = 0.241$ *(0.024)*	
▶ *CUB(0,2)*	$\hat{\pi} = 0.831$ *(0.008)*	$\hat{\gamma}_0 = 2.330$ *(0.348)*	$\ell_{02} = -17593$
Attendance		$\hat{\gamma}_1 = 0.258$ *(0.023)*	
ln(Age)		$\hat{\gamma}_2 = -1.092$ *(0.115)*	
▶ *CUB(0,3)*	$\hat{\pi} = 0.831$ *(0.008)*	$\hat{\gamma}_0 = 2.028$ *(0.356)*	$\ell_{03} = -17587$
Attendance		$\hat{\gamma}_1 = 0.253$ *(0.023)*	
ln(Age)		$\hat{\gamma}_2 = -0.973$ *(0.119)*	
Exams		$\hat{\gamma}_3 = -0.005$ *(0.001)*	

(*Standard errors in parantheses*).

Table 7.2 Asymptotic tests for the estimated $CUB(p, q)$ models

Model comparisons	Deviances difference	g
$CUB(0, 1)$ versus $CUB(0, 0)$	$2\ (\ell_{01} - \ell_{00}) = \ 99.01$	1
$CUB(0, 2)$ versus $CUB(0, 0)$	$2\ (\ell_{02} - \ell_{00}) = 191.63$	2
$CUB(0, 3)$ versus $CUB(0, 0)$	$2\ (\ell_{03} - \ell_{00}) = 203.53$	3
$CUB(0, 2)$ versus $CUB(0, 1)$	$2\ (\ell_{02} - \ell_{01}) = \ 92.62$	1
$CUB(0, 3)$ versus $CUB(0, 2)$	$2\ (\ell_{03} - \ell_{02}) = \ 11.91$	1

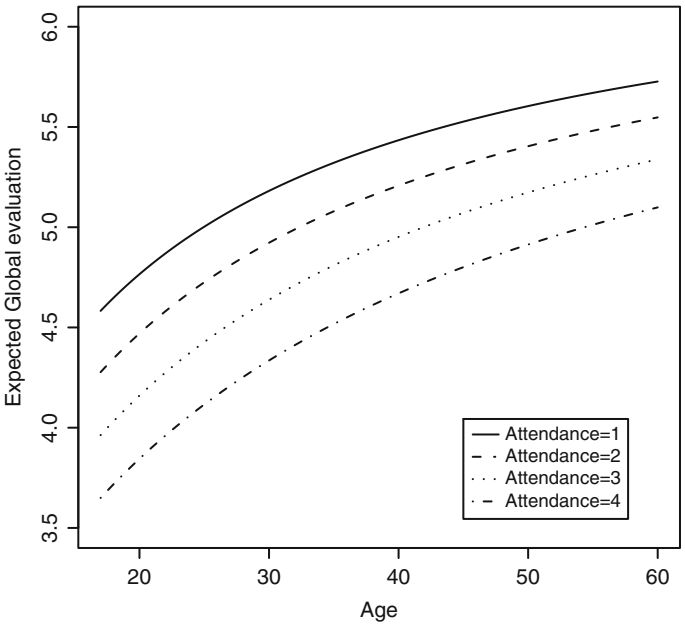


Fig. 7.6 Expected global evaluation as a function of covariates

Given that ξ is an inverse measure of satisfaction, the expressed global evaluation increases with *Age* and also it remarkably raises with the *Attendance* rate; in a lower extent, it increases also with the number of passed *Exams*. For instance, when $Exams = 0$, these results are well summarized in Fig. 7.6 where the expected global evaluation, according to the estimated $CUB(0, 3)$ model, is plotted as a function of selected covariates.

7.7 Concluding Remarks

The proposed *CUB* models are characterized by a sensible fitting performance achieved with few parameters (parsimony) and an immediate possibility to interpret results in terms of evaluation features and uncertainty, as well as by means of covariates.

In this area, we are currently looking for adequate formalizations that allow to integrate the *CUB* models approach in the latent trait environment, in order to gain the multivariate dimension of rating data analysis. Similarly, multilevel considerations for *CUB* models should be necessarily introduced for improving the interpretation of real situations where clusters of respondents generate similar evaluations.

Anyway, this kind of analysis (both conceptually and empirically based) have convinced us that some operational implications may be suggested to institutions charged to make more effective the evaluation process of University teaching. These considerations may help to generate few, simple and useful rules:

- Questionnaires must be largely simplified since all analyses confirm that results are based upon just one or two latent variables. These constructs concern the satisfaction towards a *personal* component (ability of teacher judged in a favorable way, even if structures are not adequate) and/or the criticism towards a *structural* component of courses (rooms, times, availability and adequacy of laboratories, often negatively judged even if teaching is positively evaluated). Thus, few questions may effectively capture most of data information without wasting time and/or lowering the accuracy of responses.
- Sample size may be reduced in favour of stratified procedure in order to achieve more accurate answers. In fact, collection of data among students in a classroom induces internal correlation, high dispersion of respondents and a large amount of useless questionnaires. Of course, any selection mechanism must respect the requirement that surveyed students attend lectures and courses to be evaluated.
- Simple outputs for intermediate and final users should be based on effective indicators that are related to fitted models (reported with goodness of fit measures) which are derived from hypotheses on the generating mechanism of data.

The final message is that evaluation of University teaching is both a complex task and a difficult challenge for statisticians. As in other fields, knowledge should be based on sound theory and extensive experience that lead to iterative and interactive processes. In any case, simple and effective models should be encouraged for supporting correct decisions.

Acknowledgements This chapter has been presented at the Conference: "DIVAGO: la Statistica, la Valutazione e l'Università", University of Palermo, 10–12 July 2008. We thank discussants and referees whose constructive considerations improved the final version. The research has been realized within the 2006 PRIN-MIUR project: "Stima e verifica di modelli statistici per l'analisi della soddisfazione degli studenti universitari" and with the scientific support of CFEPSR, Portici. Authors thank University of Naples Federico II, and especially the "Nucleo di Valutazione di Ateneo" and UPSV, for providing the data set which has been partly analyzed in this chapter. This is a joint work; however, M. Iannario wrote Sects. 7.4, 7.5 and 7.6 and D. Piccolo the others.

References

1. Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York, NY
2. Amemiya T (1981) Qualitative response models: a survey. J Econ Lit XIX:1483–1536

3. Aiello, F. and Capursi, V. (2008), Using the Rasch model to assess a university service on the basis of student opinions. Applied Stochastic Models in Business and Industry, 24:459–470. doi: 10.1002/asmb.730
4. Andrich D (1978) A rating formulation for ordered response categories. Psychometrika 43:561–573
5. Andrich D (1985) An elaboration of Guttman scaling with Rasch models for measurement. In: Tuma NB (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 33–80
6. Andrich D (1988) A general form of Rasch's extended logistic model for partial credit scoring. Appl Meas Educ I:363–378
7. Balirano G, Corduas M (2006) Statistical methods for the linguistic analysis of a humorous TV sketch show. Quaderni di Statistica 8:101–124
8. Balirano G, Corduas M (2008) Detecting semeiotically-expressed humor in diasporic TV productions. HUMOR. Int J Humor Res 21:227–251
9. Bartholomew DJ (1980) Factor analysis for categorical data. J R Stat Soc Ser B 42:293–321
10. Bartholomew DJ (1987) Latent variable models and factor analysis. M. Dekker, New York, NY
11. Bartholomew DJ, Knott M (1999) Latent variable and factor analysis, 2nd edn. Kendall's Library of statistics, vol 7. Arnold, London
12. Bernardi L, Capursi V, Librizzi L (2004) Measurement awareness: the use of indicators between expectations and opportunities. In: Proceedings of XLII SIS meeting, Cleup, Padova, pp 315–326
13. Biggeri L (2000) Valutazione: idee, esperienze, problemi. Una sfida per gli statistici. In: Proceedings of XL SIS meeting, CS2p, Firenze, pp 31–48
14. Biggeri L, Bini M (2001) Evaluation at University and State level in Italy: need for a system of evaluation and indicators. Tertiary Educ Manage 7:149–162
15. Bock RD (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37:29–51
16. Bock RD, Moustaki I (2007) Item response theory in a general framework. In: Rao CR, Sinharay S, (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 469–513
17. Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs. I. Method of paired comparisons. Biometrika 39:324–345
18. Cagnone S, Gardini A, Mignani S (2004). New developments of latent variable models with ordinal data. Atti della XLII Riunione Scientifica SIS, Bari, I:1–12
19. Cappelli C, D'Elia A (2004) La percezione della sinonimia: un'analisi statistica mediante modelli per ranghi. In: Prunelle G, Fairon C, Dister A (eds) Le poids des mots - Actes de JADT2004, Presses Universitaires de Louvain, Belgium, pp 229–240
20. Capursi V, Porcu M (2001) La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni di giudizi. In: Proceedings of SIS meeting on: "Processi e Metodi Statistici di Valutazione", Roma, pp 17–20
21. Cerchiello P, Iannario M, Piccolo D (2010) Assessing risk perception by means of ordinal models, in: Perna C. et al. editors, Mathematical and Statistical Methods for Insurance and Finance, Springer, New York, pp 65–73
22. Chiandotto B, Bertaccini B, Bini M (2007) Evaluating the quality of the University educational process: an application of the ECSI model. In: Fabris L (2006) Effectiveness of university education in Italy: employability, competences, human capital. Springer, Heidelberg
23. Chiandotto B, Bertaccini B (2008) SIS-ValDidat: a statistical information system for evaluating university teaching. Quaderni di Statistica 10:157–176
24. CNVSU (2002) Proposta di un insieme minimo di domande per la valutazione della didattica da parte degli studenti frequentanti, Comitato Nazionale per la Valutazione del Sistema Universitario. MIUR Doc. 9/02, http://www.cnsvu.it
25. Corduas M (2008a) A testing procedure for clustering ordinal data by CUB models. In: Proceedings of Joint SFC-CLADAG 2008 meeting, ESI, Napoli, pp 245–248

26. Corduas M (2008b) A study on University students' opinions about teaching quality: a model based approach for clustering ordinal data. DIVAGO meeting proceedings, University of Palermo 10–12 July 2008, this book
27. Corduas M (2008c) A statistical procedure for clustering ordinal data. Quaderni di Statistica 10:177–189
28. Cramer JS (2001) An introduction to the logit model for economists, 2nd edn. Timberlake Consultants Ltd., London
29. De Battisti F, Nicolini G, Salini S (2005). The Rasch model to measure service quality. J Serv Mark III:58–80
30. De Battisti F, Nicolini G, Salini S (2008). Methodological overview of Rasch model and application in customer satisfaction survey data. Dipartimento di Scienze Economiche, Aziendali e Statistiche, University of Milan, Working paper n.2008-04
31. De Battisti F, Nicolini G, Salini S (2010). The Rasch model in customer satisfaction survey. Qual Technol Quant Manage 7(1):15–34
32. D'Elia A, Piccolo D (2002) Problemi e metodi statistici nei processi di valutazione della didattica. Atti della Giornata di Studio su "Valutazione della Didattica e dei Servizi nel Sistema Universitario", Università di Salerno, Fisciano, pp 105–127
33. D'Elia A, Piccolo D (2005) A mixture model for preference data analysis. Comput Stat Data Anal 49:917–934
34. D'Elia A, Piccolo D (2006). Analyzing evaluation data: modelling and testing for homogeneity. In:Zani S, Cerioli A, Riani M, Vichi M (eds) Data analysis, classification and the forward search. Springer, Berlin, pp 299–307
35. Dobson AJ, Barnett AG (2008) An introduction to generalized linear models, 3rd edn. Chapman & Hall/CRC, Boca Raton, FL
36. Everitt BS (1984) An introduction to latent variable models. Chapman & Hall, New York, NY
37. Fabbris L (ed) (2006) Effectiveness of university education in Italy: Employability, competences, human capital. Springer, Heidelberg
38. Fischer GH (2007) Rasch models. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 515–585
39. Franses PH, Paap R (2001) Quantitative models in marketing research. Cambridge University Press, Cambridge
40. Greene WH (2000) Econometric analysis, 4th edn. Prentice Hall International, Inc., Englewood Cliffs, NJ
41. Hensher DA, Rose JM, Greene WH (2005) Applied choice analysis. A primer. Cambridge University Press, Cambridge
42. Iannario M (2007) A statistical approach for modelling Urban Audit Perception Surveys. Quaderni di Statistica 9:149–172
43. Iannario M (2010) On the identifiability of a mixture model for ordinal data, METRON, LXVIII:87–94
44. Iannario M (2008b) A class of models for ordinal variables with covariates effects. Quaderni di Statistica 10:53–72
45. Iannario M (2010) Modelling *shelter* choices in ordinal surveys, submitted
46. Iannario M, Piccolo D (2010) A new statistical model for the analysis of customer satisfaction. Qual Technol Quant Manage 7(2):149–168
47. Johnson MS, Sinharay S, Bradlow ET (2007) Hierarchical item response theory models. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Boston: Elsevier, pp 587–606
48. Jöreskog KG, Moustaki I (2001) Factor analysis of ordinal variables: a comparison of three approaches. Multivariate Behav Res 36:347–387
49. Lewis C (2007) Selected topics in classical test theory. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Boston: Elsevier, pp 29–43
50. Lord FM (1980) Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates, Hillsdale, NJ

51. Lord FM, Novick MR (1968) Statistical theory of mental test scores. Addison-Wesley, Reading, MA
52. Luce RD (1959) Individual choice behavior. Wiley, New York, NY
53. King G, Tomz M, Wittenberg J (2000) Making the most of statistical analyses: improving interpretation and presentation. Am J Pol Sci 44:341–355
54. Masters GN (1982) A Rasch model for partial credit scoring. Psychometrika 47:149–174
55. McCullagh P (1980) Regression models for ordinal data (with discussion). J R Stat Soc Ser B 42:109–142
56. McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London
57. McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontiers of econometrics. Academic Press, New York, NY, pp 105–142
58. Mignani S, Cagnone S (2008) University formative process: quality of teaching versus performance indicators. Quaderni di Statistica 10:191–203
59. Monari P, Mignani S (2008). Modalità per la valutazione e il monitoraggio del processo formativo, dalla didattica all'apprendimento: l'esperienza dell'Universit'a di Bologna. Meeting at Università di Napoli Federico II, 6th March 2008, available at http://www.dipstat.unina.it
60. Moustaki I (2000) A latent variable model for ordinal data. Appl Psychol Meas 24:211–223
61. Moustaki I (2003) A general class of latent variable model for ordinal manifest variables with covariates effects on the manifest and latent variables. Br J Math Stat Psychol 56:337–357
62. Moustaki I, Knott M (2000) Generalized latent trait models. Psychometrika 65:391–411
63. Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A 135: 370–384
64. Penfield RD, Camilli G (2007) Differential item functioning and item bias. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 125–167
65. Petrucci A, Rampichini C (2000) Indicatori statistici per la valutazione della didattica universitaria. In: Civardi M, Fabbris L (eds) Valutazione della didattica con sistemi computer-assisted. Cleup, Padova
66. Piccolo D (2003) On the moments of a mixture of uniform and shifted binomial random variables. Quaderni di Statistica 5:85–104
67. Piccolo D (2006) Observed information matrix for MUB models. Quaderni di Statistica 8:33–78
68. Piccolo D, D'Elia A (2008) A new approach for modelling consumers' preferences. Food Qual Prefer 19:247–259
69. Piccolo D, Iannario M (2008) A package in R for CUB models inference, Version 1.1, available at http://www.dipstat.unina.it
70. Rao CR, Sinharay S (eds) (2007) Psychometrics. Handbook of Statistics, vol 26. North-Holland, Amsterdam
71. Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Nielson Lydiche, Copenhagen
72. Reckase MD (2007) Multidimensional item response theory. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 607–642
73. Reeve BB (2002) An introduction to modern measurement theory. National Cancer Institute, USA
74. Samejima F (1969) Estimation of latent trait ability using a response pattern of graded scores. Psychometrika Monogr Suppl 17:1–139
75. Sijtsma K, Hemker BT (2000) A taxonomy of IRT Models for ordering persons and items using simple sum scores. J Educ Behav Stat 25:391–415
76. Sijtsma K, Meijer RR (2007) Non parametric item response theory and special topics. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 719–746
77. Thissen D, Steinberg L (1986) A taxonomy of item response models. Psychometrika 51:567–577

78. Thurstone LL (1927) A law of comparative judgement. Psychol Rev 34:273–286
79. Train KE (2003) Discrete choice methods with simulation. Cambridge University Press, Cambridge
80. von Davier M, Rost J (2007) Mixture distribution item response models. In: Rao CR, Sinharay S (eds) Psychometrics, Handbook of statistics, vol 26. Elsevier, Amsterdam, pp 643–661

Chapter 8
Students' Evaluation of Teaching Effectiveness: Satisfaction and Related Factors

Michele Lalla, Patrizio Frederic, and Davide Ferrari

8.1 Introduction

Student evaluation of teaching (SET) has been widely studied in the past century and considerable research has been devoted to investigate its reliability, validity, and unbiasedness [7, 25]. Often, the overall goal of the evaluations is to gauge teaching effectiveness, understood as the extent to which a given learning objective is accomplished. Effectiveness can be evaluated through (i) direct assessment of knowledge and skills acquired by the students or (ii) a questionnaire designed to survey students' opinion about the teaching styles and behaviours of teachers and/or their satisfaction [34]. Since there is no universally accepted strategy to achieve the measurements of effectiveness, students' ratings are usually employed as a primary source of data as they are easier to collect than measurements of learned knowledge/skills. As a consequence, they represent the basis for measuring not only teaching effectiveness, but also active participation and students' attitude toward academic activity, which are critical factors for the success of any teaching system. However, it is often claimed that students' evaluations do not reveal true teaching performance and can only gauge the satisfaction with their instructors.

Despite various studies showing that reliability and validity of SETs do not change significantly over time [3, 9, 26], many authors argue that effectiveness ratings are biased by teacher characteristics unrelated to effectiveness itself, including a teacher's popularity, grading style, or level of class material [6, 14, 22]. Although there is agreement that a properly designed rating system can be a valuable source of information, clearly students cannot judge all the aspects of teachers' performance [25]. For example, students' ability to detect the need for updated class materials or evaluate a teacher's depth of knowledge in a subject are questionable [33].

In addition, the role played by communication skills is a thorny subject: even an actor playing the part of a teacher can receive outstanding evaluations in spite of the

M. Lalla (✉)
Dipartimento di Economia Politica, Università di Modena e Reggio Emilia
Modena, Italy
e-mail: michele.lalla@unimore.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_8,

fact that his/her performance had little or no educational content [32]. This observation, however, should not diminish the relevance of students' responses, simply because a necessary condition of the teaching process is the communication from teacher to students. Sometimes, students' background and experience are insufficient to answer particular questions. Other times, there are questionnaires where one or more items are not well formulated. In order to establish the questionnaire completion date, one should solve the problem of whether or not the data should be collected before or after the exam. For each given solution, it is clear that if students' ratings are likely to depend on the outcome of the exams, the interpretation of the students' responses needs to be modified accordingly. Other issues of SET concern how the survey results are used (i) by teachers to improve their performance and (ii) by administrators to make decisions about tenures, promotions, and salary bonuses. Both issues demand attention to the meaning of a high score, appropriate rewards for teaching effectiveness, and the relationship between teaching and research skills [18, 20, 21]. Currently, different opinions on this matter outline various scenarios among different universities. In some cases, both research and teaching are considered equally important by the academic institutions, while individual teachers prefer only one over the other. Sometimes, a good teaching performance corresponds to little or no reward. Other times, the research productivity is better remunerated. Consequently, the procedure of SET can be strengthened only if a certain reward system is properly adopted without weakening the support for research excellence [35].

Although SET has other potential drawbacks, it is still of paramount importance. In some cases, it is the only viable method for evaluating teachers' performance inside the classrooms. Students can benefit from its presence as it creates an incentive for teachers to perform well. Besides, it is natural to expect that students' satisfaction and teaching effectiveness will continue to play a key role in many important decisions taken by administrators, teachers, researchers, and students themselves. In fact, SET pursues many objectives of an academic institution: transparency, control and monitoring of the teaching process, attention to students' needs, and effectiveness.

The relationship between students' attitude towards teaching evaluation and the success of the evaluation procedure itself has been frequently investigated in education sciences. Nevertheless, accurate studies of the factors driving the rating are relatively rare. In our view, the understanding of such factors is crucial as they can seriously affect students' rating. Hence, a correct interpretation of students' evaluation in light of such factors is a fundamental tool for developing the potentiality of the evaluation data. In the present chapter, we explore the relationships between student ratings and various characteristics of students, courses, and teachers. In particular, we are mostly concerned with students' overall satisfaction as well as clarity of the lectures.

The chapter is organised as follows. In Sect. 8.2, we present a literature review and outline some statistical methods applied to the analysis of evaluation data. In Sect. 8.3, we describe the questionnaire used at the University of Modena and Reggio Emilia and the variables used in the model set up. The analysed data refer only

to the Faculty of Economics, where Faculty hereafter includes both teaching staff and schools and courses. In Sect. 8.4, we illustrate the fitted models and discuss our results. In Sect. 8.5, concluding comments are given.

8.2 Literature Review

Evaluation of teachers' performance has been a matter of concern for a long time because of the lack of confidence in students' capability to provide unbiased and competent judgement. The trustworthiness of evaluation responses is often conditional on the number of students involved in the evaluation process, which is considered a key factor to ensure a robust and (improperly) reliable measurement. From the students' point of view, factors that may motivate them to be involved in the evaluation process are: effective teaching style, well-organized class content and format, presence of meaningful feedback from the teacher and availability of evaluation results. From the policy makers' point of view, such factors are not so relevant when making decisions about tenures, promotions, and salary raise [12]. Furthermore, there are subtle philosophical and methodological issues concerning the measurement of teaching activity and student learning, as they can be considered as two distinct but closely related objects. From a philosophical standpoint, one can say that teachers' evaluation based on students' achievements is unfair, as evaluations represent only a partial description of the actual achievements. However, in the opposite direction one can argue that the teachers' evaluation based on students' achievements is fair just because it is based on the attainment of teaching purposes. Methodologically, in order to ensure objectivity of the analyses, it is important to: (i) ensure proper control on characteristics regarding student, teacher, and course, (ii) choose a suitable standardized questionnaire and (iii) agree a priori on what students are supposed to learn.

In general, SET serves two purposes: formative and summative evaluation of teaching. The first refers to the feedback to teachers who desire to improve their teaching performance based on suggestions on the style, content, format, and overall structure of their courses [3, 27]. In other words, teachers must perceive the evaluation questionnaire as helpful. The summative function provides information for administrative decisions regarding tenures, promotions or pay raise and for students' selection of teachers or courses [10, 24, 27]. In many institutions, teaching evaluations are publicly available to students. Thus, groups of students can request these data and circulate them to other students [12]. In summary, administrators find them useful for decision making and students feel that the effort involved in filling the questionnaires is worthwhile. The availability of the evaluation results is expected to positively affect their involvement and motivation and guide them in developing their curricula.

Knowledge of the mechanisms underlying the SET process and the factors affecting its outcomes can help to interpret correctly the empirical results. Therefore, the determinants of a good teaching performance have been subject of extensive investigation. However, an overstatement of their relevance can be a trap, as the

characteristics of good teaching are based only on the data and not based on "ideal" behaviour. The outcomes of the process derive from a complex interaction between the personality traits of the teachers and those of students, involving also psychological aspects and characteristics of the course and other related factors.

The determinants of teaching effectiveness can be analysed following one of two approaches. The "ideal type" approach is based on a survey on preferred/exemplar characteristics of an effective teacher. The target populations of the survey are both teachers and students. Empirical studies showed high correlations between the responses from students and teachers, based on ordinary attributes, such as preparation, organization, clarity, comprehensibility, fairness, and sensitivity to class level and progress. However, due to their different academic roles, it was observed that students preferred teachers being interesting and skilful in presentations, whereas teachers preferred to involve the students in intellectual challenges and encourage self-initiated learning [16]. As far as which aspects of student feedback is of greatest use, teachers prioritize their interaction with students, while administrators focus on the structural issues of the courses [7].

The "causal model" approach is also based on the surveyed opinions of students about the many aspects of teaching activity. Here, the characteristics of effective teachers are identified through a statistical model, often a linear regression model, which is employed to unveil the relationship between overall effectiveness evaluations and specific questionnaire items [7]. In this framework, it is common to assume that SET is a good measure for teacher effectiveness.

A single-equation regression has been often used to ascertain the relationship between student evaluations and certain characteristics of students, teachers, and courses. However, some variables could be dependent and contemporaneously independent variables; therefore, a simultaneous equation system has been applied. For example, Nelson and Lynch [30] developed a three-equation model to verify the hypothesis that teaching evaluations contribute to grade inflation, where the dependents were the average students' expected grades, the average teacher evaluation, and the overall course quality evaluation. The model included ratings on selected questionnaire items to control for the impact of teacher characteristics on overall course quality evaluation, course, teacher, and student characteristics (including students' expected grades).

Data reduction techniques, like principal component and factor analysis, are often used to process a multiplicity of indicating variables, as is the case of information collected by SET questionnaires. In particular, structural equation modelling combines multiple regression, factor and path analyses to investigate the pattern of causal connection between both observed and latent variables. For example, it has been applied to determine the relationships between the teachers' evaluation scores and students' learning [32].

In most cases, students' evaluation of teachers and courses are expressed through a set of discrete alternatives because the items' scales range form one to four, five or seven, i.e., they are measured through an ordinal scale. Therefore, the appropriate statistical methods are those able to handle qualitative variables, such as

(multinomial) logit or probit. For example, DeCanio [15] compared the multinomial logit and linear regression specifications to analyse the impact of teacher characteristics on the effectiveness ratings of teachers. Regardless of the model's specification, many of the questionnaire items had a significant influence on ratings of teacher effectiveness. Boex [7] applied an ordered probit model to ascertain to what degree the identified teacher attributes contribute to the overall effectiveness rating. Mehdizadeh [29] applied a loglinear model, but the latter presents some limitations as regards the analysis of the interrelationships of all variables that could be included in the model because the sample size required would increase exponentially and the investigation becomes unfeasible.

The multidimensional view of the education process implies that, because of the complexity of teaching, instruction simply cannot be represented by one single measure, such as an effectiveness rating [25, 28]. Therefore, only multiple measures of teacher attributes could characterize properly the effectiveness of teachers. Thus, particular attention should be devoted to the definition and quantification of teachers' attributes, without relating them to a single measure of overall effectiveness. Conversely, the unidimensional view of the education process implies that instruction can be appropriately represented by a single effectiveness measure, even if it recognises that effective teaching can vary across teachers, courses, students, and settings [14].

8.3 Questionnaire and Data

Student evaluations of teaching activity are mandatory in Italian universities and the National Committee for University System Evaluation (Italian acronym CNVSU) proposed a course-evaluation questionnaire containing 15 items, reported in Table 8.1, with a four-point Likert scale: (i) *Definitely not*, (ii) *No, rather than yes*, (iii) *Yes, rather than no*, (iv) *Definitely yes*. Each category was translated into the values of a decimal scale ranging from 2 to 10, where the complete set of each item is {2, 5, 7, 10}, as suggested by Chiandotto and Gola [13]. They simply proposed to evaluate the teacher in each course by the mean of the decimal scores, including its standard deviation and the number of cases. This approach has the usual problems related to such data: absence of the middle category, arbitrariness of the numbers assigned to alternatives, incomprehensibility of the labels for many students and high level of uncertainty about their intensities [23, 31]. Moreover, since the variables are ordinal, their mean and standard deviation should not be used, although this would require a more detailed discussion beyond our present scope. Nevertheless, the university of Modena and Reggio Emilia adopted only the suggested 15-items and, in addition, introduced nine dichotomous observations or suggestions. The timing of the surveys followed the academic calendar and the collection periods were three weeks before the end of the term. The questionnaires were accessible to students via the Internet on a voluntary participation basis.

Table 8.1 Questionnaire items with median (Mdn) and mean or observed proportion (OP) and standard deviation (SD): $n = 4111$

Evaluation items (ordinal: 2, 5, 7, 10)	Acronym	Mdn	Mean
I01: Adequacy of the Work Load requested by the course	AWL	7	6.80
I02: Adequacy of the Teaching Materials for learning	ATM	7	7.51
I03: Usefulness of Supplementary Teaching Activity (STA)	USTA	7	7.53
I04: Clarity of the Forms and rules of the Exams	CFE	7	7.64
I05: Keeping of the Official Schedule of Lectures	KOSL	10	8.77
I06: Teacher Availability for Explanations	TAE	10	8.43
I07: Motivations and Interests aroused by Teacher	MIT	7	7.28
I08: Clarity and Exactitude of the Teacher's Presentations	CETP	7	7.49
I09: Adequacy of the Lecture Room	ALR	7	7.59
I10: Adequacy of the Room and Equipment for the STA	ARESTA	7	7.44
I11: Sufficiency of the Background Knowledge	SBK	7	6.84
I12: Level of Interest in the Subject matter	LIS	7	7.59
I13: Level of Overall Satisfaction with the course	LOS	7	7.18
I14: Adequacy of the required Total semester Work Load	ATWL	5	5.74
I15: Total Organization Sustainability (lectures and exams)	TOS	7	5.90
Observations items (dichotomous: 1/0)		**OP**	**SD**
O1: Improvement in the Coordination between Courses	ICC	0.15	0.36
O2: Reduction of the Work Load requested by the course	RWL	0.26	0.44
O3: Providing more Basic Knowledge	PBK	0.16	0.37
O4: Improvement of Teaching Materials	ITM	0.14	0.34
O5: Removal of Redundancies	RR	0.05	0.22
O6: Increase of Practice	IP	0.20	0.40
O7: Teaching Materials Before the Beginning of the course	TMBB	0.14	0.35
O8: Increase of Supplementary Teaching Activity	ISTA	0.10	0.29
O9: Introduction of Intermediate Examinations	IIE	0.07	0.25

In the present study, the data concern all classes in business and economics offered by the Faculty of Economics during the academic year 2006/2007. The sample of the web-based survey showed some differences from that of the paper-based survey carried out in the classroom: (i) an increase of 45.7% in the number of courses being evaluated, i.e., from 162 to 236 and (ii) a decrease of 27.4% in the number of participating students with respect to the traditional paper survey. In order to reduce the sample size effect on the variables referred to the teacher and course, we considered only evaluations of courses with at least five responding students. The total sample size was $n = 4{,}111$ responding students. The 15-items and the 9-observations, the core of SET, are reported in Table 8.1 with their median and mean of the decimal scores. The mean, in spite of its theoretical misuse, appears more informative than the median [23]. The quartile deviation and the standard deviation were not reported, but their values were about 1.5 and 2, respectively.

The evaluation questionnaire contained several sections to identify the faculty, school, teacher, course, and some student characteristics. The descriptive statistics of the variables generated from these sections are reported in Table 8.2 with supplementary data about teachers and courses.

Table 8.2 Variables concerning students, teachers, and courses with observed proportion (for binary variables) or mean (OP/*M) and standard deviation (SD): $n = 4,111$

Variables	Acronym	OP/*M	SD
Student characteristics			
Female	F	0.668	0.471
Liceo specialising in Classical Studies	LCS	0.173	0.379
Liceo specialising in Scientific Studies	LSS	0.162	0.369
Industrial Technical Institute	ITI	0.050	0.219
Commercial Technical Institute	CTI	0.422	0.494
Other type of School-Leaving Certificate	OSLC	0.192	0.394
Enrolment Year (1, 2, 3, 4, 5)	EY	*2.664	1.215
Type of Enrolment	TE	0.010	0.098
UG: Business Economics (UG = Under Graduate)	BE	0.331	0.471
UG: International Economy and Marketing	IEM	0.382	0.486
UG: Economic Sciences and Society	ESS	0.037	0.190
G: Economics (G = Graduate)	ES	0.000	0.000
G: Public Policies and Territory Evaluation	PPTE	0.015	0.121
G: Consulting and Management of Firms	CMF	0.041	0.198
G: Labour Relations	LR	0.014	0.119
G: Financing Analysis, Consulting, and Management	FACM	0.041	0.199
G: International Management	IM	0.138	0.345
Percentage of Attended Lectures (1, 2, 3)	PAL	*2.665	0.635
Teachers characteristics			
Female Teachers	FT	0.336	0.472
Full Professor	FP	0.396	0.489
Associate Professor	ActP	0.250	0.433
Assistant Professor	AstP	0.240	0.427
Non Academic Teacher (Lecturer)	NAT	0.113	0.316
Course characteristics			
Class Size: median divided by ten (0.5 ÷ 20.0)	CS	*9.101	4.355
Proportion of Evaluating Students (0.03 ÷ 1.96)	PES	*0.392	0.318
Juridical Sciences	JS	0.124	0.330
Business Economics	BE	0.139	0.346
International Economy and Marketing	IEMk	0.144	0.351
Organisation	O	0.042	0.201
Banking and Finance	BF	0.096	0.295
Languages	L	0.083	0.276
Mathematics and Statistics	MS	0.137	0.344
Micro-Macro Economics	MME	0.074	0.263
Economics (Courses)	EC	0.084	0.277
Public Finance	PF	0.048	0.214
History and Sociology	HS	0.027	0.162

*Mean of the non-binary variable, the support of which appears in parenthesis.

The available variables describing students' characteristics were gender, type of education level, enrolment year, type of enrolment, form of enrolment, percentage of attended lectures, and the class size estimated by respondents. We encoded gender as a binary variable, where 1 represents women and 0 represents men. The type of education level was specified by five dichotomous variables: *liceo* specialising

in classical studies, *liceo* specialising in scientific studies, industrial technical institute, commercial technical institute, and the residual category "other type of school-leaving certificate".

The enrolment year took values of 1, 2, or 3 for undergraduates and 4 or 5 for graduates. The type of enrolment was a binary variable taking the value of 1 if the student did not pass his/her exams within the prescribed time and 0 otherwise. The form of enrolment was binary: 1 if the student had a part-time enrolment and 0 otherwise. However, due to the extremely small number of part-time students, the form of enrolment was not included in the models. The attended school was represented by a binary variable and it is possible to distinguish between undergraduate and graduate, as indicated in Table 8.2. The percentage of attended lectures consisted of a set with three values: 1 if it is less than 50%, 2 if it is between 50 and 75%, 3 if it is over 75%.

Among other questions, each student was asked to estimate the class size. The median of the students' estimated sizes was adopted as the actual size of the course because of its robustness. In fact, the empirical distribution of the estimated sizes showed a long right tail with evident outliers. In the model it was divided by 10 and introduced as a polynomial of the second order. The proportion of evaluating students was defined as the ratio between the number of respondents and the class size given by the median from students' evaluations. Therefore, the proportion could be greater than 1 and, actually in two classes was 1.07 and 1.96 with medians of 30 and 50, respectively. Two factors were available for teachers: gender and professional position. The courses were grouped based on their scientific disciplinary field (see Table 8.2).

8.4 Models and Results

Student's level of overall satisfaction (LOS) and clarity of teachers' presentations (CETP) were assumed as dependent variables because the first is a proxy for teaching effectiveness and the second is an important aspect of instructors. All the items concerning the instructors were possible candidates, but only these two variables were selected for the sake of brevity and simplicity.

The selection of the explanatory variables was carried out combining aprioristic and statistical considerations. The items concerning the usefulness of supplementary teaching activity (I03) and the adequacy of the room and equipment for this activity (I10) were not included in the model. The main reason is that they are not specifically designed for students attending the courses taught in the economic schools. Actually, they are intended for evaluation of scientific activity or laboratory sections. In fact, the number of respondents is very low (less than 45%). The remaining questions in the 15-items battery were always included as explanatory variables, even when the fitted model coefficients were not statistically significant. The selection of the other explanatory variables was carried out using statistical procedures: backward and forward selection in a linear regression model with a single equation,

used as an explorative tool [2]. As a result, for the 9-observation binary items and other characteristics, only those with significant coefficients were included in the final model.

In general, the effect of student/teacher gender on the ratings of the overall satisfaction is not completely clear. Some empirical findings do not show impact, but indicate that it may interact with other factors to generate low ratings for women. For example, an interaction between the professor gender and the student gender, often described as same-sex-preference, is plausible. In fact, students tend to score better same-sex teachers' vocal quality and other related factors [5, 17]. In some studies, male professors received the same evaluation by their male and female students, while female professors received higher evaluations by their female students [4, 11]. Searching for meaningful interactions is a complicated task and their interpretation is not always straightforward. Thus, in this study we preferred to account only for the main effect of gender. In particular, we considered the combinations of the levels of gender (female or male) and the levels of academic role (instructor or student). Each combination was encoded by a separate binary variable: male student and female professor (MS-FP), female student and male professor (FS-MP), female student and female professor (FS-FP). Only the combination male student and male professor (MS-MP) was excluded from the model and assumed as the reference group.

Since the dependent variables are expressed through a four-point Likert scale, the Ordered Logit Models (OLM) is appropriate. The interested reader can find an extensive review of this topic in Agresti [1]. Let Y and X be the response variable and the vector of predictors, respectively. Moreover, let $F_Y(j)$, for $j = 1, \ldots, 4-1$, be the cumulative distribution function (cdf) of Y, where the index j denotes the level of the response. The OLM is formulated as

$$\text{logit}\left[F_Y(j)\right] = \log\left(\frac{F_Y(j)}{1 - F_Y(j)}\right) = \alpha_j - \beta'\mathbf{x} \tag{1}$$

where α_j parameters are required to be such that $\alpha_j < \alpha_{j+1}$, $\forall j$, and the linear coefficients β describe the effect of the covariate vector $\mathbf{X}$ on the response. One can justify the effect of β for different j by assuming that a model holds when the response is measured more finely. Let Y^* be a latent continuous variable having cdf $G(y^* - \eta)$ with location parameter depending on $\mathbf{x}$, $\eta(\mathbf{x}) = \beta'\mathbf{x}$. Then, the ordinal variable Y is equal to j, when $\alpha_{j-1} < Y^* < \alpha_j$. Hence, we have $F_Y(j) = G(\alpha_j - \beta'\mathbf{x})$ and by choosing G to be the logistic distribution, one obtains model (1).

The results of the parameter estimates are reported in Table 8.3 for the level of overall satisfaction (LOS). The interpretation of the coefficients is not straightforward. However, some interesting considerations follow just by looking at their sign and their significance. All the predictors I01–I15 showed significant coefficients for explaining students' level of overall satisfaction. The adequacy of the lecture room (ALR), however, turns out to be an exception. This result is somewhat surprising because such a factor is usually expected to be an important element of satisfaction. A possible interpretation of this finding is that the lectures were overall satisfactorily

Table 8.3 Estimated coefficients (β), standard errors (SE), and p-values for the OLM and SUR models (dependent: I13-LOS)

	OLM: Dependent I13-LOS			SUR: Dependent I13-LOS		
Variables by acronym	β	SE	p-val.	β	SE	p-val.
Items/intercept				−1.440	0.202	0.000
101-AWL	0.143	0.021	0.000	0.071	0.011	0.000
I02-ATM	0.257	0.026	0.000	0.100	0.013	0.000
I04-CFE	0.145	0.022	0.000	0.058	0.011	0.000
I05-KOSL	0.057	0.027	0.034	0.014	0.014	0.314
I06-TAE	0.091	0.029	0.001	0.013	0.015	0.395
I07-MIT	0.408	0.027	0.000	0.113	0.014	0.000
I08-CETP	0.490	0.028	0.000	0.489	0.013	0.000
I09-ALR	−0.011	0.020	0.586	−0.011	0.010	0.294
Ill-SBK	0.123	0.021	0.000	0.052	0.011	0.000
I12-LIS	0.281	0.023	0.000	0.152	0.012	0.000
I14-ATWL	−0.065	0.029	0.023	−0.036	0.015	0.016
I15-TOS	0.109	0.028	0.000	0.049	0.014	0.001
Observations						
04-ITM	−0.379	0.122	0.002	−0.093	0.066	0.159
05-RR	−0.306	0.172	0.074	−0.151	0.086	0.078
06-IP	−0.180	0.102	0.077	−0.118	0.053	0.027
Students						
LCS	0.193	0.110	0.081	0.087	0.054	0.105
LSS	0.243	0.112	0.030	0.113	0.055	0.040
PAL	0.284	0.064	0.000	0.110	0.032	0.001
ESS	0.368	0.221	0.096	0.137	0.106	0.197
LR	−0.962	0.335	0.004	−0.419	0.174	0.016
Teachers						
ActP	−0.185	0.106	0.081	−0.090	0.053	0.087
NAT	0.412	0.156	0.008	0.174	0.078	0.026
MS-FP	−0.036	0.144	0.805	−0.007	0.075	0.928
FS-MP	0.221	0.106	0.036	0.154	0.055	0.005
FS-FP	0.078	0.125	0.534	0.024	0.065	0.714
Courses						
CS	0.114	0.044	0.010	0.048	0.022	0.028
CS^2	−0.005	0.002	0.028	−0.002	0.001	0.036
PES	*			*		
JS	0.365	0.159	0.021	0.105	0.081	0.196
IEMk	0.461	0.152	0.002	0.204	0.077	0.008
O	0.391	0.223	0.079	0.094	0.117	0.422
BF	0.589	0.165	0.000	0.218	0.085	0.010
MS	0.382	0.156	0.014	0.194	0.077	0.012
L	*			*		
MME	0.548	0.189	0.004	0.326	0.095	0.001

* The variables of courses significant for CETP are reported without coefficients.

organised in the classroom, when all else is considered constant (including the attribute ratings). Furthermore, all the coefficients were positive, except that which referred to the adequacy of the total work load (ATWL). Thus, an increase in the work load appears to be associated to a decrease in LOS for the different levels of the categorical response.

Three out of nine observation items – improvement of teaching materials, removal of redundancies, and increase of practice – showed negative impact on satisfaction. Therefore, student satisfaction could be increased by inducing teachers to improve these aspects.

The estimated coefficients for non-academic teachers (NAT) were positive, implying a stronger impact on LOS than institutional professors. However, this kind of teacher was mainly represented by lecturers in foreign languages, who generally had scores higher than teachers in other subjects. Actually, the binary variable indicating whether the classes were in foreign languages were not included in the set of class characteristics. The reason is that its explanatory effect was completely absorbed by the binary variable of non-academic teachers. The coefficient of associate professors (ActP) was negative and significant only at a 0.1 level. The interaction between the gender of teachers and students did not always produce the expected signs and p-values according to the hypothesis of same sex-preference with respect to overall satisfaction: the binary variable for female students evaluating male professors yielded a positive coefficient implying that females were more satisfied than males in evaluating male professors.

The students who attended *liceo* – particularly those specializing in scientific studies (LSS) – turned out to be more satisfied than students in other educational levels. The students who had a larger percentage of attended lectures (PAL) were more satisfied than those attending less often. Undergraduates enrolled for the degree in Economic Sciences and Society (ESS) tended to be more satisfied than those pursuing other degrees. The opposite tendency resulted for graduate students pursuing a degree in Labour Relations (LR).

The class size (CS) had a negative effect on LOS, but also on teachers' attribute ratings. This is likely due to the fact that larger class sizes reduce a teacher's opportunity to interact with students on a one-to-one basis. This prevents the teacher to provide better explanations for the portion of the class who have more difficulties in grasping the concepts. However, an interesting nonlinear relationship emerged from data. Lower class size and large class size corresponded to lower satisfaction level than that observed for medium class size. Students may tend to limit their interactions in smaller classes to prevent revealing their inadequacies to the teacher. The low satisfaction for large classes has a different nature and is often due to circumstances beyond of teacher's control. Changing the mechanism that assigns teachers to courses might positively affect this pattern. If department heads or deans could find a way to assign effective teachers (with teaching qualities that were observed by students) to larger course sections consistently, the teaching effectiveness could achieve some improvements.

Almost all of the scientific disciplinary fields proved to have significant and positive coefficients, implying that classes in those fields met with more satisfaction than those in the reference field, that is, business economics, which contained the most popular courses. For most fields, the coefficients were similar in size and, thus, the change of reference field corresponded to a change in the number of significant binary variables.

As far as the clarity and exactitude of teachers' presentations (CETP) is concerned, the estimated parameters of which are reported in Table 8.4, the explanatory items I01–I15 showed significant coefficients, except for the adequacy of the work load requested by the course (AWL), the sufficiency of the background knowledge (SBK), and the adequacy of the total work load requested by the current courses (ATWL). The coefficient of AWL was negative implying that an increase in the work load was detrimental to clarity. Surprisingly, the coefficient of SBK was not statistically significant, but it could be connected with (and hidden by) the level of interest in the subject matter (LIS), which has a negative expected coefficient. Furthermore, SBK is clearly connected with the observation "providing more basic knowledge" (PBK) and its significance could be weakened by the presence of PBK in the model (see below). The problem of background knowledge refers substantially to the first year, as students come from high schools providing different education. The subjects taught in the first and the subsequent years should be consistent with those learned earlier. Therefore, one could have expected the lack of significance for the coefficient of SBK.

Two out of four binary observations included in the model – providing more basic knowledge (PBK) and improvement of teaching materials (ITM) – had a negative impact on clarity because students having difficulties in grasping concepts were limited by these factors and not only they expressed their difficulties through the items, but they reinforced them by filling the observations PBK and ITM as well. The other two observations – increase of practice (IP) and introductions of intermediate examinations (IIE) – had positive coefficients which are not easy to interpret. Perhaps, IP represented also the effect of dummy variables characterising the fields that did not enter in the included set. They were courses in mathematics, statistics, public finance, and economics. The latter subject should be appreciated by students, but the question now becomes whether they really wanted to have more practice or they were just complaining. This is a hard one to answer. Yet, the significant coefficient for IIE was complicated enough because the intermediate examinations existed in the organisation of the Faculty of Economics. Therefore, the significance of this coefficient might involve a presence of difficulties in the courses (inadequacy of the workload or complexity of the subjects). The expected sign of the coefficient should have been negative, but it was positive and the same interpretation as for IP could apply.

The hypothesis of same sex-preference for the clarity (CETP) seemed strengthened by data: the dummy variable of female students evaluating the clarity of male professors had a negative coefficient implying that females were stricter than males in evaluating male professors. The coefficients of other dummy variables were not significant, but the sign of the dummy variable for male students evaluating clarity of

Table 8.4 Estimated coefficients (β), standard errors (SE), and p-values for the OLM and SUR models (dependent: I08-CETP)

	OLM: Dependent I08-CETP			SUR: Dependent I08-CETP		
Variables by acronym	β	SE	p-val.	β	SE	p-val.
Items/intercept				0.581	0.208	0.005
101-AWL	−0.022	0.021	0.284	−0.034	0.013	0.008
I02-ATM	0.154	0.025	0.000	0.046	0.016	0.003
I04-CFE	0.053	0.021	0.010	0.012	0.013	0.359
I05-KOSL	0.045	0.025	0.075	0.018	0.016	0.254
I06-TAE	0.161	0.026	0.000	0.080	0.017	0.000
I07-MIT	0.561	0.026	0.000	0.274	0.015	0.000
I09-ALR	0.043	0.018	0.016	0.028	0.011	0.013
Ill-SBK	0.003	0.021	0.901	−0.021	0.013	0.100
I12-LIS	−0.084	0.022	0.000	−0.101	0.014	0.000
I13-LOS	0.531	0.030	0.000	0.642	0.017	0.000
I14-ATWL	0.004	0.027	0.894	0.006	0.017	0.738
I15-TOS	0.056	0.027	0.035	0.010	0.016	0.547
Observations						
03-PBK	−0.277	0.109	0.011	−0.159	0.066	0.017
04-ITM	−0.364	0.116	0.002	−0.187	0.076	0.014
06-IP	0.195	0.099	0.050	0.150	0.062	0.016
09-IIE	0.292	0.145	0.045	0.159	0.087	0.069
Students						
OSLC	0.318	0.101	0.002	0.146	0.060	0.015
EY	−0.123	0.041	0.003	−0.066	0.025	0.008
IEM	−0.251	0.092	0.007	−0.111	0.055	0.044
CMF	0.503	0.208	0.016	0.284	0.124	0.022
LR	0.789	0.344	0.022	0.521	0.200	0.009
FACM	0.897	0.222	0.000	0.497	0.131	0.000
Teachers						
MS-FP	−0.168	0.141	0.233	−0.072	0.088	0.412
FS-MP	−0.328	0.103	0.001	−0.213	0.063	0.001
FS-FP	0.035	0.123	0.775	0.025	0.076	0.738
Courses						
CS	*			*		
CS^2	0.002	0.001	0.002	0.001	0.000	0.014
PES	−0.398	0.151	0.008	−0.187	0.090	0.038
JS	0.512	0.127	0.000	0.287	0.079	0.000
IEMk	0.599	0.128	0.000	0.287	0.078	0.000
O	0.605	0.199	0.002	0.365	0.126	0.004
BF	0.465	0.133	0.000	0.210	0.084	0.012
MS	*			*		
L	0.550	0.156	0.000	0.303	0.095	0.001
MME	−0.268	0.153	0.079	−0.210	0.095	0.026

* The variables of courses significant for LOS are reported without coefficients.

female professors were negative again, according to the hypothesis of the same sex-preference, i.e., male students evaluating clarity of female professors were stricter than female students.

Other types of school-leaving certificates showed a positive impact on the clarity (CETP), while the enrolment year showed a negative impact, i.e., by advancing in their studies they decrease the score of CETP. This may be the result of courses increasing in difficulty or that previous preparation was not adequate for students to meet current course requirements. Only the squared term of class size proved to be significant and positive, but this is the reverse of satisfaction: low class size and high class size induced high CETP. The undergraduates enrolled for International Economy and Marketing (IEM) had a negative coefficient meaning they were reluctant to provide high CETP scores. The opposite behaviour was observed for graduates enrolled for Consulting and Management Firms (CMF), Labour Relations (LR), and Financing Analysis-Consulting-and-Management (FACM). The number of evaluating students over the class size (PES) yielded a negative coefficient meaning that CETP decreased when PES increased. This could be intended as a protest for the difficulty in attending the courses and understanding the concepts taught; in fact, when the clarity of the teacher decreases, the percentage of evaluating (participating) students increases.

Almost all the scientific disciplinary fields proved to have significant and positive coefficients again, implying that courses belonging to those fields conveyed more satisfaction than the reference field, business economics. For most of fields, the coefficients were almost equal and, therefore, the change of reference field will change the number of significant dummy variables too.

This approach could be interpreted in the direction of the unidimensional view, as the model reveals to what degree teacher, student, and course attributes determined satisfaction and, thus, the effectiveness ratings. Although this forced parallel is questionable, discarding it would imply to use a unique measure of SET. Therefore, most of items would have been excluded from the model resulting in the loss of some interesting patterns.

In order to validate the findings from the OLM analyses, we also employed more familiar linear regression tools [8, 12, 19, 30], which have the advantage to be easily interpretable. Although linear regression do not directly apply, at least in principle, its use is convenient for explorative purposes [2]. Particularly, we use seemingly unrelated regressions model (SUR) as we were interested in two equations, where the regressors in each of them include the dependent variable of the other.

The coefficients of models for LOS and CETP were estimated simultaneously. However, the results are reported separately in Tables 8.3 and 8.4 to facilitate the comparison between OLM and SUR in terms of the estimated coefficients and their p-values. Specifically, the estimated parameters of the SUR model for LOS are reported in Table 8.3, together with those of the OLM for LOS, while the estimated parameters of the SUR model for CETP are reported in Table 8.4, together with those of the OLM for CETP.

The OLM and SUR gave the same signs of the coefficients. The coefficients of determination in the model with LOS as response were $R^2(Adj) = 0.662$ for the

single equation, $R^2(Adj) = 0.638$ for SUR, and Pseudo-$R^2 = 0.465$ for OLM. The coefficients of determination in the CETP models were $R^2(Adj) = 0.618$ for the single equation, $R^2(Adj) = 0.591$ for SUR, and Pseudo-$R^2 = 0.395$ for OLM. In the single equation models, the histograms of the residuals for LOS and CETP were approximately normal. However, in the SUR model, the residuals appeared to depart sensibly from normality assumption for CETP. In particular, the histogram showed three evident spikes centred at about ± 2 and 0.

The nature of OLM and SUR does not allow for a direct comparison in terms of performance. However, the results from our explorative analysis via SUR are consistent with that based on single-equation models and OLM. This suggests an extension of the present work using multiple equation models for categorical responses such as multilevel ordered logit models. Other popular models, like optimal scaling or structural equation models focus on latent factors. Therefore, they do not show directly the effect of regressors on observed responses, which is the very purpose of our analysis.

8.5 Conclusions

Competitiveness is generally assumed as an aim to be pursued, even though it is a controversial point. Therefore, in order to be competitive (in the educational market), the universities need to have highly satisfied students (customers) to continue their mission. Accordingly, they cannot ignore the usefulness of teaching evaluations. Empirical findings demonstrate that elements of teachers' behaviours can be modified to achieve better results: adequacy of the work load requested by the course, adequacy of teaching materials, clarity of the forms and rules of the exams, and teacher availability for explanations. Therefore, teachers should be aware of the content and the quality of their course materials, the context of teaching process, the expectations and the skills of students, the organisation of the exams, and the corresponding grading system because they are important in student learning and satisfaction [35]. Improvements in teaching is an opportunity for the future of Italy and other European countries as promoting student learning directly enriches the cultural and professional assets of society.

Teaching, research, organisation, and services are the key missions of university. Generally, ratings of teaching effectiveness are currently used by persons in charge (like the dean, department chair or teaching evaluation committee) for personnel decisions, such as tenure, promotion, and merit pay. In Italy, at this stage, SET is not employed for such decisions, but this use requires much cautions and care. In fact, this practice is not widely applied in the world and the relationship between teaching and financial rewards is weak [35]. Often, teaching is less valued than research and research ability is more visible, simple to transfer among institutions, and easy to assess through publication. Teaching ability is invisible, harder to document, and much less transferable, i.e., research publications typically stimulate job offers over excellence in teaching. Therefore, the users of evaluation data should be aware of the reliability and validity of SET. Hence, the subsequent step should be

the correct interpretation and usage of such ratings in order to make comparisons between teachers and courses.

Evaluation data generally show complex patterns between satisfaction and teacher, student, course, and setting characteristics. This knowledge is useful to improve the organisation of teaching aimed at increasing current learning and motivation in the future. Moreover, it enables to achieve the optimum level of teaching effectiveness based on the available resources. However, the latter could fail for lacking in comprehension of the factors influencing teaching effectiveness. Therefore, the reality of limited financial resources may restrict the scope of teaching improvements. For example, while a medium class size proved to be optimal, without a sufficient number of teachers, large class sizes will continue to be a necessity in spite of the empirical findings.

Acknowledgements Davide Ferrari is supported by a research scholarship provided by Associazione Nazionale Cavalieri del Lavoro, group of Emilia Romagna, Italy, at the University of Modena and Reggio Emilia.

References

1. Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken, NJ
2. Amemya T (1981) Qualitative response models: a survey. J Econ Lit XIX(4):1483–1538
3. Arubayi EA (1987) Improvement of instruction and teacher effectiveness: are student ratings reliable and valid? Higher Educ 16(2):267–278
4. Basow SA (1995) Student evaluations of college professors: when gender matters. J Educ Psychol 87(2):656–665
5. Basow SA, Montgomery S (2005) Student ratings and professor self-ratings of college teaching: effects of gender and divisional affiliation. J Pers Eval Educ 18(2):91–106
6. Becker WE Jr, Watts M (1999) How departments of economics evaluate teaching. Am Econ Rev 89(2):344–349
7. Boex LFJ (2000) Attributes of effective economics instructors: an analysis of student evaluations. J Econ Educ 31(3):211–227
8. Bosshardt W, Watts M (2001) Comparing student and instructor evaluations of teaching. J Econ Educ 32(1):3–17
9. Byrne CJ (1992) Validity studies of teacher rating instruments: design and interpretation. Res Educ 48 (November):42–54
10. Cashin WE, Downey RG (1992) Using global student rating items for summative evaluation. J Educ Psychol 84(4):563–572
11. Centra JA, Gaubatz NB (2000) Is there gender bias in student evaluations of teachers? J Higher Educ 71(1):17–33
12. Chen Y, Hoshower LB (1998) Assessing student motivation to participate in teaching evaluations: an application to expectancy theory. Issues Account Educ 13(3):531–549
13. Chiandotto B, Gola MM (2000) Questionario di base da utilizzare per l'attuazione di un programma per la valutazione della didattica da parte degli studenti. Rapporto finale del gruppo di ricerca (RdR 1/00), MURST, Osservatorio (ora Comitato nazionale) per la valutazione del sistema universitario, Roma (http://www.cnvsu.it)
14. D'Appollonia S, Abrami PC (1997) Navigating student ratings of instruction. Am Psychol 52(11):1198–1208
15. DeCanio SJ (1986) Student evaluations of teaching: a multinominal logit approach. J Econ Educ 17(3):165–176

16. Feldman K (1989) Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. Res Higher Educ 30(2):137–94
17. Feldman K (1993) College students views of male and female college teachers: Part II – evidence from students' evaluations of their classroom teachers. Res Higher Educ 34(2):151–211
18. Gomez-Mejia LR, Balkin DB (1992) Determinants of faculty pay: an agency theory perspective. Acad Manage J 35(5):921–955
19. Greene WH (2003) Econometric analysis, 5th edn. Prentice Hall, Upper Saddle River, NJ
20. Kasten KL (1984) Tenure and merit pay as rewards for research, teaching, and service at a research university. J Higher Educ 55:500–513
21. Katz DA (1973) Faculty salaries, promotions and productivity at a large university. Am Econ Rev 63:469–477
22. Kwan K-P (1999) How fair are student ratings in assessing the teaching performance of university teachers? Assess Eval Higher Educ 24(2):181–195
23. Lalla M, Facchinetti G, Mastroleo G (2004) Ordinal scales and fuzzy set systems to measure agreement: an application to the evaluation of teaching activity. Qual Quant 38:577–601
24. Lin YG, McKeachie WJ, Tucker DG (1984) The use of student ratings in promotion decisions. J Higher Educ 55(5):583–589
25. Marsh HW (1987) Students' evaluations of university teaching: research findings, methodological issues, and directions for future research. Int J Educ Res 11(3):263–388
26. Marsh HW, Hocevar D (1991) Students' evaluations of teaching effectiveness: the stability of mean ratings of the same teachers over a 13-year period. Teaching and Teacher Educ 7(4):303–314
27. Marsh HW, Roche LA (1993) The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. Am Educ Res J 30(1):217–251
28. Marsh HW, Roche LA (1997) Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias and utility. Am Psychol 52(11):1187–97
29. Mehdizadeh M (1990) Loglinear models and student course evaluations. J Econ Educ 21(1):7–21
30. Nelson JP, Lynch KA (1984) Grade inflation, real income, simultaneity, and teaching evaluations. J Econ Educ 15(1):21–37
31. Schuman H, Presser S (1996) Questions and answers in attitude surveys: experiments on question form, wording, and context. Sage Publications, Thousand Oaks, CA
32. Seiler VL, Seiler MJ (2002) Professors who make the grade. Rev Bus 23(2):39–44
33. Seldin P (1993) The use and abuse of student ratings of professors. Chron Higher Educ (July 21) 39(46):A40
34. Simon B, Haghirian P, Schlegelmilch BB (2003) Enriching global marketing education with virtual classrooms: an effectiveness study. Mark Educ Rev 13(3):27–39
35. Tang TLP (1997) Teaching evaluation at a public institution of higher education: factors related to the overall teaching effectiveness. Public Pers Manage 26(3):379–389

Part III
The Evaluation in the Italian Universities: Statistical Methods for Careers and Services Evaluation

Chapter 9
Modeling Ordinal Item Responses via Binary GLMMs and Alternative Link Functions: An Application to Measurement of a Perceived Service Quality

Vito M.R. Muggeo and Fabio Aiello

9.1 Introduction

Evaluation of a service on the basis of consumer opinion is a widespread practice in many fields. The assessment of perceived quality [7] of a service is generally carried out through administration of a questionnaire, composed of several items with responses posed on an ordinal scale, whereby each item represents an important feature of the evaluated service [3, 7]. In this context, the aim is to evaluate something similar to the external effectiveness, that is the part of efficacy related to the satisfaction expressed by the service users for the provided service. A particular and important example of service users is represented by students' responses measuring the perceived quality of services at the Reception Office of their College or Faculty. The service provided to the students by Reception may consist of several front office operations, including access and admission to University, student examination certificates, and career certificate credit courses. Student perceive whether or not Reception service is responsive to their administrative needs. Typically service evaluation occurs on the basis of the student's perception which is measured by means of the questionnaire administration. In such types of service is fundamental the "interaction" between the user (i.e. the student) and the service provider [24]. The service is interactive because of the student asks for information, certificates, and suggestions and the service provider may or not satisfy the requests.

Evaluation of such a type of service, named "Public Utility Services" provided by government to its citizens is important [5]; examples include, among others, assessment of Health care and Education services. The assessment of the efficiency, the efficacy and the resources' allocation has assumed increasing relevance over the last years as users' opinions may provide feedbacks and inputs for making larger

V.M.R. Muggeo (✉)
Dipartimento Scienze Statistiche e Matematiche "S. Vianelli", Università di Palermo, Palermo, Italy
e-mail: vito.muggeo@unipa.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_9,

changes in policy and programming. Hence this type of evaluation is very useful to practitioners, program planners, and policy makers [25].

A useful model for analysing this type of data is the Rasch model for polytomous (ordinal) item responses [4, 10, 20]. The Rasch model (hereafter RM) constitutes a probabilistic approach to measurement of latent variables or traits, such as the ability to pass a test in Psychometrics or the satisfaction for a provided service in the context of quality assessment. The RM enables us to obtain quantitative measures of the personal satisfaction for each service feature; in particular, the model describes by means of a logistic function, the conditional probability of a category response given the satisfaction level of the student and the perceived quality of the service feature [10].

In this chapter we use an alternative approach based on the mixed model framework, and in particular binomial GLMMs (generalised linear mixed models) [8, 21]: the fixed effects represent the effect of the personal characteristics (such as sex or age) on the satisfaction, the subjects-specific effects instead represent the individual level of satisfaction. The chapter has both a substantive and a methodological aim: extending the binomial GLMM to include an "alternative" link function beyond the standard ones, and obtaining estimates of the satisfaction levels for specific features of students and/or service. For instance, possible questions to be answered are "does the declared satisfaction level depends on gender?" or "which feature of the service is associated with higher satisfaction levels for the student?"

The chapter is structured as follows. Section 9.2 is devoted to the description of data, and in Sect. 9.3 we describe the methods employed in the present analysis. The results are discussed in Sect. 9.4 and some comments and a brief discussion are included in the last section.

9.2 Data

The data used in this chapter concern the responses to four items from a larger questionnaire composed of 23 items. The data come from a sample of $n = 273$ students enrolled at the Faculty of Economics at the University of Palermo in the 2004: the questionnaire was administrated from 20th April to 19th May of 2004. For each item, a four level scale is used to obtain students' ratings of their satisfaction for the service features, i.e. 1 = "absolute dissatisfaction", 2 = "moderate dissatisfaction", 3 = "moderate satisfaction", 4 = "absolute satisfaction". The ordinal responses measure the degree of student's satisfaction for some features (items) of Students Reception Office. The four items analysed in this chapter represent the following service features:

1. i39 = *the staff at the desk are polite*;
2. i26 = *the waiting time at the counter is acceptable*;
3. i34 = *student reception records examination results in acceptable times*;
4. i38 = *the staff at the desk provide clear information*.

Covariates employed in the present analysis refer to some students characteristics as gender, age, frequency of access, and year of matriculation. Students were 21.6 ($\pm$1.8 standard deviation) years old on average, and 42% of them were male. They matriculated in four successive academic years, 2000–2001 (13.2%), 2001–2002 (35.7%), 2002–2003 (26.8%) and 2003–2004 (24.3%). Table 9.1 reports frequency distributions by item and explanatory variables used in this chapter. The frequencies show how the students' satisfaction levels are mostly moderate for the items i39 and i38 as opposed to the satisfaction levels for the items i26 and i34; in fact for items i39 and i38 the most frequent answers fall in the central categories of responses, while for items i26 and i34 the answers fall mostly in the first two categories of responses.

Previous analyses on such data aimed at measuring the perceived quality for the service features and to calibrate the original questionnaire have been carried out via the RM using all the 23 items [3]. However, results from the RM [2], concerning these four items have provided evidence of the so-called "differential item functioning" [9] with respect to some of the variables here considered; that is, the responses appear to depend on some students' characteristics. The mixed model described in the next section is aimed at modeling simultaneously the person-specific and the covariate effects and at providing relevant quantitative measures.

Table 9.1 Frequency distributions of item responses by covariates and row percentages (in italic)

	i39				i26				i34				i38			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Gender																
Female	30	56	61	12	78	51	29	1	117	36	5	1	17	79	57	6
	18.9	*35.2*	*38.4*	*7.5*	*49.1*	*32.1*	*18.2*	*0.6*	*73.6*	*22.6*	*3.1*	*0.6*	*10.7*	*49.7*	*35.8*	*3.8*
Male	17	31	47	19	53	42	16	3	87	19	6	2	12	46	48	8
	14.9	*27.2*	*41.2*	*16.7*	*46.5*	*36.8*	*14.0*	*2.6*	*76.3*	*16.7*	*5.3*	*1.8*	*10.5*	*40.4*	*42.1*	*7.0*
Age																
≤ 21.5	26	46	58	21	70	49	30	2	109	32	7	3	15	63	63	10
	17.2	*30.5*	*38.4*	*13.9*	*46.4*	*32.5*	*19.9*	*1.3*	*72.2*	*21.2*	*4.6*	*2.0*	*9.9*	*41.7*	*41.7*	*6.6*
> 21.5	21	41	50	10	61	44	15	2	95	23	4	0	14	62	42	4
	17.2	*33.6*	*41.0*	*8.2*	*50.0*	*36.1*	*12.3*	*1.6*	*77.9*	*18.9*	*3.3*	*0.0*	*11.5*	*50.8*	*34.4*	*3.3*
Access																
≤ 5	24	44	68	21	72	54	28	3	110	39	6	2	13	64	70	10
	15.3	*28.0*	*43.3*	*13.4*	*45.9*	*34.4*	*17.8*	*1.9*	*70.1*	*24.8*	*3.8*	*1.3*	*8.3*	*40.8*	*44.6*	*6.4*
> 5	23	43	40	10	59	39	17	1	94	16	5	1	16	61	35	4
	19.8	*37.1*	*34.5*	*8.6*	*50.9*	*33.6*	*14.7*	*0.9*	*81.0*	*13.8*	*4.3*	*0.9*	*13.8*	*52.6*	*30.2*	*3.4*
	17.1	*32.9*	*37.9*	*12.1*	*47.1*	*35.8*	*16.3*	*0.8*	*73.8*	*20.4*	*4.6*	*1.3*	*10.0*	*44.2*	*40.4*	*5.4*
Acd year																
2000	10	13	10	3	24	10	2	0	33	2	1	0	6	20	9	1
	27.8	*36.1*	*27.8*	*8.3*	*66.7*	*27.8*	*5.6*	*0.0*	*91.7*	*5.6*	*2.8*	*0.0*	*16.7*	*55.6*	*25.0*	*2.8*
2001	21	29	42	6	44	37	15	2	76	19	3	0	13	52	30	3
	21.4	*29.6*	*42.9*	*6.1*	*44.9*	*37.8*	*15.3*	*2.0*	*77.6*	*19.4*	*3.1*	*0.0*	*13.3*	*53.1*	*30.6*	*3.1*
2002	8	26	29	10	44	20	9	0	57	14	1	1	5	26	38	4
	11.0	*35.6*	*39.7*	*13.7*	*60.3*	*27.4*	*12.3*	*0.0*	*78.1*	*19.2*	*1.4*	*1.4*	*6.8*	*35.6*	*52.1*	*5.5*
2003	8	19	27	12	19	26	19	2	38	20	6	2	5	27	28	6
	12.1	*28.8*	*40.9*	*18.2*	*28.8*	*39.4*	*28.8*	*3.0*	*57.6*	*30.3*	*9.1*	*3.0*	*7.6*	*40.9*	*42.4*	*9.1*

9.3 Methods

Modeling the K-level ($r = 1, 2, \ldots, K$ levels) ordered response Y_{ij} for subject (student) $i = 1, \ldots, n$ for several items ($j = 1, \ldots, J$) as a function of individual factors x_{ij}, say, calls for an ordinal regression model [12, 13],

$$P(Y_{ij} \geq r|x_{ij}) = h(\alpha_r + \gamma_i + x_{ij}^T\beta), \tag{1}$$

where $\alpha_1 > \alpha_2 > \ldots > \alpha_{K-1}$ are the threshold parameters, γ_i are subject specific intercepts and $h(\cdot)$ is a response function relating the linear predictor values onto the scale [0, 1] of the responses. Note that model (1) applies simultaneously to all $r = 1, \ldots, K-1$ cumulative probabilities $P(Y_{ij} \geq r)$, and moreover an identical effect of the predictors for each cumulative probability is assumed: in this case, when $h(\cdot)^{-1} = g(\cdot)$ is the logit function, model (1) is usually referred as Proportional Odds model [13]. The parameters γ_i are, in general, nuisance parameters, and their total number increases proportionally with the sample size. In the traditional Rasch terminology, they measure the the "ability" of the subjects, but in our context the γ_i's represent the individual satisfaction level not affected by possible explanatory factors which are modeled via the β parameters.

9.3.1 The GLMM Framework

Agresti and Lang [1] and Mukherjee [16] propose to translate the problem of modeling ordinal responses into a binary one by collapsing the response categories: let $Y_1^* = I(Y \geq 2), Y_2^* = I(Y \geq 3), \ldots, Y_{K-1}^* = I(Y \geq K)$ be the $K-1$ dummies associated with the K-level ordinal variable Y; then the corresponding regression model for the binary pseudo-response is

$$P(Y_{rij}^* = 1|x_{ij}) = h(\alpha_r + \gamma_i + x_{ij}^T\beta). \tag{2}$$

Model (2) is substantially equivalent to ordinal model (1); interpretation of parameters is unchanged, however estimation is quite different. Model (2) is based on the augmented data set and it uses dichotomic (pseudo) response Y^*. Using the bernoulli likelihood allows to exploit some advantages: there exist sufficient statistics and a conditional likelihood approach may be advocated, provided the canonical link (the logit for binary data) is employed [1, 16]; moreover, from a practical point of view software/code/general optimization facilities for binary likelihood are easy available and they can be a good starting point for writing code to fit more complex models. This leads the binary-outcome model (2) to be preferred to the ordinal-outcome model (1).

We propose a random effect model based on binary model (2) for the pseudo-response Y^*; unlike the conditional likelihood approach the link function $h(\cdot)^{-1}$ does not need to be logit. By using the connection between the binary Rasch model

and the GLMM [21, 23], we suggest to model ordinal data via binary GLMM applied to the binary collapsings of the ordinal scale, i.e.

$$P(Y^*_{rij} = 1|x_{ij}) = h(\alpha_r + \gamma_i + x_{ij}^T\beta) \quad \gamma_i \sim \mathcal{N}(0, \sigma). \tag{3}$$

Estimation of model (3) is carried out by maximisation of the marginal likelihood obtained by integrating out the random effects, i.e. the subject-specific parameters γ_i's. The resulting marginal likelihood depending only on the fixed-effects (α, β) is

$$\text{Lik}(\alpha, \beta) = \prod_{i=1}^{n} \int \prod_{r=1}^{K} \prod_{j=1}^{J} \mu_{rij}^{y^*_{rij}} (1 - \mu_{rij})^{1-y^*_{rij}} f(\gamma_i) \text{d}\gamma_i \tag{4}$$

where $f(\cdot)$ is the (Gaussian) density of the random effects, and $\mu_{rij} = P(Y^*_{rij} = 1|x_{ij})$ is related to the linear predictor via the link function $g(\cdot)$. Unfortunately maximisation of marginal (log-) likelihood (4), with a few simple exceptions, is demanding and unpractical. Several approaches have been suggested, such as Penalized Quasi Likelihood (PQL), Laplace approximation or fully Bayesian methods based on MCMC computations; see for instance Ref. [17].

Given estimates of the fixed effects α, β, "predictions" of the subject-specific γ_i's are obtained via empirical bayes estimates based on the posterior density of the random effects [8, 23].

9.3.2 Alternative Link Functions

A possible limitation of model (1)–(3) concerns the link function $g(\cdot) = h^{-1}(\cdot)$. The choice of the link function is usually overlooked in practice, and a logit link is usually employed for binary responses. However the link function is paramount in item-response analysis as it defines the so called ICC, *Item Characteristic Curve* [10]; therefore the link function is critical for the model interpretation and also it may be important in identifying explanatory variables in the regression equation. It is well known that the binary observed response model are based on the underlying latent variable model

$$\tilde{Y}_i = x_i^T\beta + u_i \qquad i = 1, \ldots, n$$

where we observe the binary variable $Y^*_i = 1 \Leftrightarrow \tilde{Y}_i > s$ for some threshold value s, the u_i's are iid, and the link function for the observed binary response is the quantile function of the u_i's.

Hence it is possible to generalise model (3) by considering alternative link functions beyond the logit which assumes a logistic density for the latent variable. Possible choices are the link probit, complementary log-log, cauchy, and the less known Gosset and Pregibon links [11, 14]. The Gosset (or Student t) link is a one-parameter link function which allows to account for symmetrically distributed heavy tails in the

latent variable; the degrees-of-freedom of the latent t variable represent the parameter of the link function which simplifies to the cauchit link when it is equal to 1.

The Pregibon link function is a two-parameters function based on the generalised Tukey λ family [18, 19]. Its expression is given by

$$g(\mu; a, b) = \frac{\mu^{a-b} - 1}{a - b} - \frac{(1-\mu)^{a+b}}{a + b} \tag{5}$$

where (a, b) are the two parameters regulating the shape of the link itself; more specifically, a and b control the tails' heaviness and the skewness of the latent distribution respectively. Due to its flexibility, we focus only on the Pregibon link, since simpler cases (e.g. logit or complementary log-log) may be covered as a function of particular values of (a, b). For instance, when $a = 0$ the latent distribution is symmetrical and tail weights increase as a decreases; at limiting when a and b tend to zero, the Pregibon link reduces to the standard logit link.

9.4 Results

Model (3) with Pregibon link for fixed (a, b) may be fitted via standard algorithms employed in estimation of mixed models, for instance the Laplace approximation of integral (4) or Penalized Quasi Likelihood (PQL). Despite of that, at the best of our knowledge, we are not aware of any usage involving binary GLMM with non-standard link. We have employed PQL on a grid of values for (a, b) to fit the Pregibon Binary GLMM for item response. Given the estimates $(\hat{a}, \hat{b})$ maximizing such profile (pseudo) likelihood, the binomial GLMM has been fitted assuming $(a, b) = (\hat{a}, \hat{b})$. Relevant R [22] code is available on request from the authors.

Figure 9.1 displays the profile log-likelihood on a slice of plane $a \times b$ with relevant maximising value at $(0.22, -0.24)$. The white bottom right part of the figure refers to very different (extremely lower) log-likelihood values which have been not represented in the plot.

The binomial log-likelihoods are -895.2 and -891.4 for the Logit and the Pregibon link respectively. Table 9.2 shows the parameter estimates for the two binary GLMM's, the logit link and the Pregibon link with $(a, b) = (0.22, -0.24)$.

While findings are mostly unchanged, we observe that a somewhat important difference appears in the results of the age effect. By using a limiting Gaussian distribution for the Wald statistic "Est./SE" which appears reasonable in such large sample, we note that in the logit link model the age may be considered unimportant from an "hypothesis testing" point of view. However, according to the Pregibon link GLMM, older students appear to be significatively associated with higher satisfaction levels. Moreover, males and a low number of accesses seem also to favour high satisfaction levels. While no difference turns out between the baseline item i39 and the item i38, we observe that item i26 and especially item i34 are those with lower satisfaction levels. The relevant Wald-based 95% confidence intervals $((-4.089, -3.521)$ for i34 and $(-2.445, -1.998)$ for i26) are not overlapped, suggesting that the item i34

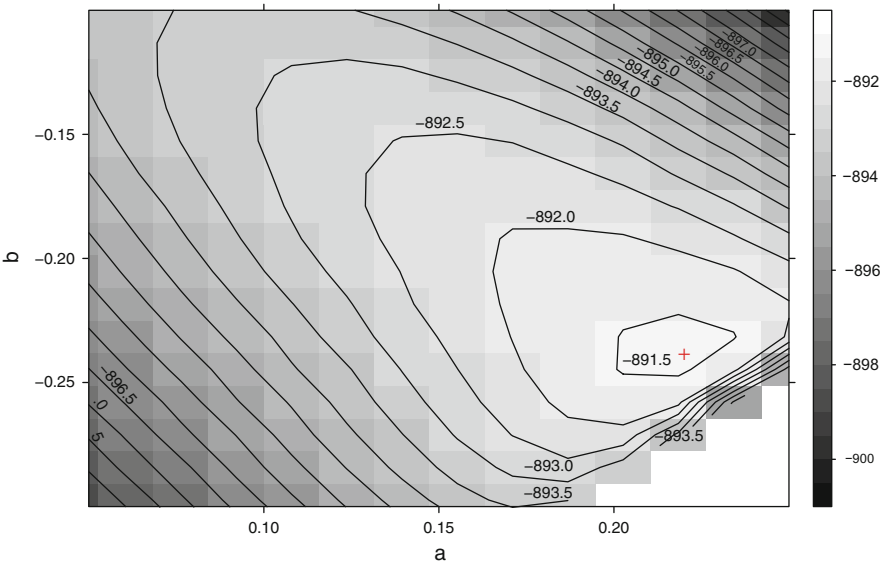


Fig. 9.1 Contour plot of the profile log likelihood for the Pregibon link parameters (a, b). The *cross* identifies the ML estimate (0.22, −0.24)

Table 9.2 Parameter estimates of fixed effects from the Logit and the Pregibon link binary GLMM

	Logit				Pregibon(0.22,–0.24)			
	Est	SE	$\frac{\text{Est}}{\text{SE}}$	p-value*	Est	SE	$\frac{\text{Est}}{\text{SE}}$	p-value*
Intercept	−0.763	1.463	0.52	0.6	−1.112	1.375	0.81	0.4
α_2	−2.393	0.108	22.1	< 0.0001	−2.335	0.092	25.4	< 0.0001
α_3	−5.363	0.171	31.3	< 0.0001	−5.517	0.166	33.3	< 0.0001
Item_{26}	−2.189	0.130	16.8	< 0.0001	−2.222	0.114	19.6	< 0.0001
Item_{34}	−3.786	0.158	24.0	< 0.0001	−3.805	0.145	26.3	< 0.0001
Item_{38}	−0.199	0.117	1.70	0.0892	−0.149	0.096	1.55	0.12
Age	0.115	0.063	1.84	0.0667	0.122	0.059	2.08	0.0389
AA_{01}	0.742	0.323	2.30	0.0224	0.668	0.307	2.18	0.0305
AA_{02}	1.114	0.355	3.13	0.0019	1.061	0.337	3.15	0.0018
AA_{03}	1.783	0.383	4.66	< 0.0001	1.655	0.362	4.58	< 0.0001
Sex_m	0.394	0.205	1.93	0.0553	0.379	0.194	1.96	0.0512
Access	−0.080	0.025	3.26	0.0013	−0.076	0.023	3.25	0.0013
σ_a	1.419				1.361			

* Using a limiting Gaussian distribution.

is actually the item with the lowest satisfaction. This agrees with the findings arose from the standard Rasch analysis where the ranking of the four service feature with respect to the satisfaction is the same to that coming from the binary GLMM.

However, what seems to us quite interesting is the different ICC which turns out from the two different models; Fig. 9.2a illustrates.

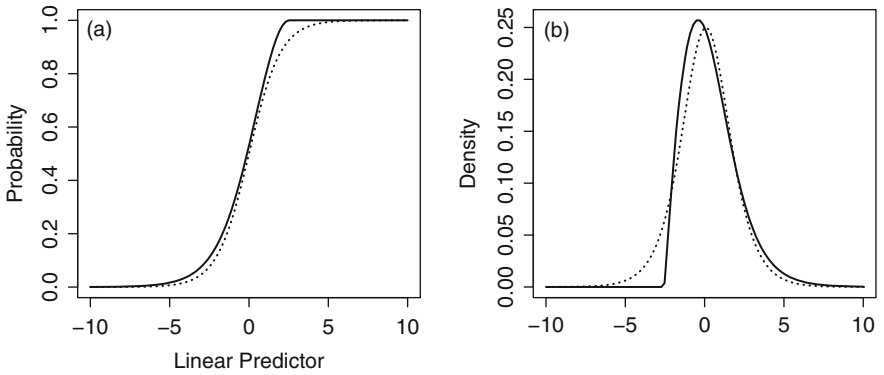


Fig. 9.2 (**a**) Logit (*dotted line*) and fitted Pregibon(0.22, −0.24) (*continuous line*) link functions; (**b**) Underlying distribution of the latent variable, logistic (*dotted line*) and Pregibon(0.22, −0.24) (*continuous line*)

The ICC from Pregibon(0.22, −0.24) reflects the positive asymmetry of the latent distribution which remarkably differs from the logistic one. A direct consequence of the different link functions concerns the fitted values; Fig. 9.3 portrays the scatter-plot of the fitted values from the two binary GLMMs with a somewhat slight prevalence of over-estimated probabilities, in particular in the range 0.4–0.6. Differences in the fitted values (i.e. cumulative probabilities) of the binary pseudo-responses Y^* may be of major concern as we are also interested in the profiles of the students' evaluator. At this aim we have to compute the fitted probabilities on the original ordinal scale via the inverse link for a given covariate vector $\bar{x}$, i.e. $\pi_r(\bar{x}) = P(Y = r|\bar{x})$ for $r = 1, \ldots, 4$. Note that model (3) is actually a model for the cumulative probabilities, therefore the category-specific probabilities have to be computed as a difference between consecutive fitted values, $P(Y_r^*) - P(Y_{r+1}^*) = \mu_r - \mu_{r+1} = P(Y = r) = \pi_r$. In our case it turns out.

$$\pi_4 = \mu_4 \qquad \pi_3 = \mu_3 - \mu_4 \qquad \pi_2 = \mu_2 - \mu_3 \qquad \pi_1 = 1 - \pi_2 - \pi_3 - \pi_4$$

To illustrate, the four response probabilities have been estimated for the item i39, here assumed as the reference among the four service features. Figure 9.4 shows the response probabilities as a function of the underlying personal satisfaction level. As the latent trait gets larger (or smaller), the probability of being absolutely satisfied (or absolutely unsatisfied) tends to 1. The figure emphasizes the differences with respect to the same response probabilities coming from the logit link model which are reported in grey. For the same item i39, we have also computed the probabilities for a number of access and student age fixed at 5 and 21.6 years, respectively. Table 9.3 shows the results for the two aforementioned binary GLMM's.

Results from two models are substantially in agreement, however for some profiles there exist non-negligible differences. This is not surprising, as the fitted values μ_r for some profiles strongly depend on the link function as shown in Fig. 9.3. As it can be seen in Table 9.3, the greater differences are between the fitted probabilities

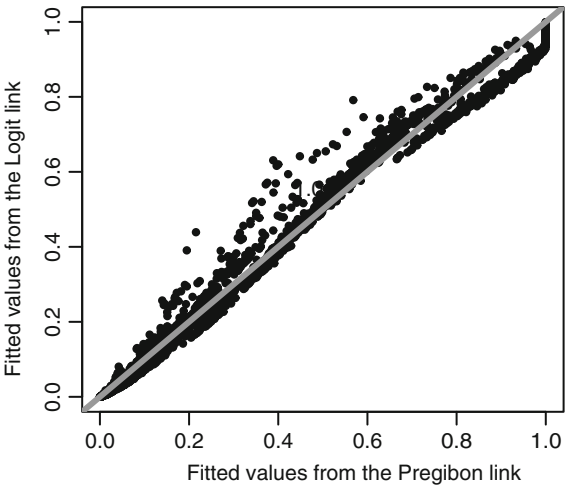


Fig. 9.3 Contrasting fitted values (cumulative probabilities) from the two GLMM's with Logit and Pregibon link

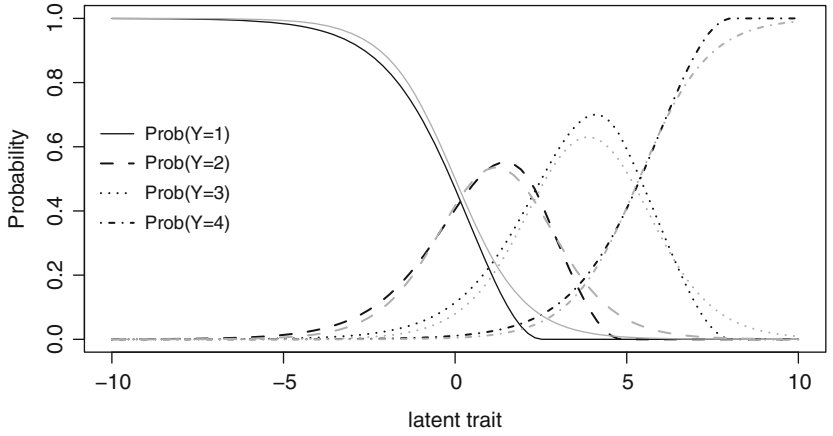


Fig. 9.4 Probability for the four response categories as a function of the satisfaction level (latent trait). *Black lines*: fitted probabilities from the Pregibon link model. *Grey lines*: fitted probabilities from the Logit link model

of the first category of response (item i39). The logit link provides higher estimates of the probability of $Y = 1$ (being absolutely unsatisfied) with respect to the Pregibon link: when academic year equal to 2002 and 2003 the fitted probabilities from the Logit model are quite different from those returned by the Pregibon link model, especially for the females. We have also computed probabilities for i34, the item with lowest satisfaction level. Here the largest differences between Logit and Pregibon pertain the categories "moderately satisfied" and "absolutely satisfied": the logit link yields fitted probabilities lower than those from the Pregibon link.

Table 9.3 Fitted probabilities of ordinal responses by gender and enrolment (academic) year from Logit and Pregibon link GLMM (at n. of accesses = 5 and age = 21.6) for item i39

			Logit				Pregibon(0.22, −0.24)			
Profiles	Gender	AA	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$
1	M	2000	0.153	0.511	0.311	0.025	0.104	0.556	0.305	0.036
2		2001	0.079	0.405	0.464	0.052	0.013	0.497	0.430	0.060
3		2002	0.056	0.337	0.534	0.074	0.000	0.412	0.507	0.081
4		2003	0.029	0.219	0.617	0.134	0.000	0.260	0.615	0.125
5	F	2000	0.211	0.534	0.238	0.017	0.183	0.546	0.243	0.027
6		2001	0.113	0.469	0.383	0.036	0.055	0.543	0.357	0.045
7		2002	0.081	0.409	0.460	0.051	0.012	0.495	0.433	0.061
8		2003	0.043	0.286	0.576	0.095	0.000	0.357	0.549	0.095

9.5 Conclusions

We have presented a binomial GLMM framework to model ordinal item responses as a function of subject-specific "abilities", item and individual covariates. Like the RM, the parameters in the GLMM have a "subject-specific" interpretation. The random effects in the GLMM may be interpreted as the baseline satisfaction of the students, adjusted for the effects of covariates. This is analogous to the RM where the person location parameters are estimates of the satisfaction levels expressed by the subjects, without considering any person characteristic or covariate [10]. However with respect to the RM, the GLMM framework appears to be more flexible, and moreover possible concordant responses for each subject do not need to be discarded like in the classical RM. In the GLMM framework we have investigated possible impact of the choice of the link function, and other possible differences and extensions are discussed in [8]. For instance we have simultaneously modeled latent traits, and effects of student- and item-specific features without employing the so-called DIF analysis [6].

The link functions with one or two-parameters, such as the Pregibon link, may be considered as a "middle point" between the standard links with no parameter (such as the logit, probit or complementary log-log) and the very flexible unspecified link models, where no parametric function is assumed and some nonparametric smoothers are employed for estimation, see [15] for instance.

Although from a statistical point of view (in terms of likelihood values), better results are achieved using the Pregibon, a possible limitation concerns the interpretation of the parameters which are *not* log odds ratios in the Pregibon link model; however they still measure the effect size of the covariates and sometimes such information will suffice; moreover a more "appropriate" link (as measured by the likelihood) is expected to provide better estimates of the probabilities of the satisfaction levels.

References

1. Agresti A, Lang JB (1993) A Proportional odds model with subject-specific effects for repeated ordered categorical responses. Biometrika 80:527–534
2. Aiello F (2005) Il modello di Rasch per la costruzione di uno strumento di misura della qualitá di un servizio. PhD thesis, University of Palermo, Italy
3. Aiello F, Capursi V (2008) The Rasch model performance to assess a University Service according to students' opinion. Appl Stochastic Models Bus Ind 24:459–470
4. Bond TG, Fox CM (2007) Applying the Rasch model: fundamental measurement in the human sciences, 2nd edn. Lawrence Erlbaum Associates Publisher, Mahwah, NJ
5. Capursi V, Ghellini G (2008) Dottor DIVAGO: Discernere, Valutare e Governare la nuova Università. Collana dell'Associazione Italiana di Valutazione: 1900.2.4 - Teoria, metodologia e ricerca. Franco Angeli, Milano
6. Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, Hays RD, Teresi JA (2007) A comparison of three sets of criteria for determining the presence of differential item functioning using logistic regression. Qual Life Res 16:69–84
7. Cronin JJ, Taylor S (1992) Measuring service quality: a re-examination and extension. J Mark 56:55–68
8. Doran H, Bates D, Bliese P, Dowling M (2007) Estimating the multilevel Rasch model with the lme4 package. J Stat Soft 20(2):1–18
9. Fayers PM, Hand DJ (2002) Casual variables, indicator variables and measurement scales: an example form of quality of life. J R Stat Soc A 165:233–261
10. Fischer GH, Molenaar IW (1995) Rasch models - foundations, recent developments, and applications. Springer, New York, NY
11. Koenker R (2006) Parametric links for binary response. R News 6:32–34
12. McCullagh P (1980) Regression models for ordinal data. J R Stat Soc B 42:109–142
13. McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall/CRC, Boca Raton, FL
14. Morgan BJT, Smith DM (1992) A note on Wadley's problem with overdispersion. Appl Stat 41:349–354
15. Muggeo VMR, Ferrara G (2008) Fitting generalized linear models with unspecified link function: a P-spline approach. Comput Stat Data Anal 52:2529–2537
16. Mukherjee B, Ahn J, Liu I, Rathouz PJ, Sanchez BN (2008) Fitting stratified proportional odds models by amalgamating conditional likelihoods. Stat Med 27:4950–4971
17. Ng ESW, Carpenter JR, Goldstein H, Rasbash J (2006) Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. Stat Model 6:23–42
18. Pregibon D (1980) Goodness of link tests for generalized linear models. Appl Stat 29:15–24
19. Prentice RL (1976) Generalization of the probit and logit methods for dose response curves. Biometrics 32:761–768
20. Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Danish Institute of Educational Research, Copenhagen
21. Rijemen F, Tuerlinckx F, DeBoeck P, Kuppens P (2003) A nonlinear mixed model framework for item response theory. Psychol Methods 8:185–205
22. R Development Core Team (2008). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
23. Skrondal A, Rabe-Hensketh S (2004) Generalized latent variable modeling. Chapman & Hall, Boca Raton, FL
24. Weiss CH (1988) Evaluation: methods and studying program and policies, 2nd edn. Prentice Hall, Upper Saddle River, NJ
25. Weiss CH (1998) Have we learned anything. Am J Eval 19(1):21–33

Chapter 10
Analyzing Undergraduate Student Graduation Delay: A Longitudinal Perspective

Paola Costantini and Maria Prosperina Vitale

10.1 Introduction

In Italy the number of years in which undergraduate students should complete their education programme is established by law. However, many students obtain their degree after the expected time: a well-known issue affecting numerous Italian universities. In 1999, therefore, the Italian government introduced a reform that, among other aims, intended to reduce the gap between the average number of years in which a student completed the education programme and the official deadline established by the university regulations.

Recent statistical reports show that, after the introduction of the reform in the academic year 2001/2002, the number of undergraduate students in 2006 who succeeded in completing the full length first level degree course in 3 years increased significantly with respect to the past [1, 9, 10]. Despite this improvement, many students who enrolled after the reform still have an irregular university career and a fair number fail to earn any credits in their first year of enrolment.

Consequently, there is a need for a more detailed analysis into "graduation delay", i.e. the phenomenon of students enrolled in the final year of a higher education degree course who fail to complete it successfully during the reference year, regardless of their age [17]. Using different delay indicators, several models have been proposed in the literature to explain the determinants of the delay in student careers of the Italian university system (see e.g. [4, 8, 19]). However, they perform a cross-sectional analysis on longitudinal data, producing a potential loss of information.

For this reason, we aim to introduce a longitudinal statistical methodology that allows researchers to investigate patterns in graduation delay and its potential determinants. For our purposes, we first define a graduation delay longitudinal indicator that measures the student delay at different time points. In this way, a delay pattern may be analysed and interpreted.

P. Costantini (✉)
Dipartimento di Economia, Università di Cassino, Cassino, Italy
e-mail: p.costantini@unicas.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_10,

Furthermore, we propose to use Latent Curve Models (LCM) in order to study the systematic changes over time of this student career delay indicator. The LCM approach was developed in the framework of Structural Equation Modelling in order to model individual growth trajectories for repeated measured data, summarizing many curves in a single average trajectory [6, 13, 16]. In addition, they provide a means for testing the contribution of other variables or constructs in order to explain variability in initial levels and in patterns of growth [14, 18].

This methodological approach will be presented through the study of the careers of an undergraduate student cohort enrolled at an Italian university. We provide a linear unconditional LCM to evaluate whether any significant pattern is present in the observed delay indicator. Then a conditional LCM model, with socio-demographic and educational background covariates, is estimated in order to investigate any difference in student career trajectories.

The chapter is organized as follows. In Sects. 10.2 and 10.3 the graduation delay issue related to student careers in the Italian university system, as reported in official documents and in literature, is briefly reviewed. The proposed new indicator of graduation delay is offered in Sect. 10.4. In Sect. 10.5, the main characteristics of Latent Curve Models are described, while in Sect. 10.6 the data source and the estimated models are discussed for the analysis of data concerning the careers of an undergraduate student cohort enrolled in an academic field at Cassino University. Some concluding remarks are provided in Sect. 10.7.

10.2 The Graduation Delay Issue

Graduation delay has connoted the Italian university system for years, especially if compared with the other Organisation for Economic Co-operation and Development (OECD) countries. This phenomenon regards students enrolled in the final year of a higher education degree course who fail to complete it successfully during the reference year [17].

In order to tackle this and other issues, the Italian Higher Education system has undergone substantial changes over the past 10 years following the first large reform introduced in 1999. The main innovations were the introduction of first and second level university degrees and the adoption of the European Credit Transfer System (ECTS) to evaluate the students' learning activities.

To monitor the effects of this reform, official reports by the OECD and the Italian National Committee for the Evaluation of the University System (CNVSU) were made available. As discussed in the recent OECD document *Education at a Glance 2007*, in Italy the percentage of individuals who succeed in graduating has more than doubled from 19.0% in 2000 to 41.0% in 2005. This increase is largely attributed to the Italian university reform that now allows university students enrolled in short programmes to obtain a degree in 3 years.

The VIII document of the CNVSU [10] reports some interesting results concerning student career performance in terms of both "process indicators" and "outcome

indicators". Considering the process indicators (see e.g. *drop-out rate*, *duly performing students* and so on), the percentage of students with a regular career out of the total number of enrolled students is 57.7%. This percentage rises to 71.0% if we consider only the students enrolled after the reform in a first level degree. Therefore, the percentage of undergraduate students after the reform with an irregular career is still 29.0%. In addition, we find that 15.5% of the students fail to earn any credits in their first year of enrolment. The percentage of students not enrolled in the second year is still about 20.0% on average.

With respect to the outcome indicators (see e.g. *average age at graduation*, *number of years to obtain a degree* and so on), the document considers the duration of student careers to complete an education programme. The percentage of students involved in a first level degree who completed the programme in the standard number of years specified in the regulations was 30.3% in 2006, while 54.3% completed the education programme within 1 or 2 years after the 3 years established. The average number of years to obtain a first level degree is 4.4 with a median of 3.6. With respect to the students enrolled before the reform on 4, 5 or 6 year programmes, there has been an improvement: in 2000 the percentage of students enrolled before the reform who obtained a degree within the expected time was respectively 1.8, 5.0 and 29.9%, while the average duration of programmes was about 8 years for all programmes. Finally, the number of graduates who were awarded the degree within the expected time out of the total number of graduates highlights that 30.3% obtained a degree within the 3 years provided by law, while for the students enrolled before the reform on 4, 5 or 6 year programmes, this ratio shows a worse student career performance (with only 4.1% of students completing the education programme within the standard number of years).

From this perspective, the graduation delay issue was only partially resolved in Italy with the introduction of the 1999 University reform. There is therefore still a need for detailed study of this phenomenon and its determinants, and with this aim we first introduce a graduation delay indicator and then propose to examine its determinants through a model for longitudinal data analysis.

10.3 Measuring and Analyzing Graduation Delay

In this section, focusing on the graduation delay measurement issue, we briefly look at the indicators proposed in the research reports of Almalaurea [1] and CILEA [9] projects to study irregular university careers of Italian undergraduate students. We also present some models proposed in the Italian literature to explain the determinants of the delay in student careers from a cross-sectional perspective [4, 8, 19].

The *graduation delay* is computed by Almalaurea [1] as the difference between the number of years in which a student completed the education programme and the official deadline established by university regulations. A delay *indicator* is defined as the ratio of graduation delay out of the standard number of years provided by regulations in which a student can complete the education programme.

The graduation delay fell from 2.9 in 2001 to 1.7 years in 2007 (registering a decrease of 1.2) and the age at graduation decreased from 28 to 27 years. The introduction of first level degrees has involved a reduction in the number of years to obtain a degree, estimated by Almalaurea as falling from 4.4 in 2001 to 3.7 years in 2007. From this point of view, the improvement that occurred between 2001 and 2007 is relevant: the percentage of students who successfully completed the course programme more than trebled (from 10.2 to 37.9%), but for 26.7%, the time to graduate is still 3 years or more above the standard. The Almalaurea delay indicator from 2001 to 2007 shows a decreasing pattern: its value has slipped from 0.69 in 2001 to 0.45 in 2007.

A *survival indicator*, described in the Stella report [9], is defined as 1 plus the ratio of the graduation delay of standard programme duration. This indicator measures the actual duration of student careers related to the standard duration of an education programme provided by the regulations. For students who successfully complete the degree, this indicator assumes values close to 1. In the period 2004–2006, Stella estimated that about 50% of students successfully completed the degree programme in 3 years, while about 20% took more than three years after the legal programme duration to complete the degree.

While the Almalaurea and Stella indicators considered a university programme as the basic statistical unit, some authors defined indicators that measure the delay for each student. In this way, they were able to evaluate determinants of delay in terms of the students' features.

Boero et al. [4] estimated an econometric model in order to identify the determinants of student progression. The proportion of credits achieved by a cohort of students, enrolled for the first time in the 2001/2002 at the faculties of Viterbo and Cagliari Universities, is computed as the number of accumulated credits earned at the end of December 2003 by each student out of the total number of credits that should have been earned at the end of the second academic year.

Schizzerotto and Denti [19] defined "irregular students" as students enrolled at the first level degree courses on the academic year t_0, who on February 1st of the next academic year had not earned at least 95% of the credits specified in the regulations. With this strict definition, they performed a logistic regression model for a cohort of students enrolled at Milan University in the academic year 2001/2002. In order to explain the probability of their being in a position of regularity at the end of the second year, the authors considered as possible determinants certain socio-demographic variables (*gender*, *enrolment*, *final secondary school grade* and *type of school attended prior the university*, *residence*).

Finally, Boscaino et al. [8] evaluated the undergraduate student careers/delay in terms of credits earned over the 4 years after enrolment. Careers are assessed according to the number of credits earned each year. The authors defined 4 threshold values of ECTS (one for each academic year) in order to consider different groups of irregular students. In brief, student profiles are evaluated based on the number of times they exceed the threshold values (i.e. from 0 to 4 times). Students are assumed to be "suffering" if they do not exceed at least one threshold. On this basis, three

categories of students were defined: *positive performer* (a student who has suffered at most once), *wavering performer* (a student who has suffered twice), and *negative performer* (a student who has suffered at least 3 times). They applied their analysis to the student cohort enrolled in the academic year 2001/2002 at the Humanities Faculty of the University of Palermo, analysed retrospectively from the academic year 2004/2005. A proportional odds logit model is exploited to explain student suffering with respect to socio-demographic characteristics, secondary school performance, and career information at the time of enrolment. In addition, they analyzed the cumulate earned credits through Zero-Inflated models to take into account the presence of students who earned zero credits within the 4 years.

In these studies, it emerges that high age at enrolment time usually exerts a negative influence on student career progression; female students are faster in the university career than male students; students who did not attend a "lyceum" type secondary school have a worse performance and the proportion of earned credits increases in correspondence to a high final secondary school grade.

Analysis of the academic performance of undergraduate students also has a long tradition in other countries [15, 20, 21]. These analyses have focused, among other determinants, on the influence of such factors as *age*, *gender*, and *prior educational attainment* of students on university degree performance. In particular, some papers have highlighted the significant effect of these personal characteristics on students' level of performance at university.

We note that in all these studies the problems relating to career progression at University are essentially analyzed by means of methods suitable for cross-sectional data and, in so doing, they miss potentially useful information.

10.4 Defining a Longitudinal Graduation Delay Indicator

The assessment of student careers may gain further insights if a longitudinal perspective is adopted in both the graduation delay measurement and in the corresponding statistical analysis. For this reason, we define a delay indicator that provides not a single measure for each student as reported elsewhere, but as many measures as there are time occasions. Hence the researcher is able to describe delay trajectories and to analyze their determinants over time. We propose to define a delay measure considering the number of ECTS obtained at the end of the first, second and third academic years. In particular, we define the *graduation delay indicator* y, computed as the difference between the total number of credits required per year by the university rules and the actual value of credits earned by each student. For the i-th student at time t, it is thus defined as:

$$y_{it} = ECTS_t^e - ECTS_{it}^o$$

where $ECTS_t^e$ is the total number of credits required per year to obtain a first level degree and $ECTS_{it}^o$ is the observed total number of credits earned by the i-th

student. In particular, given the Italian universities' rules for the first level degree, we will have:

$$y_{i1} = 60 - ECTS^{o}_{i1}$$
$$y_{i2} = 120 - ECTS^{o}_{i2}$$
$$y_{i3} = 180 - ECTS^{o}_{i3}$$

For a student who earned the due number of credits in each year, this indicator will assume a value of zero over the three time occasions. On the other hand, "irregular" students will be characterized by positive values of these indicators: the higher the value, the greater its delay. We note also that for students who are "regular" in their delay, the three values will define a linear increasing trajectory. Students with some delay in the first year who are able to overcome it in the second and third years will show a decreasing trend. A few students may present some negative values for these indicators if they are able to earn more credits than expected in a given year (this may occasionally occur).

10.5 Latent Curve Model to Monitor Student Careers

In this chapter, we propose to analyze the student delay patterns by means of the so-called Latent Curve Model (LCM). It allows the study of trajectories and is thus suitable to explain delay patterns at the individual level. In addition, it also provides a means for testing the contribution of other covariates in order to explain variability in initial levels and in patterns of growth [14, 18]. This kind of model and its variations have already been used as a tool for evaluating student performance [2, 3, 11, 12]. However, their potential has not yet been exploited to study delay within a longitudinal data framework.

In detail, LCMs are analogous to random effect models (also called multilevel models or hierarchical linear models), in which within-person variations are allowed at the first level (due to intra-individual change over time), whereas between-person variations are estimated at the second level (due to intra-individual differences). The repeated measures are thus nested within persons. The fixed effects are represented by means of the intercept and slope factors whereas the random effects are the intercept and slope factors. As a result of individual differences in these intercepts and slopes, changes occur in the relationships between individuals' data over the different time intervals. Because each individual has their own growth trajectory, time is nested within the individual. This is referred to as the within individual level or the Level 1 model [16, 23] specified as Eq. (1):

$$y_{it} = \eta_{0i} + \eta_{1i}\lambda_t + \epsilon_{it} \tag{1}$$

where y_{it} represents the measure of the response variable y for the i-th subject at time t; the random effects η_{0i} and η_{1i} are the regression coefficients at the individual level; λ_t is a parameter that can be either fixed or estimated (generally it represents

time occasions) and finally ϵ_{it} is the residual for individual i at occasion t. In Latent Curve Models, intercept and slope are random variables, with their variation over individuals modeled in the so-called Level 2 model (2):

$$\begin{aligned} \eta_{0i} &= \alpha_0 + \zeta_{0i} \\ \eta_{1i} &= \alpha_1 + \zeta_{1i} \end{aligned} \tag{2}$$

with the individual intercept η_{0i} equal to the mean of the intercepts for all cases α_0 plus a random variation; similarly η_{1i} represents the individual slope as the sum of the mean of the slope for all cases α_1, plus a random variation. The random variations ζ_{0i} and ζ_{1i} are disturbances with means of zero and are uncorrelated with the ϵ_t. The variances of the random coefficients provide measures of the extent to which individuals vary in each change feature. The covariance between the coefficients represents the linear relationship which exists between the individual intercepts and slopes. The units of the first level model are the time occasions, while the units of the second level model are the subjects. In this respect, it is possible to specify exactly the same model as an LCM or a multilevel regression model [22].

Finally, covariates can be inserted in the level 1 and/or level 2 equations, obtaining a conditional latent curve model. By means of an example, we will show below a conditional level 2 model to explain the effect of some determinants on the graduation delay of a student cohort.

10.6 A Case Study: The Delay Patterns of a Cohort of Undergraduate Students

To show the LCM potential in the analysis of the graduation delay, we apply our idea within a case study. We analyse the careers of a cohort of students enrolled at an Italian University.

The data we consider come from the internal database of the University of Cassino and concern the university career of a cohort of students who enrolled in the undergraduate programme in Economics in the academic year 2001–2002. The data are collected for 132 students across 4 years. 50% of the enrolled students are females; the average age at enrolment year is 20.5 (with a median equal to 19) and 64% have obtained a professional or a technical secondary school degree. The average of the final secondary school grade is 77.1/100 (with a median equal to 75) and only 37% have reported a final grade of more than 80/100. Among them, the percentage of students who successfully completed the degree programme in 3 years according to the duration specified by the regulations is 7%.

For our data, the values of the *graduation delay indicator* defined in Sect. 10.4 is depicted for four students in Fig. 10.1 through individual plots. In each plot, the horizontal axis represents the 3 years under analysis, and the vertical axis represents the delay indicator defined above. Different students show different patterns. In the plot (left to right, top to bottom), the delay of the first student is close to the zero line (this is one of the "best students"); the delay of the second and third students

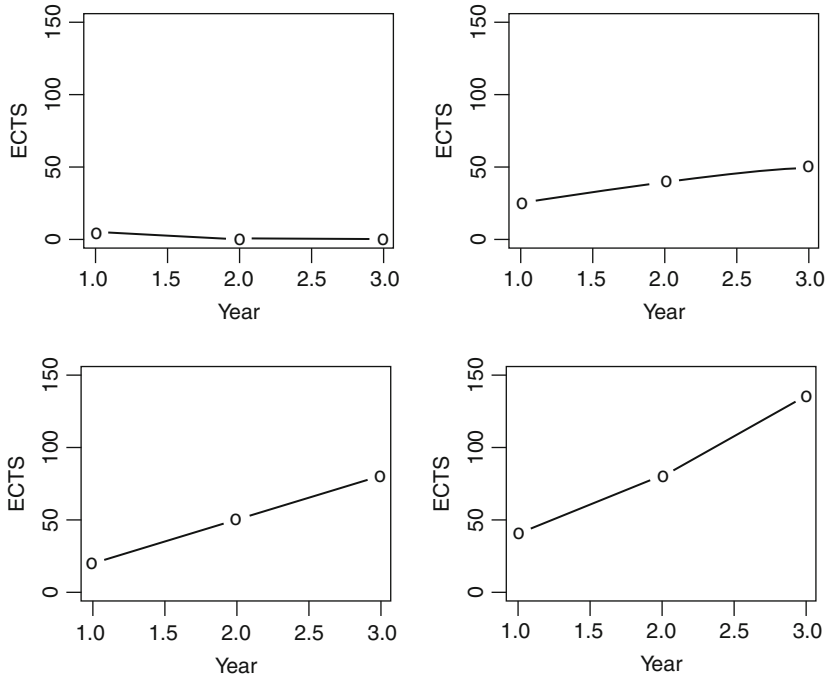


Fig. 10.1 Values of the *graduation delay indicator* at the end of first, second and third academic years for four students. In the plot (*left to right*, *top to bottom*), the delay of the first student is close to the zero line; the delay of the second and third students follows a linear trajectory (with different *slopes*); while the last plot shows a certain degree of nonlinearity for the fourth student

follows a linear trajectory (with different slopes); while the last plot shows a certain degree of nonlinearity for the fourth student.

In Fig. 10.2 we represent the distributions of the *graduation delay indicator* for all the students year by year. Over the three years, the difference between the credits actually earned with respect to those that are required by the regulations is generally increasing.

Given the information available within our official database, we investigate to what extent the graduation delay may be explained by demographic characteristics. In the conditional box-plots, showing the distribution with respect to *gender*, *age at enrolment*, *type* and *final grade secondary school*, the largest differences appear to be related to performance at secondary school (Fig. 10.3).

10.6.1 A Conditional Linear Latent Curve Model

In the following, by means of an LCM analysis, we pursue two main aims. First, we look for a model that may summarize all the 132 curves in a single average trajectory. Second, we wish to evaluate the difference in trajectories in terms of some

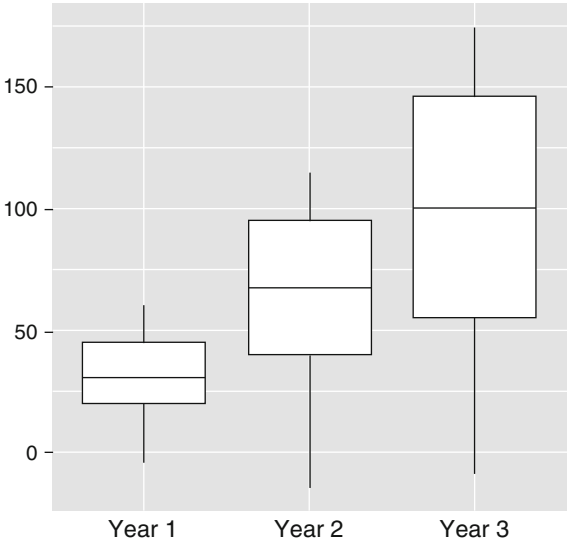


Fig. 10.2 *Box-plots* of the *graduation delay indicator* at the end of the first, second and third academic year for the 132 enrolled students

demographic and previous student career covariates. With respect to the Eq. (1), our delay measure y will then be the differences between expected and actual ECTS (year by year) for each student. We note that our graduation delay indicator is a discrete variable. However, given that the observed distributions of y_{i1}, y_{i2}, y_{i3} are not clustered around few values we decide to deal with y as if it were a continuous variable.

We assume a linear trend and hence the time points λ_t are equally spaced and set to be equal to 0, 1, 2 (with zero as the starting time, that is the end of the first year in the University programme). Furthermore, to gain more insights into this model, we consider a conditional LCM model, with covariates in the second level equation. The covariates taking part in our analysis are listed in Table 10.1.

Assuming a linear trajectory, we first estimated an unconditional linear model[1] in which covariates that may affect the trajectory are not included.[2] The unconditional model fits our data very well and this led us to consider a more complex model. Then we proceeded to estimate a conditional linear model whose level 1 and level 2 equations are defined respectively as:

$$y_{it} = \eta_{0i} + \eta_{1i}\lambda_t + \epsilon_{it} \tag{3}$$

$$\begin{aligned} \eta_{0i} &= \alpha_0 + \beta_{01}Gender + \beta_{02}SSCert + \beta_{03}SSGrade + \beta_{04}Age + \zeta_{0i} \\ \eta_{1i} &= \alpha_1 + \beta_{11}Gender + \beta_{12}SSCert + \beta_{13}SSGrade + \beta_{01}Age + \zeta_{1i} \end{aligned} \tag{4}$$

[1] All the LCM analyses presented in this work have been performed using the Lisrel software 8.8 version.

[2] Related goodness-of-fit statistics are offered in Table 10.2.

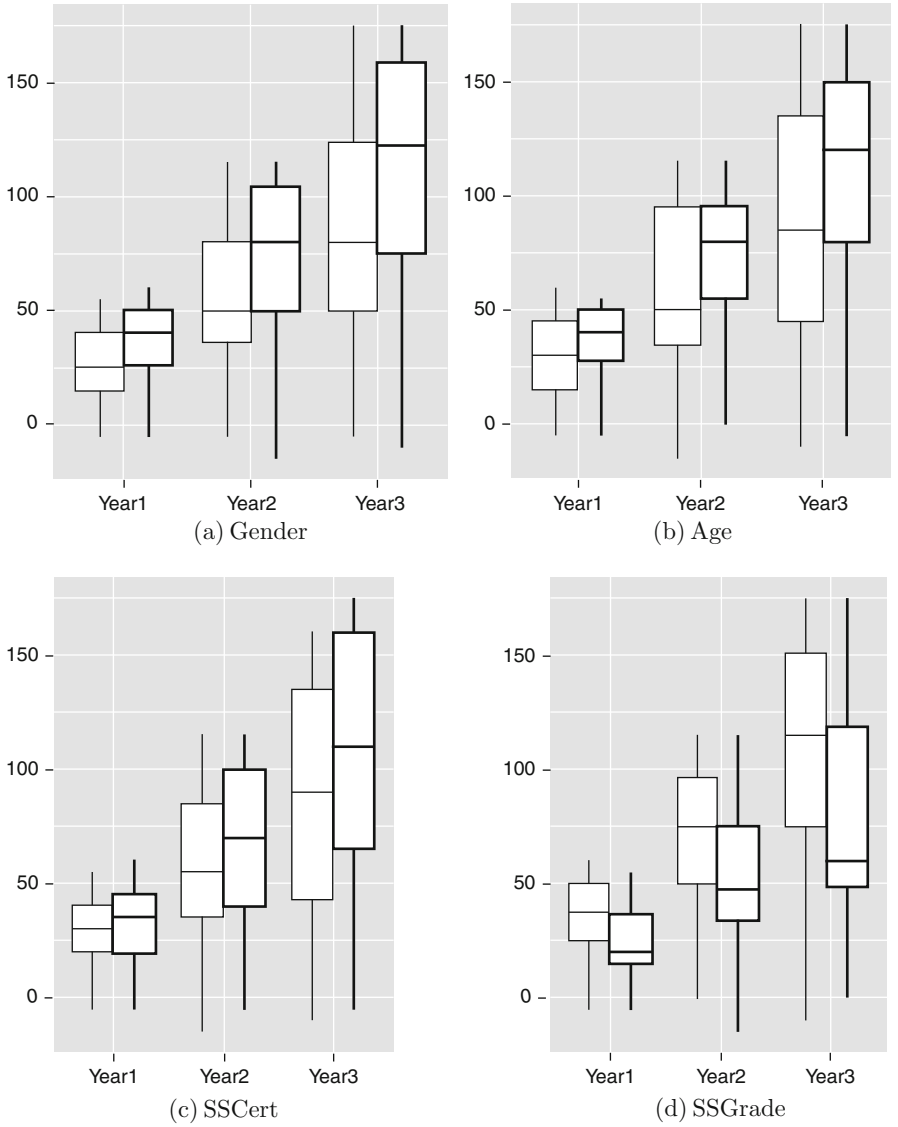


Fig. 10.3 Conditional *box-plots* of the *graduation delay indicator* at the end of first, second and third academic years for the 132 enrolled students. (**a**) *gender*: normal = female, bold = male; (**b**) *age at enrolment*: normal = $\leq$ 19, bold = $>$ 19; (**c**) *secondary school certificate*: normal = "lyceum", bold = "non lyceum"; (**d**) *secondary school final grade*: normal = $\leq$ 80, bold = $>$ 80

Table 10.1 Covariates entering the conditional linear model

Variable	Label	Measurement	Value
Gender	Gender	Binary	0 = female 1 = male
Enrolment age	Age	Continuos	Min = 18 max = 38
Secondary school certificate	SSCert	Binary	0 = not lyceum 1 = lyceum
Secondary school final grade	SSGrade	Continuos	Min = 60 max = 100

Table 10.2 Unconditional model, conditional full model and sub-model: goodness of fit

	Model1	Model2	Model3
	Unconditional model	Conditional full model	Conditional sub-model
χ^2 (df)	0.1265 (1)	6.9316 (5)	17.4765 (15)
p-value	0.7220	0.2258	0.2912
RMSEA	0.0000	0.0543	0.0355
SRMR	0.0000	0.0135	0.0460

The conditional full model seems to fit our data well, according to the goodness-of-fit statistics. The χ^2 is equal to 6.9316 (df = 5, p-value = 0.2258), the Root Mean Square Error of Approximation (RMSEA) is equal to 0.0543, and the Standardized Root Mean Square Residual (SRMR) is equal to 0.0135. Parameter estimates, standard errors (SE) and t-values for the full model are offered in Table 10.3. Given that within this model some parameters were not significant[3] we estimated a sub-model that was accepted on the basis of the χ^2 statistic comparison (p-value = 0.3940). The sub-model χ^2 statistic ($\chi^2 = 17.4765$, df = 15) and the corresponding p-value (0.2912), along with the RMSEA equals to 0.0355 and the SRMR index equals to 0.0460, lead us to accept it for our data. The obtained sub-model, for which all the parameters are significant at a 5% level, is summarized through a path diagram in Fig. 10.4. The diagram shows the response variable y with each occasion influenced by the intercept and slope factors, which are influenced by the direct effects of the covariates.

Parameter estimates, standard errors (SE) and t-values for the sub-model are offered in Table 10.4. According to our results, trajectories will vary across students in different intercepts and slopes. The significance of the intercept variance implies that at the end of the first year students already present significantly different delays. We observe a significant variance of the slopes, which means that students have different linear patterns in their delays. The covariance between intercept and slope is positive and significant. This implies that students starting with a higher delay tend to cumulate delays to a greater extent.

As for the effect of the determinants on the intercepts, we find that male students and students who are enrolled at an older age, start their career with a higher level of delay and that this difference remains constant over time. Furthermore, students

[3] The variance of ϵ_1 is negative and not significant. This may be due to a very little variation for the graduation delay measure in the first year. For this reason we fixed this residual variance equal to 0. This is justified if the negative variance is small (see e.g. [7]).

Table 10.3 Estimated coefficients for the conditional linear model on the *graduation delay indicator*. β parameters refer to Eq. (4)

Parameter	Estimate	SE	T-value
Trajectory means			
Intercept α_0	31.5690	1.3764	22.9356
Slope α_1	32.6411	1.7275	18.8951
Trajectory variances/covariance			
Intercept [VAR (η_0)]	277.6915	37.2232	7.4602
Slope [VAR (η_1)]	396.0450	49.9976	7.9213
Int-Slope [COV(η_0, η_1)]	225.0376	35.3693	6.3625
Covariance among covariates			
SSGrade-age	0.0409	0.1544	0.2646
SSCert-gender	0.0377	0.0199	1.8900
SSCert-age	−0.1192	0.1516	−0.7860
SSGrade-gender	−0.8411	0.4801	−1.7521
SSGrade-age	−7.8728	3.7095	−2.1224
SSGrade-SSCert	−0.4952	0.4666	−1.0612
Beta coefficient Intercept			
B_{01} (gender)	0.0075	0.0030	2.4740
B_{02} (age)	0.0260	0.0217	1.2001
B_{03}(SSCert)	0.0002	0.0028	0.0700
B_{04} (SSGrade)	−0.2078	0.0722	−2.8781
Beta coefficient slope			
B_{11} (gender)	0.0013	0.0026	0.5011
B_{12}(age)	0.0260	0.0193	1.3470
B_{13} (SSCert)	−0.0052	0.0025	−2.0650
B_{14} (SSGrade)	−0.0722	0.0620	−1.1645
Variance of errors			
VAR (ϵ_1)	−25.8073	16.4618	−1.5677
VAR (ϵ_2)	47.8657	18.1895	2.6315
VAR (ϵ_3)	14.3049	54.6097	0.2619

with a higher secondary school final grade start better and this difference remains constant over time. This latter estimated β coefficient $(-0,29)$ means that if student A has ten-points more than student B in the secondary school final grade, then student A will have 2.9 credits less in his delay. Considering the effects of the covariates on the slopes, we note that the students who attended a "lyceum" secondary school present a lower increasing delay trajectory with respect to the others. Finally, we note that there is no effect of gender, the secondary school final grade and age at enrolment on the slope.

Finally, it is relevant to notice that the estimated parameter values are quite small with respect to the total variation in Intercept and Slope in the estimated model. This is probably due to having disregarded some variables that may describe considerable individual characteristics in the graduation delay (e.g. *number of hours devoted to study*, *method of study*) not available for our analysis. We recall that we used an

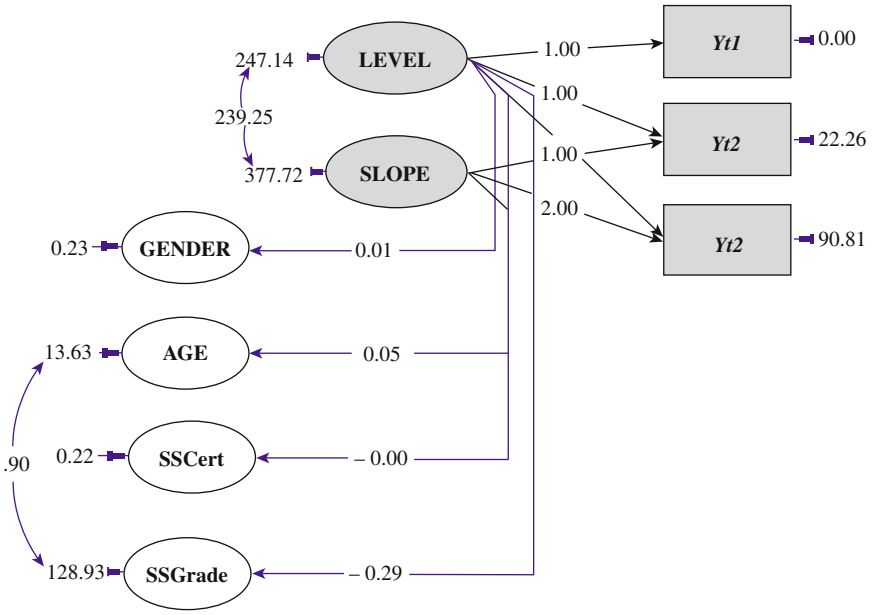


Fig. 10.4 Path diagram with estimated coefficients for the conditional linear sub-model on the *graduation delay indicator*

Table 10.4 Estimated coefficients for the conditional linear sub-model on the *graduation delay indicator*

Parameter	Estimate	SE	T-value
Trajectory means			
Intercept α_0	31.6288	1.3735	23.0277
Slope α_1	32.6999	1.7231	18.9770
Trajectory variances/covariance			
Intercept VAR (η_0)	247.1360	30.5362	8.0932
Slope VAR (η_1)	377.7226	48.0803	7.8561
Int-Slope [COV(η_0, η_1)]	239.2450	34.2157	6.9923
Covariance among covariates			
SSGrade-age	−7.8972	3.7271	−2.1189
Beta coefficient intercept			
B_{01} (gender)	0.0092	0.0027	3.4399
B_{02} (age)	0.0537	0.0205	2.6182
B_{04} (SSGrade)	−0.2905	0.0631	−4.6035
Beta coefficient slope			
B_{13} (SSCert)	−0.0049	0.0021	−2.2757
Variance of errors			
VAR (ϵ_1)	0.0000	0.0000	0.0000
VAR (ϵ_2)	22.2587	7.7969	2.8548
VAR (ϵ_3)	90.8051	31.2654	2.9043

administrative database: this has the advantage of performing analysis at a relatively low cost, but also the limit of a lack of many variables that are potentially useful to describe the phenomenon under study.

10.7 Some Concluding Remarks

In this work, we have approached the analysis of student graduation delay from a longitudinal perspective. By means of LCM models, we have evaluated different individual trajectories in student careers through the definition of a longitudinal delay indicator. The latter has been based on the difference of the accumulated standard number of ECTS and the accumulated number of ECTS obtained by students in three time occasions (at the end of the first, second and third academic years). We adopt a conditional LCM model in order to gain information on some delay determinants.

Our results have shown that, in accordance with the specialized literature, some personal characteristics (*gender*, *age* and others) and the type and the final grade of the secondary school have a significant effect on student university career. Hence, female students, students with a "lyceum" of secondary school and students with a higher secondary school final grade perform better than their peers.

However, our approach is also able to discriminate between the covariates that have a significant influence on the initial level and the ones that have effects on the delay trend. Specifically, gender and secondary school final grade significantly influence only the trajectory intercepts (i.e. the delay at the end of the first year), while the type of secondary school has effects on the slopes (i.e. in the ability to decrease/increase the delay).

Finally, we note that more complex models exist within the LCM literature that can be exploited to tackle the graduation delay longitudinal analysis, such as the Autoregressive Latent Trajectory Model [5] which incorporates autoregressive, cross-lagged and latent curve models. However, these models cannot be applied within our case study given that the presence of only three waves of data would cause identification problems.

Acknowledgement The authors wish to thank Giovanni C. Porzio for his useful discussion on the case study model.

References

1. Almalaurea (2008) Profilo dei laureati 2008: I laureati dell'Università Riformata, http://www.almalaurea.it
2. Bianconcini S, Cagnone S, Mignani S, Monari P (2007) La riuscita del percorso universitario: un'analisi longitudinale sugli studenti dell'Ateneo di Bologna. Rivista di Economia e Statistica del Territorio 3:25–38
3. Bianconcini S, Mignani S (2008) Latent variable models for longitudinal data in educational studies. Atti della XLIV Riunione Scientifica, Società Italiana di Statistica, pp 225–232

4. Boero G, Laureti T, Naylor RA (2005) An econometric analysis of student withdrawal and progression in post-reform Italian Universities. Working paper CRENoS Centro Ricerche Economiche Nord Sud, Università di Cagliari, Università di Sassari
5. Bollen KA, Curran PJ (2004) Autoregressive latent trajectories (ALT) models: a synthesis of two traditions. Sociol Methods Res 32:336–383
6. Bollen KA, Curran PJ (2006) Latent curve models: a structural equation perspective. Wiley Interscience, Hoboken, NJ
7. Bollen KA, Paxton P, Curran PJ, Kirby GB (2001) Improper solutions in structural equation models. Causes, consequences and strategies. Sociol Methods Res 29:468–508
8. Boscaino G, Capursi V, Giambona F (2007) La performance delle carriere di una coorte di studenti universitari. Working Paper n.2007.1, Dipartimento di Scienze Statistiche e Matematiche S. Vianelli, Università di Palermo
9. CILEA (2007) Laureati Stella. Rapporto statistico 2004–2006. Arti Grafiche BTZ, Cologna Monzese, Milano
10. CNVSU (2007) Ottavo Rapporto sullo Stato del Sistema Universitario. Rilevazione Nuclei 2007, http://www.cnvsu.it
11. Costantini P (2007) Analyzing learning effects thought latent growth models. In: Book of short paper of classification and data analysis. Eum, Macerata, pp 319–322
12. Costantini P (2008) Nonlinear latent curve models. PhD thesis, Dipartimento di Scienze Economiche, Università di Cassino
13. Duncan TE, Duncan SC, Stoolmiller M (1994) Modeling developmental processes using latent growth structural equation methodology. Appl Psychol Meas 18:343–354
14. Lawrence FR, Hancock GR (1998) Assessing change over time using latent growth modeling. Meas Eval Couns Devel 30:211–223
15. McNabb R, Sarmistha P, Sloane P (2002) Gender differences in student attainment: the case of university students in the UK. Economica 69:481–503
16. Meredith WM, Tisak J (1990) Latent curve analysis. Psychometrika 55:107–122
17. OECD Indicators (2007) Education at a Glance 2007, www.oecd.org
18. Rogosa DR, Willett JB (1985) Understanding correlates of change by modeling individual differences in growth. Psychometrika 50:203–228
19. Schizzerotto A, Denti F (2005) Perduti e in ritardo. L'esperienza dell'abbandono e dell'irregolarita' degli studi in cinque leve di immatricolati all'ateneo di Milano-Bicocca. Nota del Nucleo di Valutazione, Università di Milano-Bicocca
20. Smith J, Naylor RA (2001) Determinants of degree performance in UK universities: a statistical analysis of the 1993 student cohort. Oxf Bull Econ Stat 63:29–60
21. Smith J, Naylor R (2005) Schooling effects on subsequent university performance: evidence for the UK university population. Econ Educ Rev 24:549–562
22. Stoel RD, van Den Wittenboer G, Hox J (2003) Analyzing longitudinal data using multilevel regression and latent growth curve analysis. Metodologia de las Ciencias del Comportamiento, 5:21–42
23. Willett JB, Sayer AG (1994) Using covariance structure analysis to detect correlates and predictors of individual change over time. Psychol Bull 116:363–381

Chapter 11
Assessing the Quality of the Management of Degree Programs by Latent Class Analysis

Isabella Sulis and Mariano Porcu

11.1 Introduction

In the evaluation of university quality, questionnaires with multi-item scales (Likert type) are often used in order to measure specific characteristics which are known to be relevant for the evaluation. The joint distribution of multiple responses provides a complete information in order to attach an overall measure of perceived quality to each student.

The aim of this chapter is to point out classes (clusters) of students (cases) who share a homogenous perception of the quality of the management of their degree programs and to highlight profiles of responses which define each of the identified classes. Latent Class Analysis (LCA) is the modeling approach applied in order to sort out latent classes of observations from a multi-way table of polytomous variables. The ranking of classes has been made using an overall measure of dissimilarity between distributions. The procedure has been used in order to propose a composite indicator of the quality level of the degree program.

This chapter is divided into 5 sections. In Sect. 11.2 the process of building composite indicators of quality of services is discussed and some of the main critical steps are highlighted. In Sect. 11.3 the LCA approach is described. In Sect. 11.4 the proposed method is applied to data on university course evaluation. Section 11.5 provides some final remarks.

11.2 Building up a Composite Indicator

To make clearer and faster comparisons and to highlight possible critical aspects the evaluation process needs practical tools which allow to sort out objects and units (teachers, tutors, courses, facilities, etc.). These tools are usually composite indicators which summarize evaluations expressed by respondents to different indicator

I. Sulis (✉)
Dipartimento di Ricerche Economiche e Sociali, Università degli Studi di Cagliari, Cagliari, Italy
e-mail: isulis@unica.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_11,

variables. The process of building up composite indicators is characterized by a high level of arbitrariness in the definition of many critical components [6]:

1. the *indicator variables* adopted in order to operationalize the attribute;
2. the *transformations* applied in order to re-scale the set of indicator variables;
3. the *weighting scheme* selected in order to discriminate the relevance of each of the re-scaled indicator variables;
4. the *merging function* used in order to summarize multiple indicators in a single statement.

The final results are strongly influenced by researchers' choices and no definitive solutions have been so far proposed in the literature. The use of a modeling approach, especially in the explorative phase, may support and validate researchers' decisions concerning transformations, merging functions, scaling methods, weighting schemes, etc. Most of the statistical models used for measuring unobservable variables throughout indirect indicators are known as Latent Variables Models (Structural Equations Models, Item Response Models, Latent Class Analysis, Classical Scaling Methods, Partial Least Squares Regression etc. [2, 9, 10, 14, 16, 18]) and their use has widely increased in the last decade [4–7, 15, 17].

11.2.1 A Measure of the Perceived Quality of a University Service

In this work, it is assumed by hypothesis that the unobservable attribute *quality of a degree program* is measured indirectly by classifying respondents into groups which are homogeneous in terms of the perceived level of satisfaction of their members. The intensity of the attribute owned by each class needs to be assessed. Specifically, the work focuses on the following steps:

- to set up a statistical approach which allows to sort out mutually exclusive groups (classes) of students characterized by a different perception of the *quality of the management* (QM) of their degree program;
- to sketch the profile of each class (cluster) on the basis of the intensity of the latent attribute;
- to sort classes on the basis of the intensity of the attribute as perceived by students (from the *lowest* to the *highest* intensity);
- to rank degree programs on the basis of the distributions of students across classes.

The approach followed uses tools provided by LCA in order to spot out mutually exclusive classes of students. Each latent class groups together students who share the same perceived level of the quality of the managing of their degree program. Cases are classified into clusters on the basis of posterior probabilities estimated directly from students' response patterns to the items of the questionnaire. Next, classes are sorted moving from a measure of distance between distributions.

11.3 Methodological Issues

LCA aims to identify a number R of categorical classes which clusters observations characterized by a different intensity of the latent variable θ – which is supposed to be categorical – moving from individual responses to a set of categorical indicator variables (i.e. moving from the cross classification of J polytomous indicators). The model assumes that any dependency across responses provided to manifest indicators is explained "by a single unobserved 'latent' categorical variable" [12] θ which takes categories $\theta_r(\theta_1, \ldots, \theta_R)$. Responses to manifest indicators are independent conditional upon the values of the latent variable

$$\pi_{y_1 \ldots y_J|\theta_r} = P(Y_1 = y_1|\theta_r) \ldots P(Y_J = y_J|\theta_r). \tag{1}$$

In order to simplify the notation in the following we denote θ_r just with r. More specifically, by indicating with Y_{ijk} the indicator variable which assumes value 1 if student i ($i = 1, \ldots, n$) selects category (outcome) k ($k = 1, \ldots, K$ the categories) of item j ($j = 1, \ldots, J$, the manifest indicators), the probability that an individual i in class r of the latent variable θ has a particular response pattern is given by

$$f(Y_i; \pi_r) = \prod_{j=1}^{J} \prod_{k=1}^{K} (\pi_{rjk})^{Y_{ijk}} \tag{2}$$

where π_{rjk} is the probability that an observation in latent class r provides the k outcome to item j. The model maximizes the log-likelihood

$$\sum_{i=1}^{N} ln \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K} (\pi_{rjk})^{Y_{ijk}} \tag{3}$$

with respect to π_{rjk} and p_r. The latter is the probability to belong to each of the r classes.

The poLCA package in R-language [12] uses the algorithm *EM* (expectation-maximization). Moving from the estimates of $\hat{p}_r$ and $\hat{\pi}_{jrk}$, the posterior probability that a unit which provides a particular set of responses belongs to a specific class is calculated using Bayes' theorem

$$\hat{P}(r|Y_i) = \frac{\hat{p}_r f(Y_i; \hat{\pi}_r)}{\sum_{r=1}^{R} \hat{\pi}_r f(Y_i; \hat{\pi}_r)}. \tag{4}$$

The algorithm starts using (4) (setting initial guesses for the parameters $\hat{\pi}_{rjk}$ and $\hat{p}_r$) in order to estimate the posterior probability that an individual belongs to a class conditional upon the observed pattern of responses on the J items.

In the second step the log-likelihood function with the updated values of $\hat{\pi}_{rjk}$ and $\hat{p}_r$ is maximized. The two steps are automatically iterated until the convergence

in the log-likelihood is reached. The prediction of the latent class memberships can be improved by using the information available on unit characteristics. The model with covariates is known as Latent Regression Class Model [1, 13].

11.3.1 Sorting Latent Classes

Classes have been sorted moving from the vector of estimated parameters $\hat{\boldsymbol{\pi}}_{\boldsymbol{rj}}$. It has been assessed how much the observed item response probability of each item was dissimilar from the expected item response probability of a hypothetical class of "Zero Satisfied Students" (ZSS). The latter is an extreme distribution with all observations clustered in the first response category.

Different approaches can be used in order to compare two distributions [7, 15] in terms of distance. The dissimilarity index [11] for ordered variables has been adopted. Indicating with F_A and F_B the cumulated probability distributions of two categorical ordered variables "A" and "B" with K categories, the dissimilarity between distributions can be assessed by

$$Z' = \sum_{k=1}^{K-1} |F_{A_k} - F_{B_k}|, \tag{5}$$

the maximum values the index can assume is equal to $K-1$. Thus the relative index z' is

$$z' = \frac{1}{K-1} \sum_{k=1}^{K-1} |F_{A_k} - F_{B_k}|; \tag{6}$$

it varies between [0,1] (the value is 0 when the two distributions are similar). As a proxy of the overall level of satisfaction of the class the average value of the dissimilarity index calculated on the entire set of items has been used. Classes have been ranked in a *continuum* according to the values of this indicator.

11.4 The Application

11.4.1 The Data

The application deals with data on the evaluation of university courses gathered at the Faculty of Economics of the University of Palermo in 2004–2005. The evaluation form used in the survey is divided in separated sections in which students provide information on biographical details, university career, and students' assessments on several aspects of university courses (facilities, lecture programs, teaching). The second column of Table 11.1 shows the distribution of the evaluation forms

Table 11.1 Evaluations collected per degree program

Degree	♯ Evaluations	%
A	32	1.70
B	44	2.33
C	109	5.78
D	297	15.74
E	245	12.98
F	344	18.23
G	816	43.24
Total	1887	100.00

according to which degree program (DP) the course belongs to. The number of questionnaires collected by DP varies from 32 to 816.

In the main section of the evaluation form items take the form of questions (propositions) to which the student is invited to attest how much she/he agrees. All items in the main sections are measured on a four categories Likert scale: *Definitely No* (DN), *More No than Yes* (MN), *More Yes than No* (MY), *Definitely Yes* (DY). This work looks over the joint distribution of 5 items devoted to collect students' opinions on the management of the degree scheme. We selected items concerning the evaluation of general management aspects (QM): the coordination among courses (I_1), the overall workload during the term (I_2), the scheduled hours of the lectures (I_3), the overall organization (I_4), the facilities of the classroom (I_5). The internal consistency reliability of the scale has been assessed using the *Cronbach*'s α [8] coefficient, which signals the degree of the internal homogeneity of the selected indicator variables (if they are measuring or not the same dimension of the underlying variable). The coefficient assumes values between 0 and 1, the closer is the value to 1, the higher is the internal reliability of the scale. The *Cronbach*'s α has been calculated for the whole scale (0.709) and removing each of the 5 items (Table 11.2). The moderately-high level of the coefficient is consistent with the assumption that the selected indicators load on the same dimension.

Table 11.3 exhibits the rate of responses for each category of the 5 ordinal indicators. The bulk of the responses is in the categories "MN" and "MY": "MY" is always the modal and median category with a rate of observations between 30.45 and 41.57%.

Table 11.2 *Cronbach*'s α coefficients

Items	*Cronbach's* α
'*Are you satisfied about . . .*	
I_1 *. . . the degree of coordination among courses*'	Without I_1 0.733
I_2 *. . . the overall workload during a term*'	Without I_2 0.779
I_3 *. . . the scheduled hours of lectures*'	Without I_3 0.658
I_4 *. . . the overall organization*'	Without I_4 0.665
I_5 *. . . the facilities in the classroom*'	Without I_5 0.669
Whole scale	0.709

Table 11.3 Percentage of responses in each category

Item	Def. no	Mod. no	Mod. yes	Def. yes
I_1	15.21	23.72	39.11	21.96
I_2	16.86	24.96	41.57	16.61
I_3	18.39	35.37	36.89	9.34
I_4	16.67	23.80	37.12	22.42
I_5	27.17	27.57	30.45	14.81

11.4.2 The Analysis

The LCA approach is adopted in an exploratory way. The analysis starts by increasing the number of latent classes moving from a complete independent model with just one latent class. Models with a different number of latent classes are compared in terms of BIC or AIC; the first is recommended for basic latent class models [13]. The LCA measures of goodness of fit are displayed in Table 11.4. Moving from M_2 to M_4 both measures recommend an increase in the number of latent classes, whereas moving from M_4 to M_6 the AIC decreases and the BIC increases making not straightforward the choice between the two models. The procedure has been applied for the 4 and 5 classes models and results have been compared in terms of their readability and power to highlight classes of observations which are characterized by a different intensity of the latent attribute "satisfaction". Analysis in the following refers to the 5 classes model (M_5); it exhibits the greatest distance in the *continuum* between scores assigned to extreme classes (*unsatisfied*, *satisfied*). The LCA model has been estimated several times in order to avoid local maxima. Moreover, just to list classes in the output according to an ordering criteria the model has been run again applying the function poLCA.reorder implemented in the package poLCA [13].

The item response probability conditional upon the latent class memberships ($\hat{\pi}_{rjk}$) are reported in Table 11.5. We can observe that C_1 contains students who have the highest probability to score "DN" for all items, thus it represents the class of the "*unsatisfied students*". C_4 groups those students who prevalently score the category "MY" and it is labeled the class of "*moderately satisfied students*". Students in class 5 have a probability which spans between 0.43 and 0.83 to answer "DY"; thus the class represents the cluster of "*satisfied students*". It seems to be sensible to sort out

Table 11.4 Measures of goodness of fit

♯Classes	Model	♯Par.s	Measures of goodness of fit		
6	M_6	95	AIC: 23357.9	BIC: 23884.4	Dev: 929.9
5	M_5	79	AIC: 23375.4	BIC: 23813.2	Dev: 979.4
4	M_4	63	AIC: 23419.6	BIC: 23768.8	Dev: 1055.7
3	M_3	47	AIC: 23519.7	BIC: 23780.2	Dev: 1187.8
2	M_2	31	AIC: 23969.9	BIC: 24141.7	Dev: 1669.9
1	M_1	15	AIC: 24881.7	BIC: 24964.9	Dev: 2613.8

Table 11.5 Item response probability conditional upon latent class memberships

Item	$\hat{\pi}_{rjDN}$	$\hat{\pi}_{rjMN}$	$\hat{\pi}_{rjMY}$	$\hat{\pi}_{rjDY}$
	Class 1			
I_1	0.32	0.16	0.30	0.22
I_2	0.80	0.11	0.06	0.03
I_3	0.73	0.19	0.04	0.04
I_4	0.72	0.17	0.09	0.03
I_5	0.76	0.14	0.06	0.04
	Class 2			
I_1	0.09	0.25	0.50	0.15
I_2	0.20	0.54	0.25	0.00
I_3	0.24	0.68	0.08	0.00
I_4	0.17	0.53	0.29	0.01
I_5	0.37	0.39	0.24	0.01
	Class 3			
I_1	0.24	0.19	0.29	0.27
I_2	0.12	0.25	0.41	0.22
I_3	0.19	0.36	0.39	0.05
I_4	0.13	0.13	0.29	0.46
I_5	0.25	0.22	0.27	0.26
	Class 4			
I_1	0.05	0.29	0.49	0.17
I_2	0.01	0.15	0.75	0.08
I_3	0.00	0.28	0.67	0.05
I_4	0.02	0.20	0.68	0.09
I_5	0.12	0.32	0.47	0.10
	Class 5			
I_1	0.12	0.19	0.26	0.43
I_2	0.00	0.02	0.15	0.83
I_3	0.00	0.00	0.25	0.74
I_4	0.02	0.01	0.20	0.78
I_5	0.13	0.10	0.31	0.46

classes from the *least satisfied* to the *most satisfied* $C_1 < C_4 < C_5$. No clear sorting order appears for classes C_2 and C_3.

The predicted class memberships by modal posterior probability for the 5 class model are equal to $\bar{p}_r\,(0.104, 0.261, 0.232, 0.316, 0.085)$. The closeness of predicted and estimated shares of class memberships $\hat{p}_r\,(0.111, 0.232, 0.269, 0.307, 0.081)$ is a further measure of the goodness of fit of the selected model [13].

The criteria used to sort out the latent classes (i.e., to locate them in a *continuum*) and to set the order in the previously mentioned poLCA.reorder function has been described in Sect. 11.3.1. Denoting with $\hat{P}_{rjk}$ the cumulated distribution of $\hat{\pi}_{rjk}$ and with P_{ZSSjk} the cumulated distribution of p_{ZSSjk} $[P_{ZSSjk} = (1, 1, 1, 1)]$ for the class of ZSS, the dissimilarity index has been calculated for each of the J item in each of the R classes

$$z'_{rj} = \frac{1}{K-1} \sum_{k=1}^{K} |P_{rjk} - P_{ZSSjk}|. \tag{7}$$

The average value of z'_{rj} calculated for each class ($\bar{z}'_r$) is used to sort classes in a *continuum* and to handle them in further analysis as ordered categories.

$$\bar{z}_r = \frac{1}{J} \sum_{j=1}^{J} z'_{rj}. \tag{8}$$

The ranking of the classes, from the "least satisfied" – least likely to score a higher category – to the "most satisfied" – most likely to score a higher category – is: $C_1 < C_2 < C_3 < C_4 < C_5$. Table 11.6 shows the value of z' for the five items in each latent class. Almost all items show an increasing value of z'_{rj} moving from class 1 to class 5 (with the exception of item I_4 which decreases from class C_3 and C_4).

The last two rows in Table 11.6 show the value of $\bar{z}'_r$ and their standard deviations within each class. The *within* class variability of z'_{rj} can be seen as a measure of reliability of $\bar{z}'_r$. The *between class* variability of z'_{rj} (0.045) explains 81% of the *total* variability of z'_{rj}(0.055). Scores attached to each class in order to locate it in the continuum are $\bar{z}'_r$(0.195, 0.375, 0.550, 0.589, 0.824). From the values of $\bar{z}'_r$ arises that class 3 and class 4 identify students with a similar perception of the QM. A deeper look to the item response probability of both classes shows that students in the first class discriminate more (the modal category varies across items); nevertheless students in class 4 have a higher probability to provide outcomes in the category "MY".

In order to validate the use of $\bar{z}'_r$ as overall index of satisfaction of the class, the dissimilarity index across pairs of estimated item response probabilities within each class has been calculated. Results depicted in Table 11.7 show that the dissimilarity of the distributions of the items within each class is quite low and the highest value observed is 0.38 for class 2. The predicted class membership probabilities highlight that the 10.5% of surveyed students are *unsatisfied*, the 31.6 % are *moderately satisfied* and the 8.5% belong to the class of *satisfied students*.

Table 11.6 A comparison across classes using z' values (M_5)

Item	C_1: z'_{1j}	C_2: z'_{2j}	C_3: z'_{3j}	C_4: z'_{4j}	C_5: z'_{5j}
I_1	0.474	0.572	0.531	0.592	0.664
I_2	0.102	0.353	0.579	0.636	0.935
I_3	0.127	0.279	0.438	0.585	0.912
I_4	0.143	0.378	0.689	0.615	0.911
I_5	0.128	0.293	0.514	0.515	0.701
$\bar{z}'$	0.195	0.375	0.550	0.589	0.824
$sd(z')$	0.018	0.010	0.006	0.002	0.013

The procedure adopted to score classes has been replicated for the 4 classes model in order to assess the sensibility of the scoring method to the number of classes. By applying the same procedure the vector of score attached to classes is equal to $\hat{z}_r(0.233, 0.375, 0.570, 0.724)$; the range of variation of $\bar{z}'_r$ is smaller but contiguous categories differentiate more across students with a different perception of the overall quality. However, as a consequence of the narrower range of variation of $\bar{z}'_r$ the rate of variability in $\bar{z}'_r$ explained by the *between* classes variability is slightly lower (78.5%) with respect to the 5 classes model. Results in terms of scores assigned to latent classes are summarized in Table 11.8. The rate of students in each class assigned by modal posterior probability $\bar{p}_r(0.1298, 0.239, 0.4822, 0.1489)$ shows that the choice of a simpler model changes the distributions of students across classes with a greater clustering in the extreme ones. Thus even if the 4 classes

Table 11.7 Matrix of dissimilarity between pairs of items in each class

Item	I_1	I_2	I_3	I_4	I_5
	Class 1				
I_1	0.00	0.32	0.29	0.27	0.29
I_2	0.32	0.00	0.06	0.06	0.03
I_3	0.29	0.06	0.00	0.03	0.04
I_4	0.27	0.06	0.03	0.00	0.04
I_5	0.29	0.03	0.04	0.04	0.00
	Class 2				
I_1	0.00	0.27	0.38	0.24	0.28
I_2	0.27	0.00	0.12	0.03	0.12
I_3	0.38	0.12	0.00	0.15	0.19
I_4	0.24	0.03	0.15	0.00	0.13
I_5	0.28	0.12	0.19	0.13	0.00
	Class 3				
I_1	0.00	0.11	0.18	0.12	0.02
I_2	0.11	0.00	0.12	0.16	0.11
I_3	0.18	0.12	0.00	0.27	0.18
I_4	0.12	0.16	0.27	0.00	0.14
I_5	0.02	0.11	0.18	0.14	0.00
	Class 4				
I_1	0.00	0.18	0.12	0.13	0.06
I_2	0.18	0.00	0.09	0.05	0.19
I_3	0.12	0.09	0.00	0.06	0.13
I_4	0.13	0.05	0.06	0.00	0.14
I_5	0.06	0.19	0.13	0.14	0.00
	Class 5				
I_1	0.00	0.27	0.21	0.23	0.06
I_2	0.27	0.00	0.07	0.04	0.25
I_3	0.21	0.07	0.00	0.04	0.19
I_4	0.23	0.04	0.04	0.00	0.21
I_5	0.06	0.25	0.19	0.21	0.00

Table 11.8 A comparison across classes using z' values (M_4)

Item	C_1: z'_{1j}	C_2: z'_{2j}	C_3: z'_{3j}	C_4: z'_{4j}
I_1	0.470	0.571	0.574	0.603
I_2	0.156	0.347	0.614	0.800
I_3	0.153	0.267	0.545	0.685
I_4	0.204	0.398	0.612	0.883
I_5	0.181	0.292	0.505	0.648
$\bar{z}'$	0.233	0.375	0.570	0.724
$sd(z')$	0.014	0.012	0.002	0.011

model would be straightforwardly adopted in an analysis aimed to identify clusters of students, it would led to a lost of useful information if the final aim is to summarize results in a synthetic indicator.

A composite indicator of students' perceived quality of the management of the degree programs at faculty level is obtained as a linear combination of the scores assigned to each latent class $\bar{z}'_r$ weighted for the of rate of students $\bar{p}_r$

$$IS = \sum_{r=1}^{R} \bar{p}_r \bar{z}'_r. \tag{9}$$

In the following the sensibility of the composite indicator to the choice of the number of classes as been assessed. For the 5 classes model the value of the indicator at faculty level is equal to $IS = 0.503$; it is calculated as a combination of the system of weights $\bar{p}_r(0.105, 0.261, 0.232, 0.316, 0.085)$ with the scores $\bar{z}'_r(0.195, 0.375, 0.550, 0.589, 0.824)$. The composite indicator calculated for model 4 shows a similar value ($IS = 0.502$).

A comparison among the 7 degree programs of the faculty has been made moving from the rate of students in the five latent classes. The indicator IS_{DP} has been applied to the seven degree programs weighting the rate of observations in each class with $\bar{z}'_r$. Results are provided in the last columns of Tables 11.9 and 11.10. Table 11.9 shows that the main differences in the distributions of students across the five classes are observed in the rate of cases in the extreme categories: the lowest rates of *unsatisfied students* are observed for degree programs "A", "F" and "C", while the highest rates of *satisfied students* are observed for degree programs "D",

Table 11.9 Frequencies of observations in each class by degree program (M_5)

Degree program	C_1: $\bar{p}_1$	C_2: $\bar{p}_2$	C_3: $\bar{p}_3$	C_4: $\bar{p}_4$	C_5: $\bar{p}_5$	IS_{DP}
G	0.126	0.319	0.208	0.276	0.071	0.479
E	0.121	0.253	0.246	0.306	0.074	0.495
B	0.091	0.227	0.250	0.318	0.114	0.521
A	0.031	0.312	0.250	0.312	0.094	0.522
D	0.102	0.184	0.269	0.318	0.127	0.529
F	0.064	0.195	0.270	0.387	0.084	0.531
C	0.064	0.239	0.156	0.422	0.119	0.534
Faculty	0.105	0.261	0.232	0.316	0.085	0.503

Table 11.10 Frequencies of observations in each class by degree program (M_4)

Degree program	C_1: $\bar{p}_1$	C_2: $\bar{p}_2$	C_3: $\bar{p}_3$	C_4: $\bar{p}_4$	IS_{DP}
G	0.151	0.286	0.435	0.129	0.483
E	0.152	0.236	0.495	0.118	0.491
A	0.062	0.312	0.469	0.156	0.512
B	0.114	0.205	0.477	0.205	0.523
D	0.131	0.171	0.494	0.204	0.524
C	0.101	0.211	0.495	0.193	0.524
F	0.078	0.186	0.573	0.163	0.532
Faculty	0.130	0.239	0.482	0.149	0.502

"C" and "B". The main evidence comparing the seven distributions is that the degree programs have a high rate of students in the class of *moderately satisfied*. Results of M_4 are consistent with those obtained for M_5. The small variability across the IS_{DP} values could be partially explained considering that the degree programs belong to the same faculty and thus they largely share a common management.

11.5 Final Remarks

The chapter provides a method to rank degree schemes moving from students' joint response pattern to a set of ordered indicators. The LCA approach allows to skip the problem of choosing *transformation functions* and *weighting schemes* in order to summarize multiple indicators into a single measure.

Further researches are still in progress in order to: (a) improve the prediction of students' class membership by taking into account students' characteristics; (b) assess the sensitivity of the approach to the method adopted to rank latent classes; (c) validate the method on other data sets concerning the evaluation of course management; (d) explore the potentiality of the modeling approach on longitudinal studies. Furthermore LCA [1, 3, 12] classifies students on the bases of the posterior probability (modal probability) that a unit which provides a specific response pattern belongs to a specific class. In the modal assignment the variability in students' probability to belong to each class is completely ignored. This means that in a four classes model a student with a pattern of response as 0.24, 0.25, 0.25, 0.26 would be deterministically placed in class 4. A simulation analysis could be carried out in order to assess the sensibility of the overall index to criteria of assignment of the units which relies on the variability of the probability vector.

References

1. Agresti A (2002) Categorical data analysis. Wiley-Intersciencie, Hoboken, NJ
2. Bartholomew DJ (1998) Scaling unobservable constructs in social science. Appl Stat 47:1–13
3. Bartholomew DJ, Knott M (1999) Latent variable models and factor analysis. Kendall's Library of statistics. Hodder Arnold, London

4. Barholomew DJ, Steele F, Galbraith JI, Moustaki I (2002) The analysis and interpretation of multivariate analysis for social scientists. Chapman & Hall, Boca Raton, FL
5. Bayol MP, de la Foye A, Tellier C, Tenenhaus M (2000) Use of PLS path modelling to estimate the European consumer satisfaction index (ECSI) model. Stat Appl 12:361–375
6. Bernardi L, Capursi V, Librizzi L (2004) Measurement awareness: the use of indicators between expectations and opportunities. In: Atti XLII Convegno della Società Italiana di Statistica, Bari, 9–11 Giugno 2004. CLEUP, Padova
7. Capursi V, Porcu M (2001) La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni di giudizi. In: Atti Convegno Intermedio della Societ Italiana di Statistica 'Processi e Metodi Statistici di Valutazione', Roma 4–6 Giugno 2001. Società Italiana di Statistica
8. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334
9. Hagenaars JA, McCutcheon AL (2002) Applied latent class analysis. Cambridge University Press, Cambridge
10. Joreskög KG, Sorbom D (1979) Advances in factor analysis and structural equation models. Abt Books, Cambridge, MA
11. Leti G (1983) Statistica Descrittiva. il Mulino, Bologna
12. Linzer DA, Lewis J (2007) poLCA: Polytomous variable latent class analysis. R package version 1.1. http://userwww.service.emory.edu/∼dlinzer/poLCA
13. Linzer DA, Lewis J (forthcoming) poLCA: An R Package for Polytomous variable latent class analysis. Journal of Statiscal Software
14. Moustaki I, Knott D (2000) Generalized latent trait models. Psychometrika 65:391–411
15. Rampichini C, Grilli L, Petrucci A (2004) Analysis of university course evaluations: from descriptive measures to multilevel models. Stat Methods Appl 13:357–371
16. Rasch G (1960) Probabilistic models for some intelligence and attainment tests. MESA Press, Chicago, IL
17. Sulis I (2007) Measuring students' assessments of 'university course quality' using mixed-effects models. PhD thesis, Università degli Studi di Palermo, Palermo
18. Wold H (1966) Estimation of principal components and related models by iterative least squares. In: Krishnaiaah PR (ed) Multivariate analysis. Academic Press, New York, NY, pp 391–420

Part IV
Research Design and Data for Evaluation: University Between the High School and the Labour Market

Chapter 12
The Multicriteria Electre III Model Applied to the Evaluation of the Placement of University Graduates

Rosalinda Allegro and Ornella Giambalvo

12.1 Preliminary Remarks

During the last years the Italian university system has been undergoing a reform process regarding issues of governance such as the progressive financial autonomy of the University (art. 5, law 537/24 December 1993) and the reshape of the academic curricula (law 509/99 and 270/04). The most obvious result of this complex process has been the challenging attempt to manage a necessary change in a context of limited financial resources. In order to encompass the composite features of an efficient and effective management of resources, features of high relevance for the full accomplishment of the reform benchmarks, one has to take into account the academic organisation's objectives, so that to use tools and methods able to support decisions and, thus, allowing a rationalization of the decision-making processes.

The rationalization of the processes unavoidably implies both their own evaluation and the evaluation of the objectives at stake. To this end, analytical models based on decisions have been developed over the last decades and represent a solid starting point for our analysis.

Thus, we have decided to apply multi-criteria methods to the university environment in order to evaluate groups of graduates from the point of view of their placement on the labour market. Moreover, we share strong beliefs that there might be similarities between the academic reality and the business area where previous studies had tested similar methods. However, this is the second attempt in the university arena, following the previous evaluation of the degree programmes in two faculties of the University of Palermo in 2005 [5]. These methods could, for instance, identify those degree programmes that need supplementary support for increasing their performances or, on the contrary, award the outstanding ones. The analysis is structured in three parts. The first one describes the data used for the analysis. The second one describes the multi-criteria method and the model of outranking applied to the university graduates environment. The third part introduces the results of the

R. Allegro (✉)
Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università degli studi di Palermo, Palermo, Italy
e-mail: rl.allegro@libero.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_12,

previous analysis. The final chapter provides several concluding remarks aiming to stress out the advantages and limits of the employed methodology. In the end, a follow-up agenda is briefly introduced.

12.2 The Data

This analysis uses the data collected by STELLA (Statistiche in tema di laureati e lavoro) post-graduation placement survey conducted by CILEA on the 2006 University of Palermo graduates one year after their graduation [2]. The survey aims to evaluate the graduate's global satisfaction considering the whole period he/she attended the university. Thus, it provides the Italian universities with an update database of their graduates aiming to analyse the job opportunities, the level of satisfaction regarding the academic curriculum, the students' knowledge, competencies and skills, the convergence degree between academic curricula and labour market requirements, the overall degree of satisfaction regarding the quality of the academic curriculum, etc.

Based on CATI method, the survey used questionnaires with five distinct sections. With a broad grasp, the first section regards all the graduates of the new system and focused on general information on the graduates' academic curriculum, their mobility experiences, their job experiences during their studying and, more in general, information concerning their families social background. The last question was a filter aiming to identify the students' post-graduation path (the employment status was divided in two categories: employed graduates and job seeking graduates; the non-employed status was divided in two categories: the graduates who continued their education and other categories). The second, third and forth parts focused on the employed and job-seeking graduates (these two categories were gathered in the global Workforce category) and those continuing their studies. In the final part of the survey, the graduates expressed general opinions and thoughts on the university system.

Based on this survey, the analysis took into account exclusively the 3 years long university degrees belonging to all the groups of disciplines,[1] codified as such:

The main findings can be summarised as such.

From a socio-economic point of view, almost all the groups of disciplines include graduates rated with a medium-high social background. Concerning the social background, defined taking into account the median value of a variable obtained from analysing both the parents' profession and their education degree (see the methodological note in [2]), the survey shows a relationship between this variable and job and university success [2]. As an exception, the Health and Physical Education group is characterised by a medium-low social background (Table 12.1).

[1] For a broad overview of the various degrees within the single groups of disciplines see CILEA (ed. 2006), a post-graduation Placement Survey (Consorzio interuniversitario per l'elaborazione automatica).

a_1: Agrarian studies
a_2: Architecture
a_3: Chemistry and Pharmaceutics
a_4: Economy and Statistics
a_5: Health and Physical Education
a_6: Geo-biology
a_7: Law studies
a_8: Engineering
a_9: Sciences of Education
a_{10}: Literature
a_{11}: Linguistic studies
a_{12}: Medicine
a_{13}: Political and Social studies
a_{14}: Psychology
a_{15}: Science

Table 12.1 Several characteristics of the graduates by group of disciplines, social background, effective length of the studying period, graduation grades, and percentage of satisfied graduate students with their academic curriculum

Group of disciplines	Social background	Effective length of the studying period[a]	Graduation grade (average)	% of graduates reporting satisfaction with the academic curriculum[b]
Agrarian studies	3	1.65	103.3	73.9
Architecture	4	1.47	107.3	64.6
Chemistry and Pharmaceutics	3	1.47	100.3	58.1
Economy and Statistics	3	1.65	103.2	76.8
Health and Physical Education	2	1.20	107.6	40.0
Geo-biology	4	1.59	105.7	65.2
Law studies	3	1.62	102.4	70.1
Engineering	4	1.57	104.0	78.2
Sciences of Education	3	1.42	102.2	52.3
Literature	4	1.50	106.7	62.1
Linguistic studies	4	1.58	105.2	48.4
Medicine	3	1.13	107.8	73.4
Political and Social studies	3	1.51	105.8	68.6
Psychology	3	1.53	104.4	67.9
Science	3	1.74	103.4	74.7
Total	3	1.51	104.9	67.2

[a]The effective length of the studying period is calculated as the relation between the average periods employed for graduation and the legal length of the degree programme.
[b]The percentage refers to the graduates that in relation with their own academic experience consider that they would apply for the same degree at the same university.

Concerning the length of their university studying, students in Medicine followed by students in Health and Physical Education tend to graduate faster than the other categories. At the other extreme, the Scientific group tends to graduate later than the others. In addition, not only the students in Medicine and Health and Physical Education are graduating faster, but they also have the highest final grades (respectively 107.8 and 107.6 – Table 12.1), followed by the Architecture (107.3). The lowest final grades are registered among the graduates in Chemistry and Pharmaceutics (100.3).

By excluding, on the one side, the Health and Physical Education group and, on the other, the Linguistic studies group whereas less than 50% of graduates tend to be satisfied with their academic curricula (respectively 40% and 48.4%), for the

most part (more precisely over 70%), the graduates in Engineering, Economy and Statistics, Science, Agrarian Studies, Medicine and Law Studies declare themselves satisfied with their academic curricula.

Concerning their employment status, only one third of the graduates has a job one year after their graduation (30.9% equivalent to 30% of the male students and 31.4% of the female students). The gender issue does not seem to be a discriminant for the job placement after graduation. According to their groups of disciplines, strong differences can be seen when it comes to the path followed by the graduates in short-term degrees.[2] Thus, over 80% of the Medicine students (81.5%) have a fast track to the job market immediately after their graduation, while less than 20% of the graduates in Geo-Biology (10%), Engineering (18.1%) and Literature (19.6%) have a smooth insertion on the job market (Table 12.2). As illustrated by Table 12.2, by excluding the Medicine graduates, only the graduates in Health and Physical Education (60%) and those in Chemistry and Pharmaceutics (53.9%) are employed beyond the threshold of 50%. Still, while the graduates in Medicine and Chemistry and Pharmaceutics have usually full time contracts (respectively 68.7% and 41.9%),

Table 12.2 Several characteristics of the graduates by group of disciplines, type of job (A employed, B full time and for undetermined length of tie contracts), decisional autonomy (C), degree of responsibility (D), employed on the Sicily labour market (E) and medium monthly net income (F)

Group of disciplines	A %	B %	C[a] %	D[b] %	E %	F €
Agrarian studies	24.2	16.3	83.8	68.9	83.8	696
Architecture	30.1	16.9	73.5	50.4	92.7	957
Chemistry and Pharmaceutics	53.9	41.9	55.6	0.0	88.9	1139
Economy and Statistics	23.7	14.3	71.2	25.1	89.1	767
Health and Physical Education	60.0	0.0	33.3	33.3	66.7	417
Geo-biology	10.1	5.0	49.2	32.1	100.0	726
Law studies	20.7	11.4	46.8	17.6	78.1	655
Engineering	18.1	13.9	63.8	28.4	66.7	1069
Sciences of Education	47.8	17.1	63.2	11.4	73.0	671
Literature	19.6	6.2	66.7	30.8	81.4	548
Linguistic studies	34.1	16.2	47.1	25.5	87.4	678
Medicine	81.5	68.7	75.6	33.8	79.9	1169
Political and Social studies	31.0	18.3	60.8	34.1	74.0	819
Psychology	21.0	3.9	46.0	0.0	96.4	505
Science	24.7	19.8	77.0	46.0	65.5	980
Total	30.9	18.5	63.2	27.7	79.9	862

[a]% of employed graduates reporting having a decisional autonomy.
[b]% of employed graduates reporting having job responsibilities over other employees.

[2] Under the new system, the first university degree is similar to the Bologna bachelor's degree, it normally lasts 3 years.

the graduates in Health and Physic Education have a more precarious situation, with part-time or occasional contracts. Decisional autonomy and responsibility are the main characteristics of the graduates in Agrarian studies (83.8% of them report their decisional autonomy and 68.9% of them report being responsible for other employees – Table 12.2). Similarly, over 70% of the graduates in Science (77%), Medicine (75.6%), Architecture (73.5%) and Economy and Statistics (71.2%) report having a decisional autonomy. The lowest level of decisional autonomy characterise the graduates in Health and Physical Education. Concerning the responsibility in relation to other employees, with the exception of the graduates in Agrarian studies, 50% of the graduates in Architecture report having a certain level of responsibility. None of the graduates in Chemistry and Pharmaceutics or in Psychology reports having a responsibility position in their job.

For the most part, beyond the affiliation to specific groups of disciplines, most of the graduates are employed in Sicily: over 90% of the graduates in Geo-Biology (100%) Psychology (96.4%) and Architecture (92.70%). Percentages lower than 70% are registered among the graduates in Engineering (66.7%), Health and Physical Education (66.7%) and Science (65.5%) (Table 12.2).

By taking into account the monthly income of the employed graduates, only the graduates in Medicine, Chemistry and Pharmaceutics and Engineering earn more than 1,000 euros (Table 12.2), while the graduates in Literature, Psychology and Health and Physical education earn less than 600 euros monthly. In order to evaluate the opinions reported by the employed graduates regarding the coherence between study and employment, the academic curricula adequacy to the current employment and the overall satisfaction with their work, the analysis takes into account the median value of an indicator built in accordance with the graduates' remarks. Expressed on a scale from 1 to 4, these opinions have been successively transformed on a scale from 0 ("not satisfied at all") to 10 ("completely satisfied") [1]. Based on Table 12.3, according to the three analysed aspects, the graduates in Medicine are among the most satisfied (coherence study-employment 7.5, academic curricula adequacy 7.5, overall satisfaction 9).

The graduates in Health and Physical Education are the only other group reporting a higher level of overall satisfaction (9.5). Nevertheless, the same group reports lower levels of satisfaction concerning the study-employment relation (5.5) and academic curricula coherence (3.5). Beside the graduates in Medicine, five other groups of disciplines report a positive evaluation of the three analysed aspects: the graduates in Agrarian Studies, Architecture, Science of Education, Linguistic Studies and Science. Concerning the coherence between study and employment, opinions under-stating a satisfactory level characterised the groups of Law Studies (3.5), Literature (4), Chemistry and Pharmaceutics (5), Health and Physical Education, Geo-Biology Group, Social and Political Sciences and Psychology (5.5). Concerning the academic curricula adequacy to the current job requirements, under-stating opinions were reported by graduates in Health and Physical Education (3.5), Literature (4), Geo-Biology (4.5), Economy and Statistics, Law Studies, Social and Political studies and Psychology (5), Chemistry-Pharmaceutics and Engineering (5.5). Without any differences in terms of belonging group of disciplines, all the interviewees reported an opinion at least satisfactory with the current job.

Table 12.3 Opinions reported by employed graduates (median values) regarding the coherence between study and employment, the adequacy of the curriculum to the current employment requirements and overall satisfaction with their job description

Group of disciplines	Coherence study – employment	Academic curricula adequacy	(Overall) Satisfaction with the job description
Agrarian studies	6.5	7.0	6.5
Architecture	6.5	7.0	6.5
Chemistry and Pharmaceutics	5.0	5.5	8
Economy and Statistics	6.0	5.0	8.5
Health and Physical Education	5.5	3.5	9.5
Geo-biology	5.5	4.5	6.0
Law studies	3.5	5.0	7.0
Engineering	6.0	5.5	7.5
Sciences of Education	6.0	6.5	8.0
Literature	4.0	4.0	7.0
Linguistic studies	6.0	6.0	7.5
Medicine	7.5	7.5	9.0
Political and Social studies	5.5	5.0	7.5
Psychology	5.5	5.0	6.0
Science	6.0	6.0	7.0
Total	6.0	6.0	7.5

12.3 The Multicriteria Electre III Model

Several evaluation approaches aiming to identify the "best" possible solution refer to the utility theory which implies the existence of a univocal utility function. Tracing the decisional aspect back to the maximization of a utility function raises problems for the decision-maker since it does not take into account the different dimensions, the various points of view and the diverse objectives [3]. The optimisation paradigm had been abandoned in various areas of the theoretical research and is regularly criticised by the literature. For example, according to Herbert Simon (1978 Nobel Price for economy), a promoter of a critical discourse on the topic, this is not the "best" alternative one has to achieve (objectively, it might also be impossible to do it), but one should aim to identify those alternatives that "satisfy" a certain number of requirements explicitly defined (the model satisfying choice of H. Simon [11]).

In line with the above, if one intends to analyse a problem, taking into account the various aspects of the issue and its features, it is necessary to adopt a method that replaces the "optimal solution" with a group of "efficient solutions". According to this approach, defined as a multi-criteria approach, the final solutions depend on the initial conditions identified by the decision-maker him/herself. These decisions must, therefore, be defined and "justified". According to this criterion, the general approach to a decisional problem consists in using the information together with the opinions expressed by the decision-maker in order to establish a compromise or, in

other words, to help the decision-maker to choose the alternative more coherent with his/her structure of preference [7]. In general, there is no possible decision (a solution for the problem or an action to be undertaken) which is simultaneously the best choice from all the points of view considered as being relevant for dealing with the decisional problem. There is, instead, one set of solutions, generally numerous, that provides a logical framework for the choice of a "compromise" solution between the problems and the values that inspire the evaluator.

In the early 1960s, the field of the operative research laid emphasis on the need to take into account a multiplicity of criteria, sometimes in conflict with each other, in order to provide a solution. The solution in cause did not have anymore the characteristics of the "optimal" solution typical for mathematic programs; although it still was an admissible solution, given that, by substituting a single objective to optimize with a plurality of objectives, sometime in conflict with each other, there were not anymore the logical and mathematical conditions for guaranteeing the existence of an optimal solution.

The multi-criteria analysis is still a young theoretical approach, illustrating one set of diversified methodologies which are not yet homogenized in a common theoretical framework. The most recent research fills the gap between the empirical aspects and the theoretical systematization. The multi-criteria analysis integrates the following basic components: the actions and their related criteria the decision-maker(s) and the possible support for the data elaboration, the decision rule (rule used for ordering the alternatives according to the information received and the decision-maker's preferences). The decision procedure generally debouches into the choice between various elements that the decision-maker examines and evaluates according to specific criteria. These elements are considered actions or alternatives and compose the cluster A of actions among which the decision-maker has to operate his/her choice. The definition of A not only depends on the specific problem that has to be solved and the subjects involved in the decision-making process, but strongly interacts with the modelization of the preferences, the definition of the criteria, the enunciation of the problem and, last but not least, the choice between the supporting methods that are applied. Criteria are measured on each action. A criterion can directly provide indications regarding the level of a criterion; in certain cases, a criterion can have a correspondent characteristic. Thus, there might be a characteristic (a set of characteristics) that, indirectly, provides information concerning that criterion.

Among the multi-criteria methods supporting the decision process, special attention is due to the outranking method developed for dealing with problems of choice (the best action among various alternatives), of classification (assignment of actions to more classes which characteristics are known) and of ordering (construction of an order of preferences linked to the set of possible actions). These methods aim to build a relation between the actions, a so called outranking relation, and to use this relation for supporting the decision-maker in dealing with the specific problem.

In all the methods of outranking, pairs of potential actions are confronted on each individual criterion in order to establish if one of the two actions is preferable to the other or if there is no difference at all. The challenge behind the aggregation of

the results of these confrontations is dealt by building a relation of outranking (S), understood as the union of elementary relations of indifference (I), light preference (Q) and heavy preference (P). Furthermore, the method takes into account the lack comparability between actions (N), different from the indifference since caused by the existence of contrasting preferences on various criteria making impossible to establish which of the actions is better, knowing that they are not the same.

One can say that action a outranks action a' (aSa') if, according to what it is known regarding the preferences of the decision-maker and the quality of the evaluations of the actions, "there are sufficient reasons for considering that a is at least as good as a' and there are not good reasons for refuting this statement". The outranking is based on the principle of concordance/discordance,[3] in other words on testing the existence of the concordance of criteria in favour of an action instead of another and on controlling that there are not situations of strong discordance among the evaluations able to challenge (the veto issue) the concordance. All the outranking methods provide a structure in steps, in which one is focused on confronting two by two individual criteria and on the aggregation of these results with the outranking modelization (through tests or elaboration of indices of concordance and discordance). The next step uses the outranking relations to reach a final result, by adopting a procedure in order to make operative a coherent decision rule in dealing with the decisional issue. There are various outranking methods the choice among different methods is motivated by indications connected to the nature of the available data and, thus, the criteria that can be used, to the precise decision rule to be made operative, to the presence/absence of thresholds.

Two main families form the category of outranking methods: the methods Electre, oriented towards the choice (Electre I) or towards the ordering (Electre II, III and IV), and the methods of selection/segmentation, dealing with the problem of classification (as Electre Tri). The methods Electre (Elimination Et Choix TRaduisant la REalité), developed by Roy and his collaborators from the University Dauphine – Paris, starting with late 1960s, distinguish themselves by the confronted issues (choice for the first one, ordering for the others), the nature of the data and, thus, the type of criteria (criteria for the first and second, with cardinal scales for the first one and cardinal or ordinal scales for the second; pseudo-criteria are, instead, used for the others together with cardinal scales with thresholds) and the outranking modelization procedure [9].

The analysis carried out in this study uses the multi-criteria decision method Electre III [9]. This approach makes it possible to take into account the imprecision and uncertainty with which the characteristics are often evaluated and, meanwhile,

[3] The indices of concordance and discordance used in this type of models are different from the usual statistical indices of association. The concordance is not understood as linked to the variables, the criteria in this specific case, but to the alternatives. Two alternatives are concordant if picking one or the other makes no difference for the decision-maker choice. They are discordant when they are not comparable.

avoid that a solution, unacceptable for on single requirement, can prevail on other; this is obtained by applying a veto threshold to the comparison between two solutions for each considered evaluation criterion.

The third version of the model represents the first attempt of fuzzy outranking relation in the literature and it goes back to 1978 [9]. This method differentiates itself from Electre I and II mainly because it uses pseudo-criteria, in other words criteria to which are associated elements of informative and preferential uncertainty and implies, at the first stage of the method, a fuzzy outranking relation by associating to each relation between ordinate pairs of actions a characteristic function $\delta(a, a')$ that expresses the degree of credibility of the outranking relation.

According to Electre III model, the user has to employ the data (alternatives and criteria) and the decision-makers' preferences. These preferences are defined according to a weight and three threshold values for each criterion. The weight associated to each criterion corresponds to a coefficient of relative importance that represents one of the most delicate parts of the model since it is the most direct and explicit expression of the decisional preferences and can relevantly influence the results of the method. The thresholds correspond to values that are introduced for limiting two types of risks: the risk of considering distinct two situations corresponding to conditions and values which are very close and substantially equivalent and the risk of not encompassing preferential situations as different. In particular, the indifference threshold (q_j) refers to the smallest difference, among the values of the criterion j, to which the decision-maker attributes a meaning in terms of indifference. For example, if the difference between two groups of disciplines equals to two points related to the graduating average degree and the indifference threshold for this criterion equals to 3, then the two groups are, in fact, indifferent to this criterion. Only a difference beyond 3 is considered relevant. The preference threshold (s_j) expresses the minimal difference, among the values of the criterion j, to which the decision-maker attributes a meaning in terms of narrow preference. For example, if the difference between two groups of disciplines equals to 5 points related to the graduating average degree and the preference threshold established by the decision-maker for this criterion equals to 4, then the group of disciplines with the highest degree will be preferred to the other. The veto threshold (v_j) expresses the minimal difference, among the values of the criterion j, beyond which the decision-maker considers that the gap between the scores is not anymore balanceable by the performances of the other criteria. For example, if the group of discipline A surpasses the group of discipline B by 8 points, related to the graduating average degree, and the veto threshold for this criterion is established by the decision-maker to 5, then B cannot outranks A, whatever the relative value of the other characteristics might be.

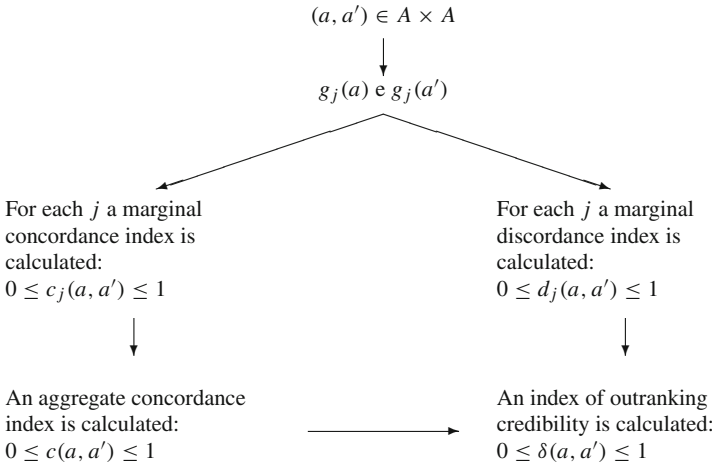
Let $A = \{a_1 : i \in I\}$ a finite set of alternatives, evaluated by a family of pseudo-criteria $g = \{g_j : j \in J\}$, then on the scale E_j of each criterion 3 thresholds are defined (q_j, s_j, v_j):

$$0 \leq q_j \leq s_j \leq v_j \tag{1}$$

respectively, indifference threshold, preference threshold and veto threshold. To each criterion is assigned a weight so that to obtain a vector of normalized weights $p = \{p_j : j \in J\}$, such as

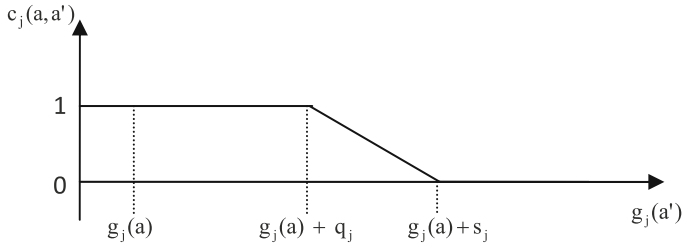
$$\forall j \quad 0 \leq p_j \leq 1 \quad \text{and} \quad \sum_{j \in J} p_j = 1 \tag{2}$$

During the first step, the model Electre III is based on the introduction of marginal indices of concordance and discordance for each criterion $j \in J$ and can be summarised as follows:



For each pair of alternatives (a, a') and for each individual criterion, the marginal concordance index is defined according to the comparison between the evaluation differences $g_j(a) - g_j(a')$ and the thresholds q_j and s_j, distinguishing the cases when the criterion is increasing (the judgment on the alternative improves as the criterion value increases) and decreasing (the judgment on the alternative worsens as the criterion value increases).

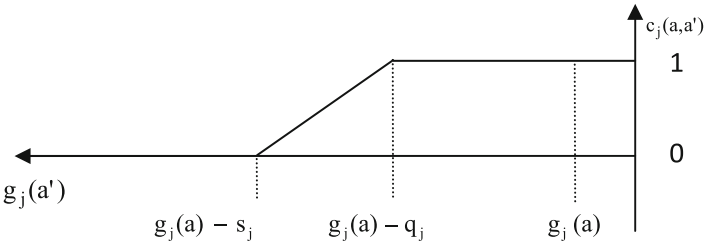
If the criterion is increasing, then



- $\forall j \in J$ if $g_j(a) \geq g_j(a')$, a outranks the action a' marginally, $aS_ja' \Rightarrow c_j(a, a') = 1$;
- if $g_j(a') \leq g_j(a) + q_j \Rightarrow c_j(a, a') = 1$ the two alternatives are indifferent;
- if $g_j(a') \geq g_j(a) + s_j \Rightarrow c_j(a, a') = 0$ the alternative a' outranks the action a;
- if $g_j(a) + q_j < g_j(a') < g_j(a) + s_j$ an interpolation has to be performed and it is possible to say that the alternative a' "weakly" outranks the alternative a. By considering, among the possible interpolations, a linear interpolation, then: $c_j(a, a') = \frac{s_j - (g_j(a') - g_j(a))}{s_j - q_j}$

(3)

If, instead, the criterion is decreasing then:



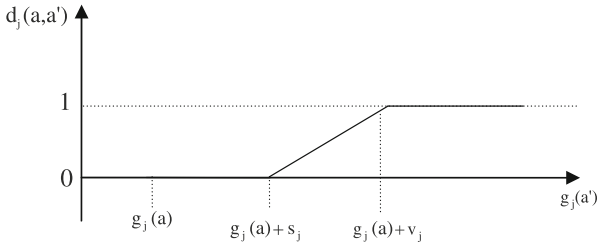
- $\forall j \in J$ if $g_j(a) \geq g_j(a')$, a outranks the action a' marginally, $aS_ja' \Rightarrow c_j(a, a') = 1$;
- if $g_j(a') \geq g_j(a) - q_j \Rightarrow c_j(a, a') = 1$ the two alternatives are indifferent;
- if $g_j(a') \leq g_j(a) - s_j \Rightarrow c_j(a, a') = 0$ the alternative a' outranks the alternayive a;
- if $g_j(a) - s_j < g_j(a') < g_j(a) - q_j$ an interpolation has to be performed and it is possible to say that the alternative a' "weakly" outranks the alternative a. By considering, as always, the linear interpolation, then: $c_j(a, a') = \frac{g_j(a') - (g_j(a) - s_j)}{s_j - q_j}$

(4)

In this way, a concordance matrix for each criterion is obtained; the elements of each matrix are the concordance indices among all the alternatives' pairs according to the considered criterion.

A similar reasoning concerns the marginal discordance indices, but in this case the veto threshold is introduced.

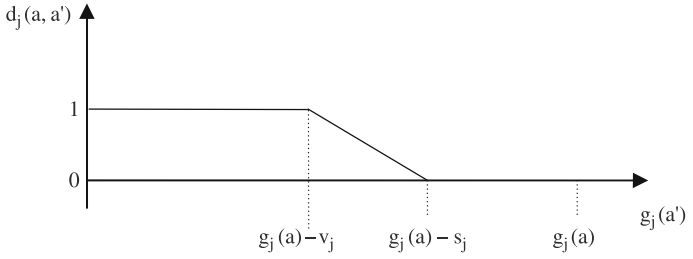
If the criterion is growing, then:



- if $g_j(a') \leq g_j(a) + s_j \Rightarrow d_j(a, a') = 0$ the two alternatives are indifferent;
- if $g_j(a') \geq g_j(a) + v_j \Rightarrow d_j(a, a') = 1$ the alternative a' outranks the alternative a;
- if $g_j(a) + s_j < g_j(a') < g_j(a) + v_j$ an interpolation has to be performed and it is possible to say that the alternative a' "weakly" outranks the alternative a. By considering the linear interpolation, then:
  $d_j(a, a') = \frac{(g_j(a') - g_j(a)) - s_j}{v_j - s_j}$

(5)

If, instead, the criterion is decreasing, then:



- if $g_j(a') \geq g_j(a) - s_j \Rightarrow d_j(a, a') = 0$ the two alternatives are indifferent;
- if $g_j(a') \leq g_j(a) - v_j \Rightarrow d_j(a, a') = 1$ the alternative a' outranks the alternative a;
- if $g_j(a) - v_j < g_j(a') < g_j(a) - s_j$ an interpolation has to be performed and it is possible to say that the alternative a' "weakly" outranks the alternative a. By considering the linear interpolation, then:
  $d_j(a, a') = \frac{(g_j(a) - g_j(a')) - s_j}{v_j - s_j}$

(6)

Once the J concordance matrices and the J discordance matrices are obtained (they are I×I matrices), one proceeds with the calculation of the I×I aggregated concordance matrix, whose elements are the weighted sum, with the weights initially assigned to each criterion, of the marginal concordance indices:

$$c(a, a') = \sum_{j \in J} p_j c_j(a, a') \tag{7}$$

By employing the aggregated concordance matrix and the discordance matrices one calculates the credibility outranking matrix, whose elements are obtained as illustrated below:

$$\left.\begin{array}{l} \text{if } \forall j d_j(a, a') = 0 \Rightarrow \delta(a, a') = c(a, a') \\ \text{if } \exists j \in J : d_j(a, a') > 0 \Rightarrow \\ \quad - \text{if } d_j(a, a') < c(a, a') \Rightarrow \delta(a, a') = c(a, a') \\ \quad - \text{if } \exists j^* \in J^* \subseteq J : d_j(a, a') \geq c(a, a') \Rightarrow \\ \qquad \Rightarrow \delta(a, a') = c(a, a') \times \prod_{j^* \in J^*} \left(\frac{1 - d_{j^*}(a, a')}{1 - c(a, a')} \right) \end{array}\right\} \tag{8}$$

Thus, the final order is established, i.e. the global classification of the alternatives. To this end the distillation algorithm[4] is used. One introduces a discrimination threshold $s(\delta)$, that is the maximal discrepancy between two credibilities, so that they can be still considered within the same order of magnitude. The distillation algorithm allows to extract from the credibility matrix the alternatives that will belong to the classification. Two distillation algorithms are applied: a descending and an ascending one. Descending distillation selects at first the best alternatives to end the process with the worst ones. On the contrary the ascending distillation selects first the worst alternatives to end the process with the best ones. Two complete pre-orders are therefore found on all the alternatives.

Within the credibility matrix, the maximum degree of credibility δ_0 is established for the extraction of the alternatives, equal to:

$$\delta_0 = \max_{(a, a') \in A^k} \delta(a, a') \tag{9}$$

[4] The distillation algorithm means that the alternatives are extracted from the credibility outranking matrix and put in a ranking.

that is the maximum among the values $\delta(a, a')$ at the k-th step (A^k is the credibility matrix at the k-th step); this determines a "value of credibility", and only the values $\delta(a, a')$ that are close enough to δ_0 will be considered. Hence, the discrimination threshold $s(\delta)$ is subtracted and, thus, δ_0' is calculated:

$$\delta_0' = \delta_0 - s(\delta) \tag{10}$$

the first level of separation is calculated δ_1, according to the set A^k:

$$\delta_1 = \begin{cases} \max\limits_{(a,a')\in\Omega} \delta(a, a'), \text{ where } \Omega = \left\{(a, a') | \delta(a, a') < \delta_0'\right\} \neq \emptyset \\ 0, \text{ if } \Omega = \emptyset \end{cases} \tag{11}$$

The qualification score $q^{\delta}(a)$ of each action $a \in A$, where A is a finite set of alternatives, is defined as the number of actions that are outranked by the action a_i minus the number of actions outranking it, i.e.:

$$\left.\begin{aligned} q_A^{\delta}(a) &= p_A^{\delta}(a) - d_A^{\delta}(a) \text{ where :} \\ p_A^{\delta}(a) &= |\left\{a' \in A : \delta(a, a') > \delta \text{ e } (\delta(a, a') - \delta(a', a)) > s(\delta)\right\}| \\ d_A^{\delta}(a) &= |\left\{a' \in A : \delta(a', a) > \delta \text{ e } (\delta(a', a) - \delta(a, a')) > s(\delta)\right\}| \end{aligned}\right\} \tag{12}$$

The descending distillation algorithm classifies the actions according to the maximal classification, following the rule:

$$q^{+} = \max_{a\in A^k} q^{\delta_1}(a) \tag{13}$$

and the following A^k subset is obtained:

$$D_1^{+} = \left\{a \in A : q^{\delta_1}(a) = q^{+}\right\} \tag{14}$$

where D_1^{+} is the first distillate from above and each class C_k^{+} will be built from above starting from this distillation unit. If D_1^{+} contains only an action, then $C_k^{+} = D_1^{+}$ and the above procedure is repeated on all the remaining actions for the next iteration. Otherwise, the algorithm is applied to all the D_1^{+}, generating, in this way, a sub-distillation until only one action will be left. The procedure is repeated starting from A^{k+1} and finishes when all actions in A have been attributed to a class. As previously, the result is a descending distillation. In the distillation from below, the

procedure is similar to the previous one but the selection is done according to the minimal qualification rule:

$$q^- = \min_{a \in A^k} q^{\delta_1}(a) \qquad D_1^- = \left\{ a \in A : q^{\delta_1}(a) = q^- \right\} \tag{15}$$

In this case, D_1^- is the first distillation unit from below and each class C_k^- will be built from below. Once obtained the two pre-orders $P(A)^+$ and $P(A)^-$ according to the distillation algorithms, the final order is established. The procedure to define the final order suggested by Schärlig [10] is an "intersection", according to the set theory, based on the following three rules. Firstly, an action cannot be ahead of another action in the final order, unless it is ahead of it, in one of the two preliminary-orders $P(A)^+$ or $P(A)^-$ and ahead of it or ex equo in the other one. Secondly, two actions cannot be ex equo in the final order unless they belong to the same class in both classifications (from below and from above). Thirdly, two actions are incompatible in the final order if one is ahead the other one in a classification (from below or from above) and is behind it in the other one. The result can be represented as a graph.

12.4 Groups of Disciplines Ranking

By putting together the data provided by the first part of the analysis, the second paragraph of this chapter aimed to identify the variables (criteria) to be used in the Electre III method. The method has been applied to various groups of disciplines according to: g_1, graduation degree; g_2, duration of effective studying period; g_3, employment percentage; g_4, percentage of full time and undermined length of time contracts; g_5, monthly net income; g_6; coherence study-work; g_7, academic curricula adequacy to the current job; g_8, (overall) satisfaction with the current job; g_9, percentage of employed graduates with decisional autonomy; g_{10}, percentage of employed graduates with job responsibilities; g_{11}, social background[5]; g_{12}, percentage of graduates satisfied with their degree; g_{13}, percentage of employed graduates working in Sicily.

If we consider the code of the groups of disciplines and the criteria associated to them then we have the performance matrix that is reported in the next page.

The matrix of the weights and thresholds correlated to the criteria is, instead, the following (see (1) and (2)):

The weights and the veto thresholds have been attributed according to the characteristics of the territory both from the point of view of the labour market and the university's offer [4]. In particular the veto thresholds are quantified after the explorative analysis of STELLA data and, also, in accordance of the evidences expressed from governmental and job market authorities often participant in round tables on this subject.

[5] See the second paragraph for the explanation of the use of the variable as criterion.

Performance matrix by group of disciplines

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}	g_{13}
a_1	103.3	1.7	24.2	16.3	696	6.5	7.0	6.5	83.8	68.9	3	73.9	83.8
a_2	107.3	1.5	30.1	16.9	957	6.5	7.0	6.5	73.5	50.4	4	64.6	92.7
a_3	100.3	1.5	53.9	41.9	1,139	5.0	5.5	8.0	55.6	0.0	3	58.1	88.9
a_4	103.2	1.7	23.7	14.3	767	6.0	5.0	8.5	71.2	25.1	3	76.8	89.1
a_5	107.6	1.2	60.0	0.0	417	5.5	3.5	9.5	33.3	33.3	2	40.0	66.7
a_6	105.7	1.6	10.1	5.0	726	5.5	4.5	6.0	49.2	32.1	4	65.2	100.0
a_7	102.4	1.6	20.7	11.4	655	3.5	5.0	7.0	46.8	17.6	3	70.1	78.1
a_8	104.0	1.6	18.1	13.9	1,069	6.0	5.5	7.5	63.8	28.4	4	78.2	66.7
a_9	102.2	1.4	47.8	17.1	671	6.0	6.5	8.0	63.2	11.4	3	52.3	73.0
a_{10}	106.7	1.5	19.6	6.2	548	4.0	4.0	7.0	66.7	30.8	4	62.1	81.4
a_{11}	105.2	1.6	34.1	16.2	678	6.0	6.0	7.5	47.1	25.5	4	48.4	87.4
a_{12}	107.8	1.1	81.5	68.7	1,169	7.5	7.5	9.0	75.6	33.8	3	73.4	79.9
a_{13}	105.8	1.5	31.0	18.3	819	5.5	5.0	7.5	60.8	34.1	3	68.6	74.0
a_{14}	104.4	1.5	21.0	3.9	505	5.5	5.0	6.0	46.0	0.0	3	67.9	96.4
a_{15}	103.4	1.7	24.7	19.8	980	6.0	6.0	7.0	77.0	46.0	3	74.7	65.5

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}	g_{13}
Direction[a]	C	*D*	C	C	C	C	C	C	C	C	C	C	C
Weight	0.11	0.06	0.13	0.06	0.06	0.11	0.06	0.11	0.06	0.06	0.06	0.06	0.06
Threshold (q)[b]	2	0.1	5	5	100	1	1	1	5	5	1	5	5
Threshold (s)[c]	4	0.2	15	15	200	2	2	2	15	15	2	15	15
Threshold (v)[d]	6	0.4	25	25	600	3	3	3	25	25	3	25	25

[a]C = growing, D = decreasing.
[b]Threshold of indifference.
[c]Threshold of preference.
[d]Threshold of veto.

The weights were an individual expression of preference expressed by the referee of STELLA initiative (see http://stella.cilea.it) [7].

Starting from the matrix of preference and taking into account the pre-established thresholds for each criterion, marginal concordance and discordance indices among all the potential pair of alternatives, have been calculated according to each criterion following the (3), (4), (5), and (6).

Based on the matrices of marginal concordance and by taking into account the weights initially established by the decision-maker, a matrix of aggregated concordance is built (Table 12.4). Its elements are given by the weighted sum of the indices of marginal concordance (see (7)).

The elements of the matrix of aggregated concordance are then used together with the matrices of marginal discordance for calculating the outranking indices of credibility (Table 12.5), the starting point of the final ordering.

By establishing a discrimination threshold $s(\delta) = 10$ such as the minimum significative difference between two credibility indices and by applying to this matrix the distillation algorithm in a ascending and descending approach, two pre-orders can be found in (16) and (17).

Table 12.4 Matrix of aggregated concordance indices

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}
a_1	0.00	0.75	0.65	0.89	0.59	0.92	1.00	0.94	0.76	0.89	0.94	0.47	0.91	0.94	0.94
a_2	0.88	0.00	0.71	0.85	0.70	0.99	1.00	0.94	0.82	1.00	1.00	0.56	1.00	1.00	0.97
a_3	0.68	0.68	0.00	0.77	0.70	0.78	0.89	0.77	0.95	0.79	0.83	0.29	0.80	0.85	0.76
a_4	0.83	0.65	0.70	0.00	0.68	0.92	1.00	0.94	0.78	0.88	0.93	0.44	0.89	0.97	0.87
a_5	0.58	0.52	0.64	0.67	0.00	0.70	0.72	0.65	0.71	0.74	0.63	0.41	0.72	0.80	0.65
a_6	0.64	0.58	0.62	0.65	0.70	0.00	0.92	0.72	0.56	0.88	0.75	0.32	0.73	0.92	0.59
a_7	0.71	0.39	0.61	0.71	0.42	0.70	0.00	0.72	0.61	0.77	0.69	0.18	0.57	0.83	0.69
a_8	0.78	0.66	0.75	0.92	0.61	0.94	0.96	0.00	0.83	0.90	0.81	0.39	0.88	0.94	0.86
a_9	0.79	0.63	0.80	0.81	0.62	0.75	0.93	0.82	0.00	0.78	0.83	0.35	0.76	0.87	0.77
a_{10}	0.61	0.56	0.70	0.68	0.65	0.84	0.98	0.72	0.65	0.00	0.65	0.35	0.75	0.88	0.58
a_{11}	0.82	0.75	0.70	0.88	0.66	0.88	0.94	0.82	0.79	0.89	0.00	0.38	0.84	0.92	0.76
a_{12}	0.92	0.89	0.98	0.97	1.00	0.94	1.00	1.00	1.00	1.00	0.99	0.00	1.00	0.94	0.96
a_{13}	0.79	0.75	0.69	0.89	0.70	0.94	1.00	0.91	0.84	0.98	0.95	0.40	0.00	0.94	0.86
a_{14}	0.72	0.61	0.61	0.65	0.57	0.88	0.90	0.70	0.54	0.86	0.69	0.21	0.65	0.00	0.75
a_{15}	0.00	0.75	0.65	0.89	0.59	0.92	1.00	0.94	0.76	0.89	0.94	0.47	0.91	0.94	0.94

Table 12.5 Outranking credibility indices

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}
a_1	0.00	0.75	0.00	0.89	0.00	0.92	1.00	0.94	0.44	0.89	0.93	0.00	0.91	0.94	0.94
a_2	0.88	0.00	0.00	0.85	0.00	0.99	0.99	0.94	0.81	1.00	1.00	0.00	1.00	1.00	0.97
a_3	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.00	0.94	0.00	0.00	0.00	0.00	0.84	0.00
a_4	0.00	0.00	0.00	0.00	0.00	0.92	1.00	0.94	0.32	0.88	0.93	0.00	0.89	0.97	0.87
a_5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.63	0.00	0.00	0.00	0.00
a_6	0.00	0.10	0.00	0.56	0.00	0.00	0.92	0.72	0.00	0.88	0.27	0.00	0.72	0.92	0.00
a_7	0.00	0.00	0.00	0.14	0.00	0.70	0.00	0.72	0.00	0.77	0.69	0.00	0.57	0.83	0.00
a_8	0.00	0.00	0.00	0.92	0.00	0.00	0.96	0.00	0.00	0.90	0.81	0.00	0.88	0.00	0.86
a_9	0.00	0.00	0.06	0.18	0.39	0.00	0.93	0.00	0.00	0.78	0.83	0.00	0.75	0.87	0.00
a_{10}	0.00	0.00	0.00	0.68	0.00	0.84	0.98	0.52	0.00	0.00	0.65	0.00	0.75	0.88	0.58
a_{11}	0.00	0.00	0.00	0.00	0.00	0.88	0.94	0.00	0.79	0.88	0.00	0.00	0.84	0.91	0.00
a_{12}	0.00	0.88	0.98	0.97	1.00	0.94	1.00	1.00	1.00	1.00	0.98	0.00	1.00	0.94	0.96
a_{13}	0.00	0.75	0.20	0.89	0.00	0.00	1.00	0.91	0.84	0.98	0.95	0.00	0.00	0.94	0.85
a_{14}	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
a_{15}	0.87	0.00	0.00	0.88	0.00	0.00	0.94	0.96	0.74	0.81	0.85	0.00	0.88	0.00	0.00

According to Schärlig [10], the complete final order results from the two pre-ordering intersection (Fig. 12.1). The final order consists in observing for each alternative how this relates to others in the two pre-order.

$$P(A)^{+} = \{a_{12}\} \succ \{a_2\} \succ \{a_1\} \succ \{a_4, a_{13}\} \succ \{a_{11}\} \succ \\ \succ \{a_3, a_6, a_8\} \succ \{a_{10}\} \succ \{a_9\} \succ \{a_{15}\}\{a_5, a_7, a_{14}\} \tag{16}$$

$$P(A)^{-} = \{a_{12}\} \succ \{a_2\} \succ \{a_1\} \succ \{a_3, a_4, a_5\} \succ \\ \succ \{a_9, a_{15}\} \succ \{a_8, a_{13}\} \succ \{a_6, a_{10}\} \succ \{a_{11}\} \succ \\ \succ \{a_{14}\}\{a_7\} \tag{17}$$

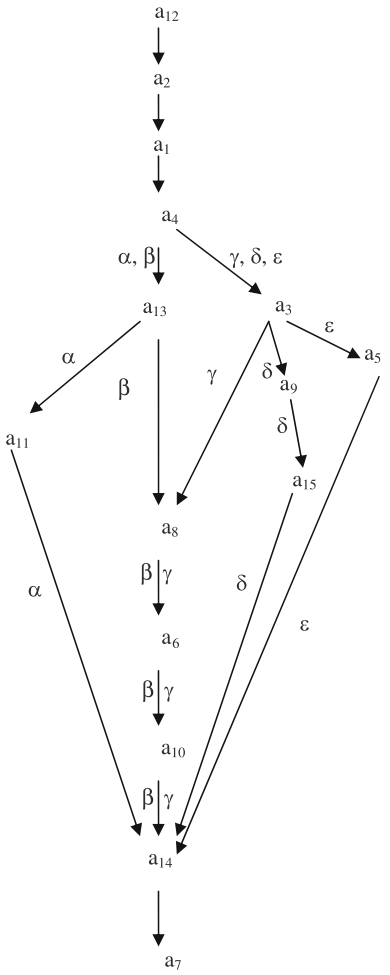


Fig. 12.1 Representation of the final ordering with α, β, γ, δ, ϵ paths

The lack of comparison between certain groups generated 5 different rankings (α, β, γ, δ, ϵ), following different paths after the Economy and Statistics group and reuniting with the Psychology group. The group of disciplines a_{12} belongs to the first position in both pre-orders and consequently can be at the top of the final pre-order. Being the second in both pre-orders, a_2 follows immediately afterwards and so on. The alternatives a_3 and a_{13} outrank each other, thence, non-comparable, just as a_3 does not compare with a_{11}, a_{13} besides observing that a_3 does not compare with a_5, a_9, a_{15}, a_{11} does not compare with a_3, a_5, a_9, a_{15}, a_8, a_6 e a_{10}, and so on.

According to this ordering, the Medicine group is better valued, followed by Architecture, Agrarian Studies and Economy and Statistics. Despite the fact that the groups of Chemistry and Pharmaceutics and Political and Social Studies place

on the fifth position, the groups are incomparable considering that the partial pre-order $P(A)^+$ a_{13} outranks a_3 while in the pre-order $P(A)^-$ a_3 outranks a_{13}. The ranking contains other incomparable groups: Health and Physical Education, Linguistic Studies and Literature, groups that mutually outrank in the two pre-orders. The last positions in the ranking are occupied by the Psychology and Law Studies group. The explanation behind the two groups' last positions may be linked to the fact that, according to most of the variables taken into account, the two groups occupy the last positions and strongly differentiate themselves from the others (Sect. 12.2).

12.5 Final Remarks

The results of the method employed in this chapter shed lights on a fundamental organisational issue, allowing thus the stakeholders to understand if the University has succeeded in transferring to the students its academic curricula and simultaneously maintain a high level of attention on the high education system's needs both at a national and local context. The knowledge and skills developed by the graduate and his/her placement result from both the individual commitment and the efficiency of the academic curriculum. The measurement of the efficiency of the academic curriculum is a delicate issue, mainly if the applied method doesn't take into account the complexity of the argument. This method has, of course, various weaknesses such as the subjectivity of the criteria weighting, the establishment of the thresholds and impossibility of "measurement" of the distance between the alternatives in the final ordering. Nevertheless, various strong points can be mentioned: the multi-criteria approach encompasses various aspects directly linked to the topic in question, the possibility to weight the criteria and to establish thresholds provides the method with a major flexibility and, thus, it can be adapted to various needs and requirements. By establishing weights and thresholds, this standard method facilitates the decision-making process without ambiguity and, last but not least, the methods easily adapt to statistic softwares (for example R).

Our approach, adapted to the university arena, aims to provide a ranking of the groups of disciplines according to the graduates' placement. This is the first step for providing the universities a strategic tool able to guarantee a qualitative improvement of the system. More specifically, it aims at providing a support for those in charge with the planning of the various degrees and the guiding of the graduates on the job market. Thus a follow-up agenda of this analysis might imply the development of strategic actions for the improvement of the quality in synergy with the available resources and the preferences of the stakeholders, among which the most relevant are the graduates and the enterprises. Our analysis lays the basis for further developments such as the enterprises opinions on the graduates and, thus, a critical overview of the professional efficiency according to the graduates' specific academic curricula.

References

1. Capursi V, Librizzi L (2008) La qualità della didattica: indicatori semplici o composti? In: Capursi V, Ghellini G (eds) Dottor Divago. Discernere valutare e governare la nuova università. Franco Angeli, Milano
2. CILEA (ed) (2008) Laureati Stella, Indagine occupazionale post-laurea, anno solare 2006. CILEA
3. Enea M (ed) (2006) Metodologie multicriterio per la selezione dei progetti in ambito F.S.E.. Università degli studi di Palermo, Regione Siciliana
4. Enea M, Giambalvo O (2002) The statistical informative system for the university. 23rd conference on regional and urban statistics and research, Lisbon, Portugal, 12–15 June
5. Enea M, Giambalvo O, Morreale G (2005) La valutazione dei percorsi formativi dei laureati attraverso l'uso del modello multicriterio Electre III. In: Crocetta C (ed) Modelli statistici per l'analisi della transizione Università-lavoro, Determinazione e previsione di rischi sociali e sanitari, N. 7. CLEUP, Padova
6. Keeney RL, Raiffa H (1976) Decision with multiple objectives; preferences and value tradeoffs. Wiley, New York, NY
7. Norese MF (2006) ELECTRE III as a support for participatory decision making on the localisation of waste-treatment plants. Land Use Policy 23:76–85
8. Roy B (1978) ELECTRE III: Un algorithme de classements fondé sur une représentaion floue des préférences en présence de critères multiples. Cahiers du Centre d'Etudes de Recherche Opérationnelle 20:3–24
9. Roy B (1996) Multicriteria Methodologie for Decision Aiding. Kluwer Academic Publishers, Dordrecht
10. Schärlig A (1996) Pratiquer Electre et Prométhée. Un complément à dèdicer sur plesieurs critères. Presses Polytechniques et Universitaires Romandes, Lausanne
11. Simon HA (1997) Models of bounded rationality. The MIT Press, New York, NY

Chapter 13
Competences and Professional Options of the Italian Graduates: Results from the Textual Analysis of the Degree Course Information Data

S. Balbi, C. Crocetta, M.F. Romano, S. Zaccarin, and E. Zavarrone

13.1 Introduction

The present chapter, developed within a research project promoted by the National Committee for the Evaluation of the University System (CNVSU, the complete report is available on www.cnvsu.it), analyses the Italian university offer, focusing on the communication adopted by the Universities to publicise their objectives and the results expected, also with respect to the potential employments. The analysed documents are the course information sheets of all the 3-year degree courses and some specialised (2-year) degree courses, contained in the OFF.F database of the Ministry of the University (MIUR) for the academic year 2005/2006. The research pursued the following aims:

1. reading the education offer focusing mainly on the competences foreseen for the graduates;
2. reading the foreseen job prospects for the graduates in a course;
3. analysing the consistency of the competences provided with the foreseen employment prospects;
4. analysing the consistency between the competences acquired by the 3-year graduates and the competences offered to whom decides for continuing with a 2-year degree course specialisation.

The analysis of points 1–3 is carried out through the use of typical *text – mining* procedures, while for point 4 the reference is to methods developed for the analysis of *multi-linguistic corpora.* In Sect. 13.2, a short overview on the adopted methods is presented, while in the following sections some general reflections on the topics are made, referring to the principal results for the degree in *Statistics* (class code 37), as an example.

S. Balbi (✉)
Dipartimento di Matematica e Statistica, Università di Napoli "Federico II", Napoli, Italy
e-mail: simona.balbi@unina.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_13,

13.2 Textual Analysis of the University Education Offer

For the university system, as well as for any other aspect of our society, the topic of *communication* has gained – over the last few years – a growing importance. In addition to the obvious considerations connected with the "Internet" revolution, which makes new tools for interaction available, there are other stimuli which derive more specifically from the transformations that the Italian University system has undergone; transformations which are still going on.

On the one hand, the University belongs to the Public Administration, to whom *transparency* is required, in terms of the visibility of its work. This requisite is therefore required for accreditation of a university course. However, on the other hand, the communication regards also the autonomy which is granted to each single institution, in order to achieve a healthy competition. In this sense, therefore, it is a *communication* to be intended in a corporate sense, aimed at hitting one's own target market.

A further element to be considered derives from the indications elaborated within the Bologna Process with the purpose of creating a European common framework, whose practicability depends not only on having common objectives and tools, but on sharing their formulation.

There are, therefore, many forms of communications, starting from the advertising campaigns of the Universities, but also of the individual programs, in different *media*. Here, as mentioned, the attention is focused on those aspects which are more institutional, such as the presentations that each degree course sends to the Ministry, to be inserted in the database of the education offer OFF.F. These *schede* are the course information data sheets with a structure defined by the Ministry, which contain a variety of information. In particular, there are two sections which can provide the reader (the school-leaver or his/her family), a picture easily understandable by non specialists: these are two paragraphs written in free-form, one about the education objectives and the other about the employment prospects.

The subsequent analysis refers to the course information data sheets about the education offer of the Italian university system in the first phase of the fulfillment of the so-called $3 + 2$ reform. Of the 2,339 3-year degree courses activated for the year 2005/2006 within the 42 classes (see Table 13.1) defined in the Ministerial Decree of the 4th August 2000, the portions of text relative to the *specific educational objectives (obiettivi formativi caratterizzanti*) and *job possibilities* (*ambiti occupazionali previsti*) were examined.

The compilation of the *schede* is, first of all, a requisite for the institution of a degree course. In this sense then, an administrative obligation exists. There is also a sort of framework, compiled by the Ministry, where, for each class of degree, the aims of the centrally specified qualifying program (*obiettivi formativi qualificanti*) and the areas of prospective employment are foreseen. The declaration that one course belongs to a specific class should, therefore, imply the adhesion to the education aims of that class.

It is, therefore, for the purposes of *competition* that each degree course is required to differentiate itself and give more space to the employment perspectives (in this sense the new reform in force from year 2008/2009 is more explicit on this point).

Table 13.1 Three-year degree classes and number of courses. Year 2005/2006

Class codes and names	Number of courses activated
1 – Biotechnologies	53
2 – Legal Services	50
3 – Linguistic Mediation	41
4 – Architecture and Construction Engineering	62
5 – Literature	68
6 – Social Work	47
7 – Town, Regional and Environmental Planning	23
8 – Civil and Environmental Engineering	81
9 – Information Technology	141
10 – Industrial Engineering	168
11 – Modern Languages and Civilizations	66
12 – Biological Sciences	53
13 – Cultural Heritage Studies	74
14 – Communication Studies	73
15 – Political Science and International Relations	56
16 – Earth Sciences	31
17 – Business Economics	169
18 – Education Sciences and Teacher Education	66
19 – Public Administration	33
20 – Agriculture, Food Industry and Forestry	114
21 – Chemistry	61
22 – Aviation and Maritime Navigation	1
23 – Visual Arts, Music, Performing Arts and Fashion Studies	33
24 – Pharmacy and Industrial Pharmacy	62
25 – Physics	51
26 – Computer Science	56
27 – Environmental Sciences	59
28 – Economics	95
29 – Philosophy	46
30 – Geography	8
31 – Law	73
32 – Mathematics	48
33 – Physical Education and Sport	34
34 – Psychology	49
35 – Social Sciences for Co-Operation, Development and Peace	19
36 – Sociology	23
37 – Statistics	32
38 – History	29
39 – Tourism	23
40 – Animal Husbandry	27
41 – Technologies for the Conservation and Restoration of Cultural Assets	21
42 – Industrial Design	20
Total	2, 339

Source: Database OFF.F, year 2005/2006 http://offf.miur.it/index.html

However, although it is necessary to compile these forms, precise indications on what they should exactly say have not yet been issued.

On the theme of the Europeanisation of the system, the putting into effect of the *Bologna Process* pushes towards the arrangement of course information data sheets according to the so-called "Dublin descriptors", built on the following elements:

- knowledge and understanding
- applying knowledge and understanding
- making judgements
- communication skills
- learning skills

As far as the methodology adopted is concerned, the first descriptive analysis of the education offer of the whole university system was carried out from the point of view of *text mining*, given the mass of texts to be examined. The following analyses are based on the multivariate statistical techniques, opportunely adapted to the circumstance of having to work on textual data, instead of numeric data.

The first consequence is that our databases (made up of the *corpus* of the two sections of the course information data sheets) are unstructured and need to be organised in order to deal with the typical data structures used to perform textual analyses. Here we think of the so-called lexical tables, i.e. matrices cross-classifying courses and terms used for describing them. In building those matrices, we have to make different choice related to the minimal units to adopt, the frequency thresholds to consider them interesting for our aims. Moreover, we need to pre-treat the documents by lexicalisation for avoiding trivial cases of ambiguity. The documental base consists of about 2,400,000 occurrences, while concerning with documents length, it is worth noting that all the texts where written for being included in the OFF.F pre-defined window. Therefore, there are no interesting differences in their length.

The results obtained in our analysis have to be intended as stimuli for further investigation. The aim of exploration of the present work must therefore be borne in mind while reading the following.

The analysis suggested here is divided into three principal sections, characterized by different aims as well as by different dimensions of available data. The first aim is to understand the *specific education offer* and the *employment perspective expected* from the 2,339 degree courses (Sect. 13.3). Afterwards, the aspects about the consistency among these texts are analysed (Sect. 13.4), and then we deal with the themes connected to the 3 + 2 structure (Sect. 13.5). For reasons of space, only results for class 37 (*Statistics*) are shown.

13.3 Identification of the Competences Offered and the Job Possibilities

The tools chosen for this analysis, given the large dimensions of the *corpus* to be analysed, are those of the analysis of the textual data [6] and of *text mining*, using the *software* TALTAC. It was furthermore decided to treat separately the texts about

the competences and those referring to the employment possibilities, because of the different dimensions of the two bases of data, as the description of the job opportunities expected has received less interest at the local level. Thus, also different methodologies were used.

We started from the analysis of the specific educational objectives (competences), in a typical perspective of *mining*, using a two stage procedure of knowledge extraction [2]: the document is considered as a complex data, with a hyerachical structure within itself, the levels of which are made up of sentences, made up of phrases, made up of words (in their turn, made up of characters). To analyse the words (or groups of words), an initial selection of the interesting sentences was made, as they were relative to competences (*specific educational objectives*). In order to do this, we used, in a content analysis perspective, the identification "a priori" of *key-verbs* (e.g., acquire, apply, learn, have, know, demonstrate, carry out, be able to supply, teach, offer, operate, participate, possess, utilize, use, ...) which it is assumed connect – within a sentence – words referring to the competences offered during the course of studies. These verbs were used to obtain a drastic reduction of the material to be analysed, as the *scheda* pertains also to organization and sometimes to considerations of a wider nature, independent from an analysis of the competences offered.

A field identifying a sentence as interesting (i.e. referred to competences) was built, through logical rules of the type: IF *acquire* = 1 THEN *interesting* = YES OR *apply* = 1 THEN *interesting* = YES OR ... OR ...). Therefore, all those sentences where *interesting* is different from YES were eliminated, and the textual analysis performed on the *sub-corpus* containing the sentences relative to competences (about 800,000 occurrences, with a 17,200 terms vocabulary).

Once the fragments of interest were identified, the construction of the repeated segments (defined as sequences of recurring words with a frequency higher than a fixed threshold (see [6]) was carried out as well as that of the nouns. For building the repeated segments, only the terms with a frequency higher than 5 were considered. Furthermore, the length of the segments was limited to a maximum of 4 terms.

In this way, the linguistic structures which characterised most of the courses belonging to the same class of degree were identified: e.g. short lists of terms used in a class of degree could suggest poor lexicon, somehow connected with a lesser (or more homogeneous?) education offer, from the point of view of the contents.

This analysis shows the necessity for more stringent indications from the Ministry about the contents to be assigned to these course information data sheets, so that they can work as a valid tool for an informed choice for those wishing to get useful information about the degree courses in which to enroll, and also for the transparency about the real education contents of a course, both in terms of competitiveness and autonomy.

The differences among the single classes of degree appear clearly: some classes of degree have an objective which is clearly defined and limited (e.g. *Industrial Design*), while others seem to want to cover large fields of human knowledge (e.g. *Agriculture, Food Industry and Forestry*). Furthermore, there is obviously the work carried out by each single University. Apart from the extreme case of class 22, in

which only one degree course was activated, more generally speaking, the different cultural matrices of the different courses expressed themselves in ways which are clearly typical of each area and definitely very variable.

The most concise vocabularies are those of the Classes 2 – *Legal Services*, 5 – *Literature*, 11 – *Modern Languages and Civilizations*, 13 – *Cultural Heritage Studies*, 18 – *Education Sciences and Teacher Education*, 23 – *Visual Arts, Music, Performing Arts and Fashion Studies*, 28 – *Economics*, 31 – *Law* and 35 – *Social Sciences for Co-operation, Development and Peace*. These vocabularies are substantially limited to a few words specific to each discipline, such as for example "legal" and "rules" for Class 2.

The Class 1 – *Biotechnologies* has a vocabulary which is well-constructed and not limited to terms which are specific to the discipline, but also relative to other disciplines ("chemistry", "physics", "mathematics", "informatics", "statistics"). Even larger vocabularies appear the Classes 4 – *Architecture and Construction Engineering*, 12 – *Biological Sciences*, 20 – *Agriculture, Food Industry and Forestry*, 29 – *Philosophy*, 40 – *Animal Husbandry*. These vocabularies are also characterized by a homogenous distribution of the occurrence of the words.

In general, it is appropriate to point out how the competences expressed in all the vocabularies are of a general character, though sometimes supported with terms which are specific to the discipline, such as "constructions", "building site", "technical and economic feasibility" (Class 4), "management of human resources" (Class 19), "commercialisation" and "technical assistance" and "quality control" (Class 20), "therapeutical preparations" (Class 24), "problems analysis" and "techniques of argumentation" (Class 29), "mathematical models" and "computational tools" (Class 32), "services to the individual" (Class 34), "properties of materials" (Class 41), "engineering", "planning" and "communication interfaces" (Class 42).

A noteworthy case is that of the degree courses in Engineering (8 – *Civil and Environmental Engineering*, 9 – *Information Technology*, 10 – *Industrial Engineering*): their vocabularies are overlapping (as well as the corresponding ministerial course information data sheets) and also the frequency of the occurrence often respects the same order, so that some doubts appear about the real need for this tri-partition.

The vocabulary of the competences for the class of degree 37 (*Statistics*) are quoted in Table 13.2. For this class, the vocabulary appears adequate, taking into account the presence of terms which describe competences which are specific to the discipline (*surveys, data sets, measurement, data analysis, data collections, statistical analysis, statistical, experimental, observational methods*) or general competences, but not generic ones (*measurement, logical-conceptual, method of research, technological-experimental*). There are also some terms referring to the areas of application or of employment (*economical, informatics, biomedical*).

As far as the employment perspectives are concerned, starting from those words which are considered to be of higher relevance on the basis of expert knowledge (*content bearing words*), we moved in a *data driven* manner, with higher attention to the linguistic aspects. The analysis carried out allows the identification of key words of the professions expected at the end of the course by the ministerial tables

Table 13.2 Vocabulary of the competences for Class 37 – Statistics (words with frequency > 1)

Words	Total occurrences	Words	Total occurrences
Statistics	102	Applications	8
Tools	45	Data analysis	8
Statistical	42	Methodological	8
Science	30	Data collection	8
Economical	28	Logical-conceptual	7
Technical	25	Statistical analysis	7
Social	24	Execution	7
Surveys	16	Method of research	4
Planning	15	Experimental	4
Applications	12	Biomedical	3
Statistical disciplines	11	Observational	3
Databases	10	Statistical methodology	3
Informatics	10	Culture of working	2
Measure	10	Technical-experimental	2

and taken up by the different courses of study, in the light of their own education offer, planned taking into account the potential demand, their internal resources and the characteristics of the geographical area.

The "repeated segments" describing the employment perspectives of the different classes were analysed, as they appear in the declarative statements of the individual degree courses. Their importance is measured by a lexicometric index IS [8], which measures the absorption of the single words in a segment.

The index is given by the rate between the frequency of the segment and the frequency of the single words which make it up, for the number of meaningful words of the segment (not articles, prepositions, conjunctions, adverbs). This tool allows us to distinguish the degree courses for the width of possible professional roles upon graduation (which is not necessarily a measure of the marketability of the degree in the world of work). Furthermore, it is necessary to distinguish the classes of degree with an exclusive professional role (doctors) and those where there are coexisting roles, such as, for example, those of Economics, which are defined generalist, as they offer a wide range of professional roles upon graduation.

From the textual analysis carried out on the employment possibilities, the picture appears to be quite varied. Given the high variability existing among the different course information data sheets proposed by MIUR, and among the latter and the degree programs activated at the different universities it is difficult to be able to gather any regularity.

The analysis carried out has highlighted the need to dedicate more attention to the description of the professional figures coming out from the different study paths proposed, also for the purpose of the verification of the congruity between the professional profiles expected and the education paths proposed (a need understood in the formulation of the *scheda* for the year 2008/2009). Furthermore, it would be opportune to add, on the *scheda*, together with the qualitative information, also some quantitative data about the effective possibilities of employment offered by the different paths and about the conformity of their contents to the working activities

carried out. In this way, useful elements could be provided to the potential student allowing him/her to make a well-informed choice.

Going back to the formulation of the course information data sheets, the duplicity of their objective emerges here even more clearly: providing *marketing* and *transparency*, in order to guarantee the potential student more information about the realistic employment perspectives, beyond simply catching labels. A higher degree of structuring (and standardization) of the course information data sheet of presentation of the employment perspectives would not affect the organizational autonomy of the Universities, however it would force them to carry out a more thought through evaluation about the future work placement of their own graduates.

Once more, the result of the analysis carried out for the class 37 is presented. Among the possible employment outcomes for the graduates in statistical sciences, the ministerial decree proposes the professional activities in the field of learning and spreading of statistical knowledge, the management of qualitative and quantitative information on behalf of public or private companies and research institutions. The degree programs stress, instead the possibilities of employment in the fields of the planning and management of the *informative systems* and of the *professional activities* either as self-employment or as an employee of *private companies* or *research institutions*, not stressing work activities with specific contents referring to the professional competences of a statistician.

13.4 Consistency of the Education Paths with Employment Perspectives

For the evaluation of the consistency between competences offered and employment perspectives, we used a variant of Lexical Correspondence Analysis (LCA, [6]), one of the most common methods for the analysis of textual data. Aiming at analysing in depth the dependence of the description of the foreseen employment with the figure of the graduate described in the section "education objectives", the non symmetrical analysis of the co-occurrences [4] was adopted. This is a factorial technique aimed at the study of matrices having as generic element the frequency with which a term is used to describe the same object, in two different conditions, one logically, or temporally, antecedent to the other one. It is a variation of the non symmetrical lexical correspondences analysis (NSLCA), proposed by Balbi [1] for studying a relation of dependence of the vocabulary from a partition induced on the corpus.

NSLCA, instead of decomposing the association index χ^2, as usual in Correspondence Analysis, deals with the predictability index τ_b, proposed by Goodman and Kruskal (see also [5]). Let us consider an aggregated lexical table F (K,V), where K is the number of categories considered for documents and V is the number of terms in the vocabulary. F general element is $f_{kv,}$ relative frequency of the v-th term in the k-th category. NSLCA studies the V conditional distribution $f_{kv}/f_{.v}$, with respect to the independence hypothesis, given by $f_{k.}$, where $f_{.v}$ and $f_{k.}$ are respectively the column- and row- marginal distribution. From a geometrical viewpoint, distances between the categories are measured in the usual Euclidean metric,

while distances between terms are measured in a weighted Euclidean metric. Compared with Lexical Correspondence Analysis, an interesting effect consists in giving a lower importance to infrequent words.

In this case, we analysed the 32 courses belonging to the class *Statistics*, in order to represent the internal consistency, by measuring how much the presence of one word in the section "specific education objectives" has influenced its use in the description of the "job possibilities". The two lexical tables cross-classifying *degree courses* (documents) and *objectives* (terms) and *degree courses* and *professional roles* (terms) were collapsed into a single matrix (terms × terms) *professional roles* × *objectives*. The generic element of the resulting matrix is given by the number of times the two terms are used jointly in the two corpora. The risk of obtaining a sparse matrix, lacking in information for the presence of too many 0's, was avoided by eliminating from the vocabulary all the so-called functional part-of-speech (i.e. conjunctions, articles, prepositions, adverbs), and limiting the analysis to the meaning words: nouns, adjectives, verbs.

In this analysis there are specific reading rules for the interpretation of the factorial planes:

- the scattering of words around the origin shows the strength of the dependence of the vocabulary from the information given on the documents, inducing their partition;
- two words are near, if they similarly depend from the partition;
- two categories of the partition variable are close if they similarly influence the use of words;
- a word is so far from the origin, the more it depends on the partition;
- a modality of the external variable is so far from the origin, the more it influences the use of the words.

The differences in metrics produces some important consequences. First of all, compared with usual LCA, the words with higher frequency are more "important" in the characterization of the documents, than those less frequent. The second characteristic is that the "joint plot" typical of LCA (the unique factorial plane where documents and terms are simultaneously represented) is difficult to be read, because of the different scales (as in principal components analysis), even though the origin (which represents the hypothesis of independence), the orientation of the axes and the percentage of the explained variability are common. The two factorial planes are therefore represented separately: Figure 13.1 represents competences, while Fig. 13.2 the employment perspectives.

It is worth noting how the reform has granted the opportunity to increase the offer of statistics training, motivating a wider differentiation of the degree courses in the different faculties. Mainly in those Universities without a Faculty of Statistics, the reform has meant planning something radically new. Unexpectedly, one of the roots which appears more frequently, on the graph describing competences, is *mercat-* (market), together with *marketing*. To these are added other words indicating further areas of application of statistical tools, such as *comunicazione* (communication), *finanza* (finance) and *assicurazioni* (insurance) at the bottom and *qualità* (quality)

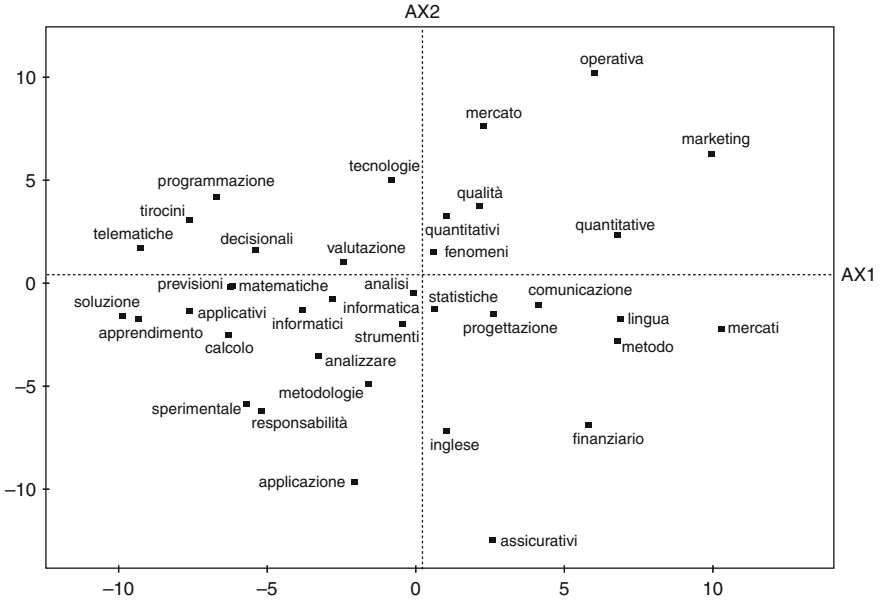


Fig. 13.1 The first factorial plane of the competences for Class 37 – Statistics

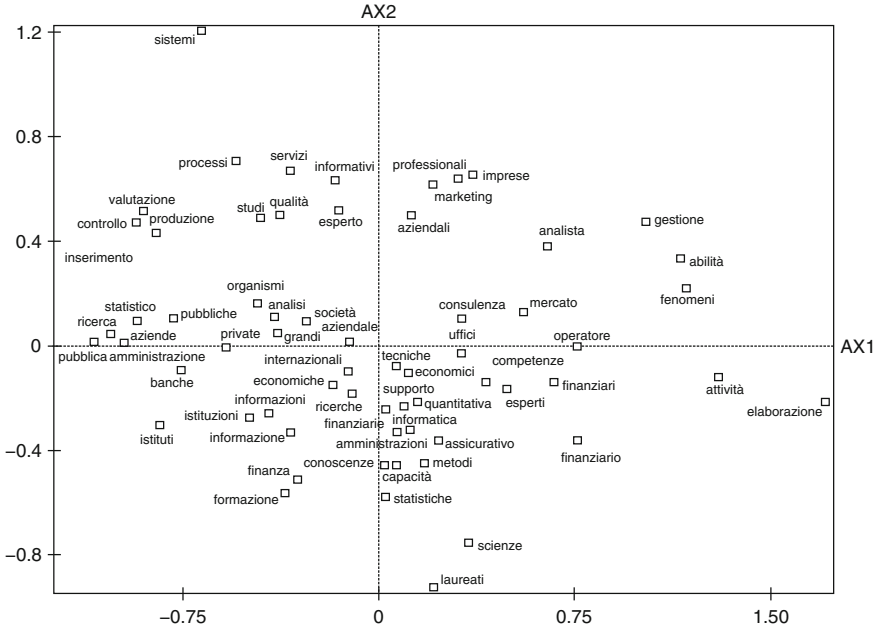


Fig. 13.2 The first factorial plane of the employment perspectives for Class 37 – Statistics

at the top. All these words are positioned on the right-hand-side of the diagram, while on the left there are words which are somehow more connected with the education of a statistician (*applicativi*-applicative, *previsioni*-forecasts, *decisionali*-decisional, *sperimentale*-experimental) next to generic terms (apprendimento-*learning*, soluzione-*solution*). It seems appropriate to point out how the root *statistic* – is very close to the origin of the axes, suggesting that it is being used generally in all degree courses.

In Fig. 13.2, describing the potential jobs for their graduates, it seems as if the statisticians are frightened of using the word which identifies their discipline and prefer turns of phrases instead (*quantitative*, *method*, *methodology*, *analysis*, *analyze*). There is a lot of mention of the disciplines ancillary to the study of statistics (*informatics*, *information technology*, *mathematics*, *calculation*), still on the left-hand-side of the graphic. In any case, the first axis opposes employees vs. consultants, but with poor relations with the opposition seen in Fig. 13.1. In conclusion, there doesn't seem to be a direct consequential link between the education offer and the employment offer, except for some specific professional roles (e.g. finance and insurance).

13.5 Coherence Between the Three Year Degree Courses and the Specialised Degree Programs

The objective to study the linguistic coherence between 3-year degrees and specialised 2-year degrees was pursued by adopting a method developed for the analysis of multilinguistic corpora, for comparing the latent semantic structures in translations, thanks to Procrustes rotations [3]. Perhaps this is a problem which will be less pressing with the new reform, as the structural connection between the 3 year degree course and the specialised degree appears to be loosening up, nevertheless considerations about the past could be profitable for the immediate future, for redefining of the degree courses.

The geometrical component of the method used is important. The documents are represented as points in a multi-dimensional system, spanned by the terms which make up the vocabulary of the analysed *corpus* (*vector space model*), or by their factorial transformation (*latent semantic indexing*). A measure of distance between two corpora can be obtained through Procustes rotations, so called because they best adapt the scatterings which represent the documents of the two corpora, through centering, standardization, reflection and rotation of the factorial axes.

Briefly, it is necessary to reduce the number of terms used in the two linguistic *corpora* being compared, identifying the individual latent semantic structures (which can be assimilated to the factors of a principal components analysis) and then to compare the two semantic structures, measuring their distance, in terms of fitting of one to the other.

In addition to the quantitative result, represented by the goodness-of-fit index, it is also possible to represent graphically the scattering of the points which represent the declaratives of the individual degree courses.

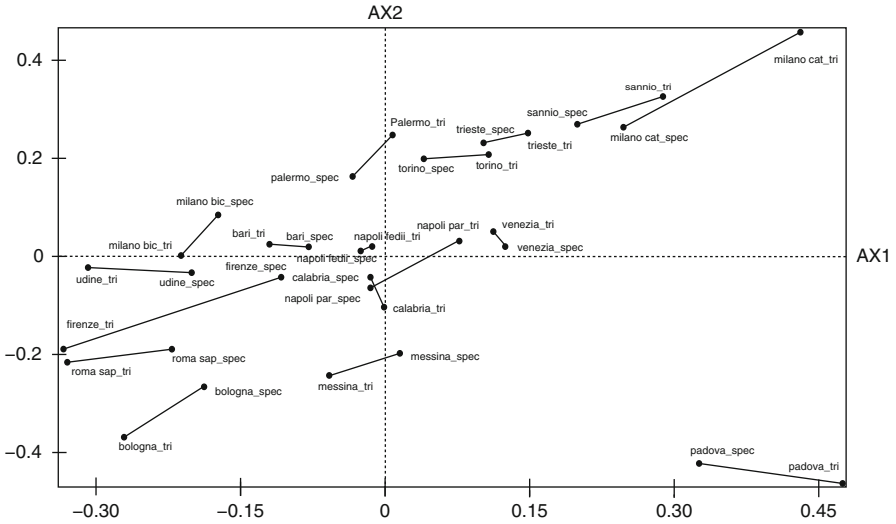


Fig. 13.3 The first factorial plane for Class 37 – Statistics and Classes 90S – Demographic and Social Statistics, 91S – Statistics for Experimental Research, 92S – Financial, Actuarial and Economic Statistics

For the class of degree under consideration (*Statistics*), thus two matrices were built, with the same number of rows, corresponding to the Universities where there are jointly activated degree courses in the 3 year degree and specialised degree courses in the connected classes and, in columns, the terms used in the course information data sheet about the education offer; in the first matrix about the 3 year degree and in the second one about the specialised degree programs. To the 3 year degree in statistics, many specialist degree courses can be connected, and only those where the word "statistics" appeared in the name were considered (Class 90S: *Demographic and Social Statistics*; Class 91S: *Statistics for Experimental Research*; Class 92S: *Financial, Actuarial and Economic Statistics*), involving a total of 17 Universities.

Once the latent structures were identified, the strength of the link for class of degree was measured, and the positions within the individual universities were shown (Fig. 13.3). In the figure, the final "tri"("spec") that follows the University name, indicates the 3 year (specialised) degree courses. The measure of fit is very low and amounts to 5.3 (the index is linked to the dimensions of the matrices analysed). The main variances are in the universities of Bologna, Milano Cattolica and Florence.

13.6 Final Remarks

From the whole of the analyses carried out, without simplifying the variety of the results obtained, some reflections have to be highlighted.

In predisposing the documentation, the universities show a substantial formal correspondence to the indications issued by the MIUR, even if sometimes we can note an excessive superficiality and, often, an excessive variability which is not always helpful in the choice. Furthermore, the areas where the model 3+2 was perceived more as extraneous show a greater difficulty in proposing more specific professional competences. In connection to the areas of employment expected, there doesn't seem to be a close consequence between the education offer and the employment possibilities indicated, except for some specific professional role. The analysis carried out has shown the necessity to dedicated more attention in the description of the professional roles at the end of the courses, trying to compare the competences offered to the students and what is effectively required in the job market, once they graduate. Finally, the connection between the 3 year degree courses and the specialist degree courses connected seems, in the cases examined, quite close; this positive judgement could, vice versa, turn itself into a contraindication in the light of the new reform, where this connection is, for some aspects, interrupted.

References

1. Balbi S (1995) Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms. In: Bolasco S et al (eds) Actes des 3es JADT, Journées internationales d'Analyse statistique des Données Textuelles. CISU, Roma
2. Balbi S, Di Meglio E (2004) A text mining strategy based on local contexts of words. In: Purnelle G et al (eds) JADT04 Le poids des mots. Press Universitaire de Louvain, Louvain
3. Balbi S, Misuraca M (2005) Visualization Techniques in non symmetrical relationships. In: Sirmakessis S (ed) Knowledge mining (studies in fuzziness and soft Computing). Springer, Berlin
4. Grassia MG, Misuraca M, Scepi G (2004) Relazioni non Simmetriche tra Corpora. In: Purnelle G et al (eds) JADT04 Le poids des mots. Press Universitaire de Louvain, Louvain
5. Lauro NC, D'Ambra L (1984) L'analyse non symetrique des correspondances. In: Diday E et al (eds) Data analysis and informatics, vol III. North-Holland, Amsterdam
6. Lebart L, Salem A, Berry L (1998) Exploring textual data. Kluwer Academic Publishers, Dordrecht
7. Misuraca M (2005) La visualizzazione dell'informazione testuale. Contributi metodologici e applicativi, Tesi di Dottorato in Statistica, Dipartimento di Matematica e Statistica, Università di Napoli "Federico II"
8. Morrone A (1993) Alcuni criteri di valutazione della significatività dei segmenti ripetuti. In: Anastex SJ (ed) JADT93, Actes des secondes Journées Internationales d'Analyse Statistique de Données Textuelles, ENSTelecom, Paris

Chapter 14
After the PhD: A Study of Career Paths, Job and Training Satisfaction Among PhD Graduates from an Italian University

Stefano Campostrini

14.1 Introduction

The aim of this chapter is to offer an initial presentation of the results from a survey conducted by the University of Pavia on its PhD graduates. Information was gathered regarding professional career paths chosen, job satisfaction and their doctorate experiences.

Doctorate programmes in Italy have not, unfortunately, been subject to extensive assessment and very little information is therefore available about their effectiveness, etc. This is in contrast with the international scenario where much attention has been paid over recent years to this subject (just scroll the 45,000+ hits that come up when one types "doctorate evaluation" into a web search engine, such as google scholar, to see just how much attention this topic has received).

This chapter focuses on two main points. The first point is the survey results – important for considering the role and the value of the doctorate, not only in Pavia University, but also more generally in Italy. It goes without saying that the data coming from a single university are of limited use due their specificity and low numbers, yet in the absence of other sources of information these results can nevertheless lead to the formation of first hypotheses regarding the general situation that exists for holders of Italian PhD degrees. The second point looks at the value of the survey itself and at the possible uses to which its results could be applied. Given these aims, further and more in depth analyses will be presented in successive papers, while the present communication discusses the descriptive data.

The survey was promoted by the *Nucleo di Valutazione* (Evaluation Committee) of the University of Pavia and it was intended to replace the formal role of an auditor who would normally assess the different opportunities/prospects that the University offers. Its overall aim was to collect a substantial volume of useful information for a more informed appraisal. With regards to the doctorate programmes in particular, the survey was designed to address a multitude of aims. It was important to gather information about the working experiences of Pavian PhD graduates, as well as to

S. Campostrini (✉)
Dipartimento di Statistica, Università Ca' Foscari Venezia, Venezia, Italy
e-mail: s.campostrini@unive.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_14,

draw out their "considered" opinions after a reasonable amount of time had elapsed since the conclusion of their doctorate programmes. Two other key objectives were (1) to gather information that would help the doctorate programmes and schools increase their offer levels, and (2) to fulfill the purpose of the evaluation committee, that is to perform "comprehensive evaluation". The motivation for the survey revolved around the belief that through the use of evaluation it is possible to enhance the quality of the teaching [10].

14.2 The Survey

Pavia is one of the oldest universities in the world, with more than seven centuries of history. In the Italian setting, it is a "medium-sized" university with around 20,000 students. It is a university that has played an important role in Italy, based on its geographical position (only 20 miles from Milan) and for the fact that it is globally acknowledged as being important in several academic disciplines. Thus, research has always been an important issue at Pavia University, and reflecting this, it has numerous doctorate programmes (almost 40) organized in five "doctorate schools".

Three groups of PhD degree holders were involved in the survey, defined as having discussed their PhD theses 1, 2 and 3 years before the date of the survey. It has been a sort of pilot study since the intention of the Evaluation Committee is to conduct an evaluation every year (indeed, a second survey has just been completed). So, if for the group who completed their PhDs just 1 year ago, the survey was a pilot for what would become a stable system, for the previous groups it was the solution for the need of information on the previous doctorate programmes, and a unique way of offering important data on what is happening 2 and 3 years after PhD graduation.

An obvious limit to such a retrospective survey is that any evolutionary effects are mixed with cohort effects. Given the introductory aims of this preliminary study and the limited numbers of the samples involved, we did not even try to separate them. Nevertheless, we believe that the cohort effect is limited, given the substantial stability in both doctorate offers and of the job market in the recent years.

The survey was conducted through CAWI (Computer Assisted Web Interviewing) that guaranteed the anonymity of the respondents and easy control of the data collection. Eligible respondents were contacted via e-mail, post and SMS in order to enhance participation. Unfortunately the poor quality of the starting list resulted in a low response rate (60%). The number of "not found" was much more than the number of refusals; the number of refusals were only estimated through qualitative information since the approach did not allow for a precise count and this should be taken into consideration when interpreting the results. In the 2008 survey (the data for which are not included in the present study) a better starting list and a compulsory recall strategy led to a 75% response rate. Analyses will pay particular attention to the "late answering" data in order to validate the results of the study presented here. We found the CAWI approach to be particularly suitable for the surveying of this type of population group.

14.3 How PhD Graduates Evaluate Doctorate Programme Teaching

A first, and in some way astonishing result, is concerned with how few doctorate programmes are organized through formal courses and lectures: only two thirds of the respondents reported to have attended their lectures (the percentage does at least increase across years, see Table 14.1) and the reason given for this was that there was no lecture course organized. Thus, even in one of the best of the Italian universities, it is evident that some departments believe that doctorate programmes do not present occasions for higher education, but instead only for "high qualification". They offer, perhaps, good opportunities for joining research programmes and, in this way to "learn from experience", but we believe some (organized) teaching is essential for any educational programme at any level.

Regarding the complexities of the teaching/training activities attended, PhD graduates were asked to judge the following features: quantity, quality, level of depth, teacher competence, teacher availability. Scores seem stable across groups but variable to a large degree between the respondents.

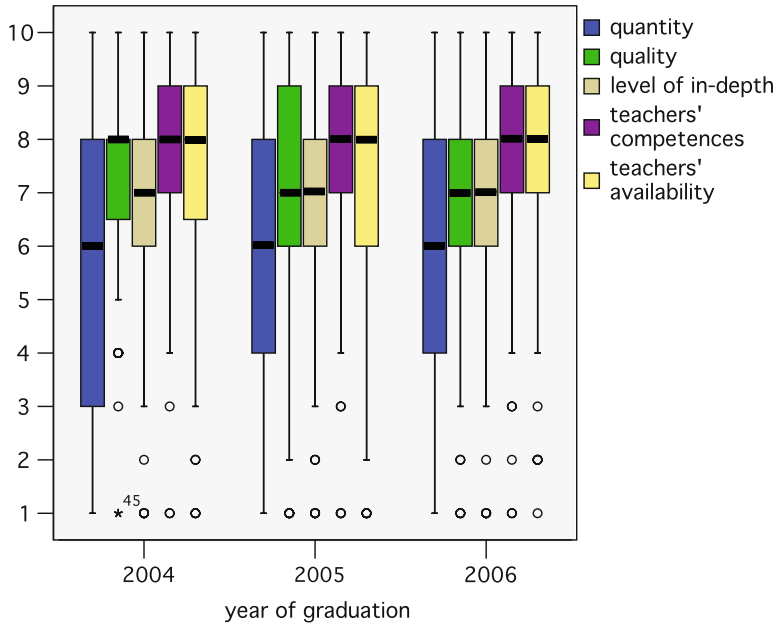


Fig. 14.1 Assessment (1–10 scores poor – high) of the teaching/training activities attended by PhD graduates by year of graduation ($n_{2004} = 79$; $n_{2005} = 91$; $n_{2006} = 90$). Features assessed: quantity; quality; level of depth (to which a subject was studied); teacher competence; teacher availability. *Boxplots* report the distribution of the answers: the *box* shows the values corresponding to the 25 and 75% of the distribution, the *bold line* in the box to the median, while the "whiskers" to the higher and lower observed value, excluding the outliers represented by *little circles or stars*

Table 14.1 Teaching/training activities attended by PhD graduates by year of graduation

	2004 ($n = 79$)			2005 ($n = 91$)			2006 ($n = 90$)		
	Lectures and courses (%)	Seminars and confer. (%)	Labs (%)	Lectures and courses (%)	Seminars and confer. (%)	Labs (%)	Lectures and courses (%)	Seminars and confer. (%)	Labs (%)
Attended	60.8	88.6	31.6	65.9	92.3	34.1	68.9	94.4	42.2
Did not attend	0.0	3.8	1.3	0.0	1.1	2.2	3.3	2.2	2.2
Was not offered	39.2	7.6	67.1	34.1	6.6	63.7	27.8	3.3	55.6

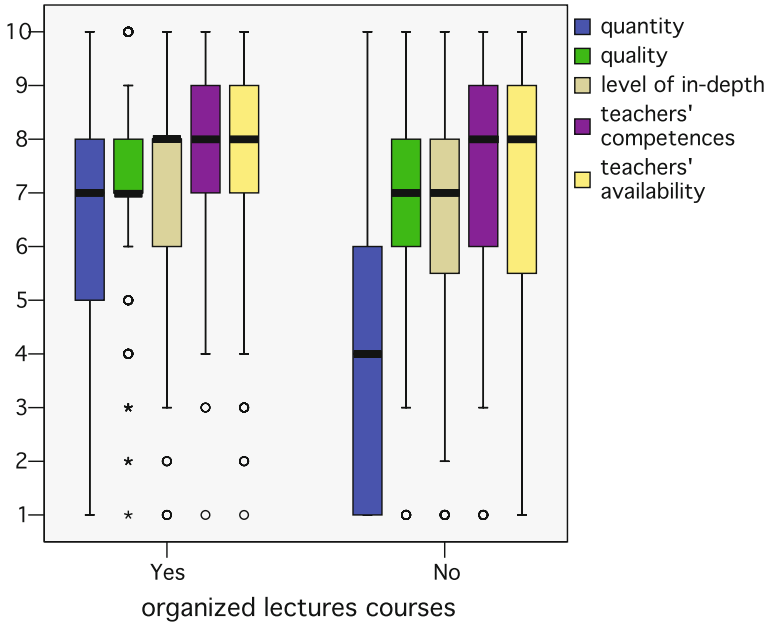


Fig. 14.2 Assessment of the teaching/training activities by the attendance of courses in the doctorate programme

Scores for all features were generally positive:

- the scores for teacher competence and availability were highest (over 75% scored more than 7 on a scale 1–10);
- quality and level of depth were generally high (50% scored more than 7);
- quantity was the feature that saw the most variability with only 50% of scores being over 6.

PhD graduates who declared that they had attended the organized lectures and courses, awarded quality, level of depth and in particular quantity with much higher scores compared to graduates who did not attend formal teaching, but only seminars and other direct contacts with teachers (Fig. 14.2).

14.4 How PhD Graduates Evaluate Their Doctorate Programme Research Experiences

A relatively high level of variability exists between average scores for the different features that characterize doctorate programme research activities: the median score across all features is consistently 8 for each of the three groups (see Fig. 14.3).

Higher levels of variability are observed for teacher availability, that probably reflect differences between the different programmes.

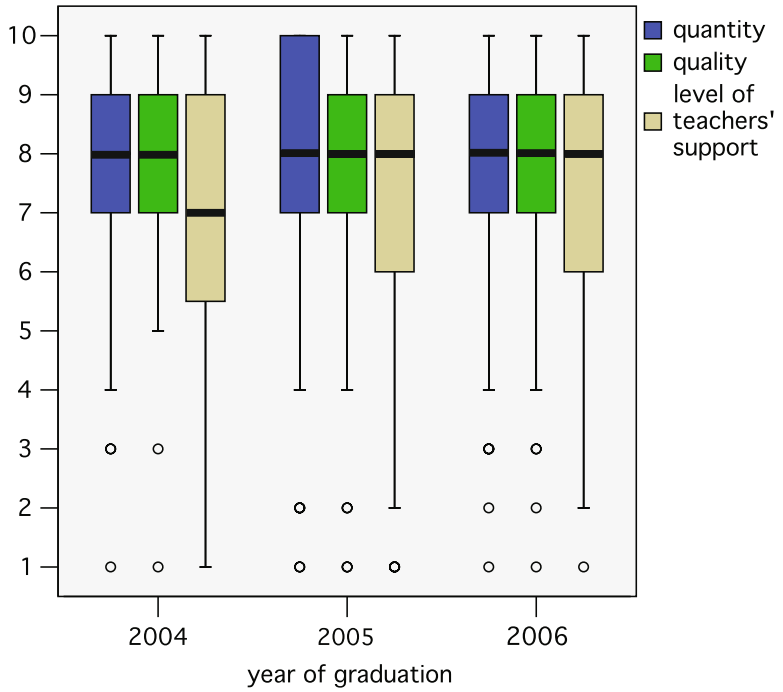


Fig. 14.3 Assessment (1–10 equals poor – high) of the PhD programme research activities by year of graduation ($n_{2004} = 79$; $n_{2005} = 91$; $n_{2006} = 90$). Features scored: quantity, quality, level of teacher support

Possible factors that could explain these differences (data not shown here) are whether or not the students had the opportunity to benefit from working within a research group within Pavia university, and whether they had the opportunity to experience working abroad as part of their doctorate programme. Again, the more organized that doctorate programmes are in terms of research experience, the more they were appreciated by the graduates.

14.5 How PhD Graduates Evaluate Their Doctorate Programme Research Experiences

One of the key areas of interests for this research lay in the potential to gain an understanding about the employment and work experiences of PhD graduates. Confirming that we are studying an "exceptional" population of students, 50% of the respondents stated that they continued to be employed by the businesses with whom they started during the doctorate programme. Among the others, many (30% of the total respondents) were anyway able to start new jobs after a relatively short delay (on average 4 months).

Two thirds of the respondents declared that they were currently work in a job that relates to their qualification. Only a minority (14%, see Table 14.3) are doing something unrelated to their research. Taking into consideration the possible "no answer" effect, this initial result indicates that the large majority of PhD graduates work within their research fields, and that extremely few are unemployed (see Table 14.2). The unemployment rate is just 3% for 1 year, 2% after 2 years and 1% after 3 years. So the problem for PhD graduates (and students) does not seem to be "if" they will continue to do research, but more (as we will soon see) "how".

Fifty percent of the respondents work in a university, but, comfortingly enough, in second place are private firms (Table 14.4). "Precariousness" seems to characterize university jobs, but this is well-known and it often depends on external factors for which, unfortunately, 3 years of observation are not enough to judge the stability of jobs within universities. The fact that it is increasingly difficulty to get a stable

Table 14.2 Type of job and time to get a job after PhD graduation, by year of graduation

	2004	2005	2006	Total
	($n = 79$)(%)	($n = 91$)(%)	($n = 90$)(%)	($n = 260$)(%)
I still do not have a paid job	1.3	2.2	3.3	2.3
I am back to the job I had before the doctorate (and that I had interrupted)	3.8	7.7	6.7	6.2
I continued to do the job I had before the doctorate (and that I had not interrupted)	15.2	12.1	16.7	14.6
I continued to do the job I got during the doctorate	48.1	51.6	41.1	46.9
I started a new job after[a] months	31.6	26.4	32.2	30.0
	100.0	100.0	100.0	100.0

[a]Mean = 4.15; standard deviation = 4.11.

Table 14.3 Consistency between job characteristics and PhD programme, by year of graduation

	2004	2005	2006	Total
	($n = 78$)(%)	($n = 91$)(%)	($n = 90$)(%)	($n = 259$)(%)
I still do not have a paid job	1.3	2.2	3.3	2.3
I have a research job relating to the doctorate programme	65.4	67.0	57.8	63.3
I have research job that does not relate to the doctorate programme	5.1	2.2	10.0	5.8
I have a job out of research, but that is still related to the doctorate programme	15.4	14.3	14.4	14.7
I have a job out of research that does not relate to the doctorate programme	12.8	14.3	14.4	13.9
	100.0	100.0	100.0	100.0

Table 14.4 Work place, by year of graduation

	2004	2005	2006	Total
	($n = 79$)(%)	($n = 91$)(%)	($n = 90$)(%)	($n = 260$)(%)
Not working	3.8	4.4	4.4	4.2
University	51.9	61.5	53.3	55.8
Other public research institutes	7.6	2.2	3.3	4.2
Other private research institutes	5.1	1.1	3.3	3.1
Public companies/organizations	2.5	3.3	5.6	3.8
Private companies	19.0	12.1	18.9	16.5
Self employed	3.8	3.3	2.2	3.1
Other	6.3	12.1	8.9	9.2
	100.0	100.0	100.0	100.0

position within an academic institution is an almost global problem [13], and even though the Italian situation is very complicated we will see that instability is not even compensated for by good pay. On the other hand, most of those working in the private sector have a stable position (see Table 14.5).

From the analysis of the work characteristics, two main issues emerge: job insecurity and low salaries. Other studies have shown that Italian graduates earn less than their colleagues in most other European countries (see Table 14.6). Nonetheless, considering the average salary of Pavian PhD holders, they are also low in comparison with average Italian salaries (even less than those declared by bache-

Table 14.5 Job conditions for respondents working in universities and private firms, by year of graduation

	2004	2005	2006	Total
University	($n = 41$)(%)	($n = 56$)(%)	($n = 48$)(%)	($n = 143$)(%)
Professor (full and associate)	4.9	1.8	0.0	2.1
Researcher	26.8	16.1	16.7	19.3
Grant (“assegno”)	41.5	48.2	56.3	49.0
Grant (“borsa”)	0.0	8.9	6.3	5.5
Short term contract	4.9	10.7	6.3	7.6
“Occasional” contract	7.3	5.4	2.1	4.8
Permanent position (not as researcher)	2.4	0.0	2.1	1.4
Temporary position	9.8	7.1	6.3	7.6
Other	2.4	1.8	4.2	2.8
	100.0	100.0	100.0	100.0
Private firms	($n = 15$)(%)	($n = 11$)(%)	($n = 17$)(%)	($n = 43$)(%)
Short term contract	26.7	9.1	5.9	14.0
“Occasional” contract	0.0	0.0	5.9	2.3
Permanent position (not as researcher)	73.3	27.3	52.9	53.5
Temporary position	0.0	18.2	23.5	14.0
Self employed (professional)	0.0	36.4	11.8	14.0
Other	0.0	9.1	0.0	2.3
	100.0	100.0	100.0	100.0

Table 14.6 Net monthly salaries, by year of graduation (in euros)

	2004	2005	2006	Total
	($n = 75$)	($n = 83$)	($n = 86$)	($n = 244$)
N. missing data (refusal)	4	8	4	16
Mean	1, 544.49	1, 494.35	1, 406.05	1, 481.72
Standard deviation	798.22	737.90	659.43	730.47
Min	192.00	300.00	100.00	100.00
Max	4, 000.00	5, 000.00	4, 000.00	5, 000.00
1st quartile	1, 200.00	1, 200.00	1, 100.00	1, 200.00
Median	1, 400.00	1, 250.00	1, 215.00	1, 250.00
3rd quartile	1, 900.00	1, 500.00	1, 500.00	1, 600.00

lor degree holders for 1–3 years after graduation; see [2]). It is worrying that even though the median value of salaries does increase 3 years post-PhD, the first quartile salaries remain stable with time; i.e. of all respondents on low salaries, 25% do not see any substantial increases with time. Those working outside universities declare to be earning higher salaries (median values are over 1,500 euros net per month, compared to less than 1,200, considering all the three groups).

Instability and salary do seem to be the only negative aspects mentioned by the majority of respondents when questioned about job satisfaction (see Fig. 14.4).

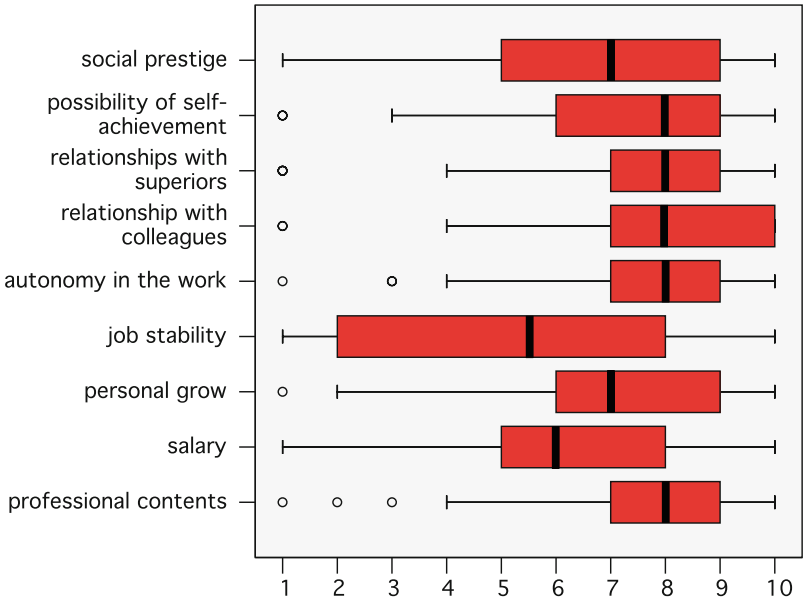


Fig. 14.4 Levels of job satisfaction expressed by PhD degree holders graduated in 2006 ($n = 91$). Features considered: social prestige, possibility for self-fulfilment, relationships with superiors, relationships with colleagues, autonomy at work, job stability, personal growth, salary and professional contents

Table 14.7 PhD qualification appreciation/acknowledgement at work (where, 1 = none, 10 = a lot), by year of graduation and place of work (at university vs. outside of university)

		Year			Structure	
1 = none – 10 = a lot	Total ($n = 249$)	2004 ($n = 76$)	2005 ($n = 87$)	2006 ($n = 86$)	University ($n = 145$)	Other ($n = 80$)
N. missing data (refusal)	11	3	4	4	0	0
Mean	6.8	7.0	6.8	6.7	7.6	5.9
Standard deviation	2.7	2.5	2.8	2.9	2.3	3.0
Min	1	1	1	1	1	1
Max	10	10	10	10	10	10
1st quartile	6	6	6	4.75	6.5	3
Median	7	7.5	7	7	8	6
3rd quartile	9	9	9	9	9	8

Most of the other components of job satisfaction present median scores of about 8 on the scale of 1–10, and this is certainly a very good result, particularly in comparison with the opinions expressed in other surveys by young workers of the same age (see, for example, [2]).

To complete the analysis on job satisfaction, we examined the answers given to the question "do you think that the value of the education and training received during your doctorate degree is recognized in your job?". On a scale that went from 1 "not at all" to 10 "a lot", average responses only reached the "sufficiency" mark (typically identified as "over 6" in Italy; from 6.7 for the 2004 graduates to 7 for those of 2004). The difference between academic and non-academic workers is quite substantial (see Table 14.7). Certainly, the difficulty that PhD graduates face in finding a job that relates to their qualification is not new and it is, in fact, a global problem. Nevertheless, in other European countries the high qualification acquired with a PhD seems to be more highly considered, also by private firms. (see [12, 14]).

14.6 PhD Holder Levels of General Job Satisfaction

In a similar format as that used for customer satisfaction surveys, we asked PhD holders to score the importance given to the possible outcomes of their doctorate experiences and their level of satisfaction perceived for each of these. Features considered included the following: life experience, increased chances of finding interesting professions, acquired competences, the provision of necessary. The results reported in Fig. 14.5 present some interesting findings: importance levels score high in every feature (with some variability) while satisfaction is high on life experience, moderate but positive on acquired competences and the provision of theoretical/basic profession training, but the possibilities for a better profession was rated as medium-low on average, although with a large response variability.

Considering satisfaction levels (assessed using the usual customer satisfaction questions, such as "if you could go back in time, would you repeat this experience?"), a relatively variable situation exists among respondents (see Table 14.8).

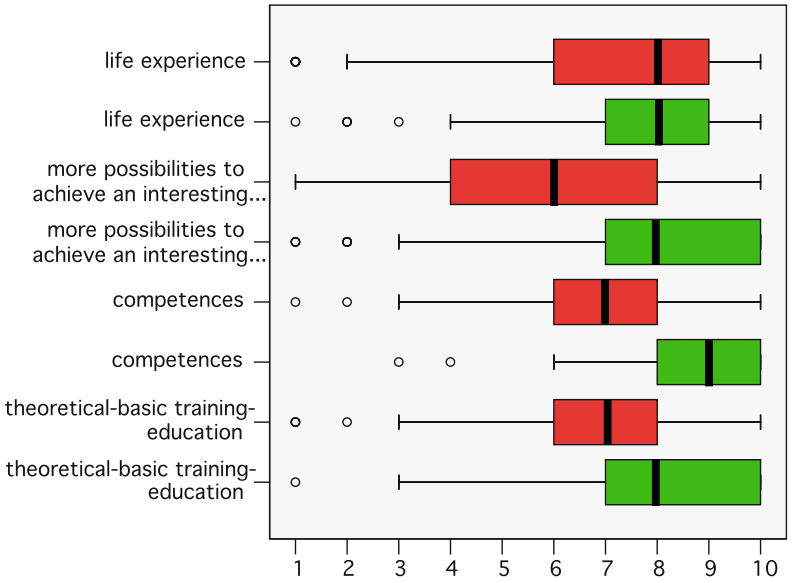


Fig. 14.5 Importance (*upper box-plot*) and satisfaction levels (*lower box-plot*) given by the PhD graduates from 2006 with regard to life experience, increased chances of finding interesting professions, acquired competences, theoretical and basic training preparatory to specific professions ($n = 91$)

The percentage of those that would repeat the same experience in the same doctorate programme was less than 50%; a bad result when compared to graduate responses from lower level degrees (e.g. bachelor or masters) at the same university and year that ranged from 75 to 85% (see [9]). The percentage of graduates that would prefer to attend programmes in countries abroad was very high (over 25%); an unsurprising result considering that respondents may have personal experiences of how a doctorate degree is considered overseas.

Table 14.8 Satisfaction levels, by year of graduation. *If you could go back in time, would you repeat this experience?*

	2004	2005	2006	Total
	($n = 79$)(%)	($n = 91$)(%)	($n = 90$)(%)	($n = 260$)(%)
Yes, in the same programme and at the same university	46.8	48.4	45.6	46.9
Yes, in the same programme but at a different university	7.6	11.0	10.0	9.6
Yes, but in a different university	11.4	2.2	2.2	5.0
Yes, but abroad	21.5	26.4	30.0	26.2
No, I would not attend a doctorate programme at all	12.7	12.1	12.2	12.3
	100.0	100.0	100.0	100.0

Although the aim of this chapter was simply to initiate a discussion based on a preliminary analysis of the data, it is worth commenting briefly on the results of the first multivariate analyses conducted on satisfaction level variables. The goal of these analyses was to identify the factors that have most influence, given all the possible interactions, upon the final scores for the doctorate programme/experience. After trying a very explorative approach, including CHAID trees (examples are not reported here) that consider almost all the variables available, usual log-linear models were applied to the satisfaction level data variables using, as explicatory variables, the data that resulted as being more interesting from the CHAID analyses. The following table (see Table 14.9) presents the results from the model that considered the following question as a dependent variable "how did you judge your doctorate experience in comparison with your expectations?" (opportunely dichotomised) and all the variables suggested as "interesting" in the first CHAID analysis (dichotomised following the results of the first step of analysis) that, at the same time, resulted as significant by the model.

It is interesting to note how in explaining satisfaction, the first important variable is research experience satisfaction – this presents strong evidence supporting genuine reasons for why doctorate students enter research. This confirms the result from the descriptive analysis that the second most important factor is satisfaction regarding teaching/training activities; once again, confirming that more organized doctorate programmes are better appreciated by PhD students.

Unexpectedly, working within a university structure is less important than the above mentioned features (and all those considered) and is not statistically significant, once all others are considered.

Table 14.9 Relative odds ratios, estimated using a log-linear analysis between declared satisfaction levels and variables that were more explicative among those collected in the survey on PhD graduates

Variable "satisfaction in comparison with expectations"
1 = *above expectations + as expected* ($n = 170$; 64.4%)
0 = *below expectations* ($n = 90$; 35.6%)

Number of observations = 260
χ^2 (4 *df*) = 86.22 (p-value = 0.000)
Log-likelihood = −125.53
Pseudo $R^2 = 0.2515$

Risk factors	Odds ratio	p-value	IC 95% lower	IC 95% upper
Satisf. of research activity ≥ 6/ Satisf. of research activity < 6 (scale 1–10)	8.61	0.000	3.76	19.70
Satisf. of teaching activity ≥ 6/ Satisf. on teaching activity < 6 (scale 1–10)	4.30	0.000	2.16	8.55
Doctor. experience useful for personal development ≥ 6/ useful for development < 6 (scale 1–10)	3.98	0.005	1.50	10.54
Does not work in a university/ works in a university	1.42	0.284	0.75	2.69

14.7 The Impact of Evaluation: A Reflection on the Use of the Evaluation

The world of the Italian doctorate remains almost unexplored,[1] despite its importance and all the efforts that universities make and are sustaining, [3, 7]. In Italy there is actually a National Committee for Evaluation of the University System that every year releases a report on all the Italian doctorate programmes, but rarely has addressed evaluation issues, being its focus on the resource allocation. Only recently it has been sponsored a study (still pilot and not systematic) addressed to study PhD holders' job trajectories (see [5, 6]). On these subject also the Consortium of University Presidents [11] produced a document that, to the best of our knowledge, has not produced any effect or any further development. With our work we hope we have given a first contribution to a too limited debate with a necessarily brief intervention on the results from a knowledge-evaluative survey conducted on a single Italian university, hoping that soon in the next future there will be the possibility to compare several ones.

As a final remark, we would like to linger over another important aspect, fundamental in our opinion, for the success of these evaluative processes. Indeed, various other authors (for example, see [1, 4, 15]) have also emphasized how important the "use" of the evaluation results is for achieving overall success of an evaluation study. Data, comments, results and, in particular, the affirmation of difficulties in specific modifiable features could have lost all importance if they had been shared only among "specialists" with vested interests.

So, above any academic publication, the most important presentation of this study is that made by the Evaluation Committee of Pavia University. Being both the sponsor and the most important "customer" of this study, this committee has shared and, importantly, disseminated the results, via both formal presentations[2] and discussing them with the various stakeholders.

If something were to change following this study, that in itself would be the best result and the ideal measure of real success. This is in the hope that the doctorate, like an "ugly duckling" [8], will be able to take off and fly, as will also hopefully happen to the rest of the research sector that is so deeply caught up in the difficulties of our country today.

Acknowledgement I would like to thank Stefano Govoni, all the other members and the staff of the Nucleo di Valutazione of the Pavia University, not only for having accepted this evaluation activity so enthusiastically, but also for having believed in it, for the support given, and for the use of the results that we have promoted. I also acknowledge the help of Francesca Pozza for the analyses and the work of Nicoletta Parise who, although busy in other fields, organized and supervised the survey and performed the checks and the first analyses on the data with the highest level of professionalism. All responsibility for content and conclusions stated within this chapter and for all possible mistakes are attributable to the author only.

[1] At the time this chapter is going to press, few national initiatives to study the effectiveness of doctorate programs have been realized (e.g. by the University consortium STELLA) or planned (e.g. by the National Institute for Statistics – ISTAT).

[2] By including them in its standard reports and awarding them maximum publicity on the web site: http://nuv.apnetwork.it/attach/file/presentazionedottori.pdf.

References

1. Alkin MC, Daillak R, White P (1979) Using evaluation: does evaluation make a difference? Sage, Beverly Hills, CA
2. AlmaLaurea (2007) IX Indagine AlmaLaurea sulla condizione occupazionale dei laureati. il Mulino, Bologna
3. Avveduto S, Cipollone PE (1998) La mobilità delle intelligenze in Europa - Internazionalizzazione della formazione e dottorato di ricerca. Franco Angeli, Milano
4. Balthasar A, Rieder S (2000) Learning from evaluation. Evaluation 6(3):245–260
5. Bianchetti M, Della Ratta F, Lanzoni M, Pischedda V, Rizzo R, Usai MC (2003) Progetto CNVSU per la ricognizione, raccolta e analisi dei dati esistenti sul dottorato di ricerca e per l'indagine sull'inserimento professionale dei dottori di ricerca. Technical report, ADI – Associazione Dottorandi e Dottori di Ricerca Italiani
6. Bianchetti M, Della Ratta F, Lanzoni M, Pischedda V, Rizzo R, Usai MC (2006) Progetto CNVSU per la ricognizione, raccolta e analisi dei dati esistenti sul dottorato di ricerca e per l'indagine sull'inserimento professionale dei dottori di ricerca. Technical report, ADI – Associazione Dottorandi e Dottori di Ricerca Italiani
7. Brandi MC, Avveduto S (2000) Risorse umane: quale futuro nella scienza? Franco Angeli, Milano
8. Cesaratto S, Avveduto S, Brandi MC (1994) Il brutto anatroccolo. Il dottorato di ricerca in Italia fra università, ricerca e mercato del lavoro. Franco Angeli, Milano
9. CILEA (2006) Laureati STELLA, indagine occupazionale post-laurea laureate 2004 – indagine 2006. Technical report, CILEA, Segrate
10. Cronbach L (1963) Course improvement through evaluation. Teach Coll Rec 64:672–683
11. CRUI (2002) Proposta di un sistema di valutazione dei corsi di dottorato di ricerca. Technical report, Conferenza dei Rettori delle Università Italiane, Roma
12. Enders J (2002) Serving many masters: the PhD on the labour market. Higher Educ 44:493–517
13. Huisman J, De Veert E, Bartelse J (2002) Academic careers from a European perspective. J Higher Educ 73(1):141–160
14. Mangematin V (2000) PhD job market: professional trajectories and incentives during the PhD. Res Policy 29:741–756
15. Patton MQ (1997) Utilization focused evaluation. The new century text. Sage, Thousand Oaks, CA

Chapter 15
Secondary School Choices in Italy: Ability or Social Background?

Dalit Contini and Andrea Scagni

15.1 Introduction

It is often held that educational expansion narrows social inequalities within nations by promoting a meritocratic basis for status attainment, yet substantial research indicates that the relative advantages of elite children over children with less privileged background have changed little in the last decades [4, 12, 19]; on average higher status children perform better in school and attain higher educational levels. In this light, inequality of opportunity (IEO) in education is still a highly relevant issue in the international educational policy agenda.

Class differentials in educational attainment are related in the sociological literature to *primary* and *secondary* effects [2]. The former refer to the influence of social origin on ability early in children's educational careers: high status parents are more likely to sustain and motivate the school work and provide a stimulating environment to their offspring. The latter operate through the choices that families make within the educational system (including exit) *given* the level of ability. The rational action approach [3, 11], assuming that families wish to avoid intergenerational downward mobility, provides a theoretical explanation for the evidence that, at given levels of ability, school choices vary across social background. Ability is intended here as an observed measure of school performance (typically grade point average) as opposed to unobserved measures of cognitive abilities, since it is held that it is the former that affects the decision process through the perceived probability of schooling success.

The evaluation of primary and secondary effects is particularly relevant at the end of compulsory schooling, where in many European countries students face the decision whether to enrol into the academic track,[1] to enrol into a vocational track, or to enter the labour market.

D. Contini (✉)
Dipartimento di Statistica e Matematica Applicata "Diego De Castro", Università di Torino, Torino, Italy
e-mail: dalit.contini@unito.it

[1] The term *track* is often used in the literature to indicate the different secondary school educational paths available to students in a certain educational system. The *academic* track is the one

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_15,

IEO is obviously affected by the institutional features of the school system. Interventions aimed at containing primary effects should enhance the performance of children of less advantaged background, especially at the primary school level. Secondary effects can be reduced by endorsing the enrolment of lower status children into the academic track or, possibly, by regulating access through ability assessments.

The assessment of the relative importance of primary and secondary effects is the aim of a growing body of literature [9, 15, 16, 20]. This research – based on surveys carried out at a national level – provides empirical evidence of the relevance of secondary effects in the creation of class differentials in educational attainment.

More specifically, the following questions are addressed: what would the differential across social strata in the transition to upper secondary education be like if the observed ability distribution was held constant? What would it be like if the transition probability given ability was held constant? The methodology is quite simple. First, the school performance distribution at the end of compulsory schooling and the probability of enrolment into the academic track given performance are estimated for each social background with standard methods. Second, these estimates are combined according to what the proponents call a "counterfactual" reasoning. For each j and k, the probability of entering the academic track that individuals would face if they had the ability distribution of class j, but the transition probability given ability of class k, is evaluated. Observed and counterfactual odds-ratio are estimated, and finally the *log*(odds-ratio) are decomposed into two components, representing the relative importance of primary and secondary effects.

Aim of this chapter is to provide an assessment of primary and secondary effects in secondary school choices in Italy. Empirical work on other countries (UK, Sweden, Germany, Netherlands) relies on panel surveys recording data on schooling careers, but no adequate prospective longitudinal data is available for Italy. The analysis is based on the data of the survey *Percorsi di studio e di lavoro dei diplomati* [14], which collects detailed information of individual educational histories up to 3 years after the attainment of the secondary school degree. Given that only secondary school graduates are interviewed, a major issue to deal with is sample selection; we estimate the relevant distributions, correcting for selection bias, by integrating the survey data with administrative and census information.[2]

A semi-parametric version of the standard approach is adopted to account for the fact that lower secondary school final marks follow a coarse 4-level scale.[3] Our main

conceived to prepare for university studies (even if in some countries it is not required to enter tertiary education).

[2] Employing data from PISA (*Programme for International Student Assessment*; [18]) would weaken the sample selection problem, since students are interviewed at 15, i.e. near the beginning of upper secondary school. However this option proves impossible since PISA does not include information on students' performance before choice. PISA may however be appropriate to evaluate the *total effect* of social background (see for example [6]).

[3] We refer to the term "semi-parametric" to account for the fact that the ability distribution is estimated non-parametrically while the transition probability is estimated by logit regression.

finding is that in Italy secondary effects are more important than primary effects in driving social origin differentials and that the relative contribution of primary effects is substantially weaker than in the other countries.

The chapter is structured as follows. In Sect. 15.2 we illustrate the methodology for decomposing total inequality in primary and secondary effects as proposed in the recent literature. In Sect. 15.3 we briefly review the main fetaures of the Italian educational system and discuss the implications of measurement error on student's ability. The sample selection problem is addressed in Sect. 15.4. In Sect. 15.5 we describe the empirical analysis, while Sect. 15.6 is devoted to conclusions.

15.2 The Methodology

Let A be a continuous measure of students' school performance before track choice and S a discrete variable representing social origins. Then $f(A|S)$ is the distribution of the performance scores for each group; assuming a normal distribution, the relevant parameters can be estimated by group sample mean and variance.

Define Y as a binary variable taking value 1 if the academic track is chosen and 0 otherwise (i.e. if the student chooses a different track or if he does not enter secondary education). Note that Y refers to the first choice after the end of compulsory schooling and not to possible subsequent changes. The transition probability *given* performance $P(Y = 1|A, S)$ can be estimated with binary logistic regression for each class separately. The integrals:

$$P_{jj} = \int_{-\infty}^{+\infty} f\,(A|S = j)\; P\,(Y = 1|A, S = j)\, dA \tag{1}$$

evaluated for each S by numerical integration, represent the predicted probability $P(Y = 1|S = j)$, whose observed counterpart is the percentage enrolling into the academic track among those in social class j enrolling into the academic track.

On the other hand, the integral:

$$P_{jk} = \int_{-\infty}^{+\infty} f\,(A|S = j)\; P\,(Y = 1|A, S = k)\, dA \tag{2}$$

is a "counterfactual" probability. Expression (2) is the transition probability that an individual would experience if he had the performance distribution of social class j and the transition probability of class k. With K social classes, there are $K(K - 1)$ counterfactual probabilities.

Observed differentials with respect to the probability to enrol into the academic track between classes j and k can be measured by the odds ratio:

$$Q_{jj.kk} = \frac{P_{jj}/(1 - P_{jj})}{P_{kk}/(1 - P_{kk})} \tag{3}$$

Define also:

$$Q_{jj.kj} = \frac{P_{jj}/(1 - P_{jj})}{P_{kj}/(1 - P_{kj})}$$

The numerator represents the odds of continuing to academic education for an individual exposed to the performance distribution and the transition probability of class j, while the denominator represents the odds for an individual with performance distribution of class k and transition probability of class j. Since the difference lies here only in the performance distributions, this quantity is informative on primary effects. Similarly:

$$Q_{kj.kk} = \frac{P_{kj}/(1 - P_{kj})}{P_{kk}/(1 - P_{kk})}$$

provides information on secondary effects, as what varies here is the transition probability while the performance distribution remains fixed.

The total effect (3) can be factorized in two distinct ways:

$$Q_{jj.kk} = Q_{jj.kj} Q_{kj.kk}$$
$$Q_{jj.kk} = Q_{jk.kk} Q_{jj.jk}$$

By taking the logarithms, we obtain:

$$L_{jj.kk} = L_{jj.kj} + L_{kj.kk} \tag{4}$$
$$L_{jj.kk} = L_{jk.kk} + L_{jj.jk}$$

The relative importance of secondary effects can be evaluated by $L_{kj.kk}/L_{jj.kk}$ or $L_{jj.jk}/L_{jj.kk}$. Estimates based on the two expressions in (4) generally differ, although in practice not to a great extent [9].

To fix ideas, let us assume there are only two social levels: H (high) and L (low). Then:

$$L_{HH.LL} = L_{HH.LH} + L_{LH.LL}$$

where the (log) total effect is given by the primary effect evaluated by considering the transition probability of the *high* class and the secondary effect with the performance distribution of the *low* class. The alternative decomposition is:

$$L_{HH.LL} = L_{HL.LL} + L_{HH.HL}$$

where the first term is the primary effect evaluated with the transition probability of the *low* class and the second term is the secondary effect with the performance distribution of the *high* class.

It is worthwhile to note that under the linear probability model:

$P(Y = 1|A, S) = \mu + \lambda S + \theta A$, with $A = \alpha + \beta S + \varepsilon$, the following holds:

$$P(Y = 1|S = j + 1) - P(Y = 1|S = j) = \beta\theta + \lambda$$

In this case primary effects are represented by $\beta\theta$ and secondary effects by λ. Instead, it can be shown that under the logistic model:

$$\ln \frac{P(Y = 1|A, S)}{1 - P(Y = 1|A, S)} = \mu + \lambda S + \theta A$$

primary and secondary effects – as measured by (4) – are functions of the parameters of both the model for A and the model for Y, although the component related to primary effects is much more sensitive to β and θ, and the component related to secondary effects is much more sensitive to λ.

15.3 The Analysis for Italy

15.3.1 Institutional Features

At present compulsory education starts at age 6 and ends at age 15; however, for the cohort considered in this work the end was still set at 14. Primary school lasts 5 years, after which there are 3 years of comprehensive lower secondary education. Students may then choose upper secondary education among a variety of different programmes. A broad distinction can be made among the academic, technical and vocational tracks. The first, conceived to prepare for university, includes different general educational programs (lyceums). The vocational track leads directly to a professional qualification and typically lasts 3 to 5 years. Technical education combines general education with a more specific vocational training and is considered to be less demanding than the academic track. There are no ability related admission requirements to enter the different programs.

After 5 years of schooling, most of all programmes give access to university [10]. In practice, around 90% of the students exiting from lyceums in 2001, 43% from technical schools and 17% from vocational schools enrol in university within 3 years from graduation.[4] Hence, there are marked differences among tracks in terms of the chances to get a tertiary level degree. In order to allow for cross-country comparability, since all the studies in the literature focus on this divide, in the empirical analysis

[4] According to the survey *Percorsi di studio e di lavoro dei diplomati* 2004.

we will distinguish between the academic track (comprising *classical*, *scientific* or *linguistic* lyceums[5]) and all other educational programmes.

15.3.2 The Data

No extensive panel survey providing information on schooling careers and parental socio-economic status is available for Italy.[6] For this reason we use data from the cross-sectional survey *Percorsi di studio e di lavoro dei diplomati* (2004), carried out by ISTAT on 2001 higher secondary school graduates with the aim to investigate the transition from secondary schools to tertiary education or the labor market. Individuals are interviewed 3 years after graduation, and information is collected retrospectively.[7] Given the survey's features, the sample is self-selected with respect to our research question; this issue is addressed in Sect. 15.4.

15.3.3 Final Marks in Lower Secondary School

According to the rational choice theory [3], families make their educational choices with the aim to avoid downward mobility, according to future employment prospects and the probability of schooling success relative to each option. This probability is assessed by taking into account children's ability, conceived as an observed measure of school performance.

In Italy, the final lower secondary school mark[8] is the main observed information on children's ability before track choice. We highlight three possible sources of measurement error:

(i) In Italy final lower secondary grades follow a 4-level scale (*pass, good, very good, excellent*). This highly discrete grading system appears to be quite a rough measurement of students' ability when compared to other countries marks, generally based on finer scales.

---

[5] The *socio-pedagogic lyceum* (formerly called *istituto magistrale*) is conceived to prepare for primary school teaching. Although university education is now required, until a few years ago this type of school gave direct access to the teaching career; for this reason we do not consider this school type in the academic track. Given its specific focus, a similar argument also applies for the *artistic lyceum.*

[6] There are few surveys recording longitudinal data in Italy: the *Indagine sui Bilanci delle Famiglie* (Banca d'Italia) does provide some information on individual's educational careers, but no data on school performance is available; moreover, the sample size is too small to allow analyses on specific birth cohorts.

[7] Interviews were carried out with CATI. Data is collected with a two stage sampling scheme; 20,408 individuals in 1,868 schools were interviewed.

[8] The mark is attributed after a national exam, detached from normal school activity, at the end of lower secondary school (*Esami di Stato conclusivi del I ciclo*).

(ii) Exams are set up by the school teachers, and are not based on standardised national tests.[9] An indirect evidence of the existence of a bias is that, although international assessments such as PISA [18] show a significantly lower average level in Southern Italy with respect to the North, in the South the percentage of *excellent* is higher than in the rest of the country.
(iii) With respect to the latter point, if marks were related to the average within school ability, higher performing schools could evaluate their students somewhat more severely. The issue is particularly relevant in highly socially segregated schooling systems, since on average high status children perform better.

The problem of measurement error is not explicitly addressed here. The reasons are twofold. First, we think that the main source of bias is likely to be given by sample selection, due to employing data on secondary school graduates. Perhaps more importantly, the second reason has to do with the rationale of the analysis. If it is true that people make their educational choices on the basis of observed school performance,[10] the "correct" measure of ability for secondary effects is given by marks, even if they are affected by measurement error. On the other hand, the "correct" measure for identifying class differentials in the performance distribution should be latent ability.

Nevertheless, it is important to note that the decomposition method described above regards the role of manifest ability in *driving school choices.* In fact the transition probability functions for each social class (1) turn out to be a weighted average of the class transition probabilities given ability (marks), where the weights are given by the relative proportion of individuals with each level of ability (again, marks) within the group. In this light, it is not relevant whether school marks represent a measurement error version of true ability (measurement error is instead very relevant if the aim is to assess ability class differentials). Thus, when we come to interpret primary effects in this context, we should acknowledge that what is here called "primary effects" has to do with the distribution of latent ability *and* the way this ability is actually translated into marks.[11] Yet, this caveats would not hold if people were aware of their true level of ability and took their decisions accordingly: transitions rates would have to be assessed given true ability and weights defined consequently. By employing marks, in this case the relative contribution of primary effects would be underestimated.[12]

[9] This issue is likely to become less relevant in the future: from 2007, in fact, final exams include two standardized tests (linguistics and mathematics) with common evaluation guidelines.
[10] Stocké [20] addresses this issue for Germany and finds that educational choices are driven mainly by school marks, although a minor effect can be ascribed to parents' perception of their children ability.
[11] It is nevertheless obvious that in the extreme case where marks were hardly related to ability, the decomposition itself would loose much if its meaning, in that secondary effects would become the only source of class differentials.
[12] We developed a simulation study (not presented here) to address this issue. The bias appears to be small for measurement error of type (i) and (ii) and somewhat bigger for type (iii).

15.4 Sample Selection

15.4.1 The Problem

As we have pointed out, no adequate panel survey recording school careers is available for Italy; for this reason we employ the ISTAT cross-sectional survey on graduates, where the relevant information is recorded retrospectively. Our aim is to evaluate class differentials in secondary school *choices*; since the survey target population does not include those who have enrolled into a secondary school and exited the educational system before completion, the sample is affected by selection bias[13].

We now deal with the consequences of sample selection; we show that without corrections we would *underestimate* both the differences in the ability distribution across social background levels, and the effect of social background on school choices. Note that traditional methods to correct for sample selection such as the propensity score or Heckman method cannot be employed here because micro-data on dropouts is not available: we can only obtain some indirect information at the aggregate level.

15.4.1.1 Primary Effects

As before let A be the school performance before track choice and S a measure of families social status. Define G as a binary variable taking value 1 if the child has attained a secondary school degree and 0 if he has exited the educational system. The distribution of interest is $P(A|S)$, while the observable distribution is $\mathrm{P}(A/S, G=1)$. The two distributions are related by:

$$P(A|S,G=1) = P(A|S)\frac{P(G=1|A,S)}{P(G=1|S)} = P(A|S)\frac{P(G=1|A,S)}{\int_A P(G=1|S,A)P(A|S)\mathrm{dA}}$$

The observable distribution and the distribution of interest coincide if the second factor in the right hand side is equal to 1, i.e. if performance A does not affect the graduation probability given social status. Since this is obviously very unlikely, the survey estimate of the performance distribution given social status is biased.

Final marks are coded as: *pass* (1), *good* (2), *very good* (3), *excellent* (4). We will make the assumption that school drop-outs come exclusively from the population of low performers (see next section for empirical evidence on this):

$$P(G=0|S,A=j) = \begin{cases} >0 & \text{if } j=1 \\ =0 & \text{if } j=2,3,4 \end{cases} \tag{5}$$

[13] Children who have chosen a vocational program and attained a *qualifica professionale* (after 3 years) but not a *diploma* (after 5 years) are also excluded from the survey. To simplify the exposition, the term "dropouts" includes them as well.

For $j = 2,..,4$ this implies that:

$$P(A = j|S, G = 0) = \frac{P(G = 0|A = j, S)P(A = j|S)}{P(G = 0|S)} = 0 \tag{6}$$

Since:

$$P(A|S) = P(A|S, G = 1)P(G = 1|S) + P(A|S, G = 0)P(G = 0|S) \tag{7}$$

by combining (6) and (7) we obtain:

$$P(A = j|S) = \begin{cases} P\,(A = j|S, G = 1)\,P\,(G = 1|S) & if\ j = 2, 3, 4 \\ 1 - \sum_{j=2}^{4} P\,(A = j|S) & if\ j = 1 \end{cases} \tag{8}$$

In order to estimate $P(A|S)$, we employ the ISTAT graduates' survey to evaluate $P(A|S, G = 1)$, but we also need to estimate the graduation probability given social status $P(G = 1|\mathrm{S})$. Since:

$$P(G = 1|S) = \frac{P(S|G = 1)P(G = 1)}{P(S)}$$

we will estimate $P(S|G = 1)$ from the graduates survey and exploit the official statistics derived from administrative data sources for the overall graduation probability $P(G)$ and the social status distribution $P(S)$ (see Sect. 15.5.2).

15.4.1.2 Secondary Effects

Let Y represent again secondary school choice: $Y = 1$ for the academic track and 0 otherwise. We are interested in $P(Y = 1|A, S)$, but we can only estimate $P(Y = 1|A, S, G = 1)$. Since:

$$P(Y = 1|A, S, G = 1) = P(Y = 1|A, S)\frac{P(G = 1|Y = 1, A, S)}{P(G = 1|A, S)} \tag{9}$$

the survey estimate is unbiased if, given ability and social status, the graduation probability does not depend on the chosen track. Note that Y refers to the *first* choice undertaken at the end of compulsory school, while graduation can be achieved in *any* track. Students may change track if they fail or if they are not satisfied with their initial choice, and then graduate. In this light, the likelihood of attaining a secondary school degree does not depend on how difficult or selective a specific track is. The enrolment into the academic track could be considered instead as a signal of higher *aspirations*.

The consequences of employing directly the graduates' survey to estimate $P(Y = 1|A, S)$ can be easily grasped by assuming the simple linear probability models:

$$P(G = 1|A, S, Y) = \alpha + \beta A + \gamma S + \delta Y$$
$$P(Y = 1|A, S) = \lambda + \xi A + \theta S$$

Then $P(G = 1|A, S)$ can be written as:

$$\begin{aligned} &P(G = 1|A, S, Y = 1)P(Y = 1|A, S) + P(G = 1|A, S, Y = 0)P(Y = 0|A, S) \\ &\quad = (\alpha + \beta A + \gamma S + \delta)P(Y = 1|A, S) + (\alpha + \beta A + \gamma S)P(Y = 0|A, S) \\ &\quad = \alpha + \beta A + \gamma S + \delta P(Y = 1|A, S) \end{aligned}$$

Then the second factor in the right hand side of (9) is:

$$\frac{P(G = 1|Y = 1, A, S)}{P(G = 1|A, S)} = \frac{(\alpha + \beta A + \gamma S) + \delta}{(\alpha + \beta A + \gamma S) + \delta(\lambda + \xi A + \theta S)}$$

This expression is never smaller than 1 (it is equal to 1 if $\delta = 0$), and is a decreasing function of A and S. In fact, $(\lambda + \xi A + \theta S) < 1$ (since it is a probability); given that parameters are positive, it is an increasing function of A and S. Thus, the observed probability is greater than the probability of interest for all status, but it is increased by a greater factor for the lower social background. As a consequence, secondary effects are *underestimated*. By employing a different data source, in the next section we will provide empirical evidence that δ is nearly 0, implying that aspirations are entirely captured by school performance and social status. Given this result, no corrections are needed here; $P(Y = 1|A, S)$ can be estimated directly from the graduate's survey data.

15.4.2 Supporting the Assumptions

By employing other data sources we now provide empirical evidence to support the assumptions made in the previous section.

15.4.2.1 Primary Effects

Let us recall the relevant assumption described by Eq. (5) that for each social background, only low performers eventually drop-out from secondary school. The marginal distribution of performance can be written as:

$$P(A = j) = P(A = j|G = 1)P(G = 1) + P(A = j|G = 0)P(G = 0)$$

from which we obtain the performance distribution for school-drop-outs:

$$P(A = j|G = 0) = \frac{P(A = j) - P(A = j|G = 1)P(G = 1)}{P(G = 0)} \tag{10}$$

This distribution can be roughly estimated by employing the graduates survey data – providing information on $P(A|G = 1)$ – and aggregate administrative data from ISTAT – which records the overall distribution of lower secondary final examination marks $P(A)$ for the year 1996, as well as an estimate of the overall national percentage of school dropouts $P(G = 0)$ for the same year. From (10) we obtain:

$$\hat{P}(A = 1|G = 0) = 0,96 \qquad \hat{P}(A = 2|G = 0) = 0,05$$
$$\hat{P}(A = 3|G = 0) = 0,005 \quad \hat{P}(A = 4|G = 0) = -0,02^{14}$$

strongly supporting the assumption.

15.4.2.2 Secondary Effects

We now evaluate the assumption:

$$\frac{P(G = 1|Y = 1, A, S)}{P(G = 1|A, S)} = 1 \tag{11}$$

As we have pointed out before, no longitudinal micro-data on schooling careers is available for the estimation of the conditional distribution of *G*. However, a survey carried out jointly by CISEM and IARD[15] in 2006 on 3,600 upper secondary school students in the area of Milan can be employed for this purpose. The sample includes students in each of the five grades of the upper secondary schools; information on schooling careers as well as family characteristics, including parental educational and occupational status are recorded. The survey is cross-sectional and does not include dropouts; nevertheless, by comparing 1st grade students (including all future dropouts) with 5th grade students (assuming that nobody will exit the school system thereafter), we can roughly assess the profile of those who do not attain a secondary school degree.

By a simple application of Bayes' theorem[16]:

$$\frac{P\,(G = 1|Y = 1, A, S)}{P\,(G = 1|A, S)} = \frac{P\,(Y = 1|G = 1, A, S)}{P\,(Y = 1|A, S)}$$

[14] Small inconsistencies among the combined data sources produce a negative probability, which is however so close to 0 to be reasonably considered negligible.

[15] CISEM stands for *Centro per l'Innovazione e Sperimentazione Educativa Milano* and is a research centre on educational problems of *Provincia di Milano*. IARD – Istituto Franco Brambilla is a research centre focusing on life problems and opportunities of young people. The authors would like to thank both CISEM and IARD for the collaboration and availability of data.

[16] Since $P(G = 1|Y = 1, A, S) = \frac{P(G=1,Y=1|A,S)}{P(Y=1|A,S)} = \frac{P(Y=1|G=1,A,S)P(G=1|A,S)}{P(Y=1|A,S)}$.

The right hand-side can be estimated by the ratio of the share of students enrolled in the academic track in 5th grade to the corresponding share in 1st grade. Considering S as the highest parental education and modelling both $P(Y = 1|G = 1, A, S)$ and $P(Y = 1|A, S)$ with binary logit regressions, these ratios are all close to 1, supporting the validity of (11).[17]

15.5 The Empirical Analysis

15.5.1 Semi-parametric Approach

In Sect. 15.2 school performance A was assumed to be a continuous variable (in most countries marks follow a fine scale and in some cases the grade point average is employed) which can be approximated quite well by a normal distribution; although not strictly necessary, this is also useful for the numerical evaluation of integral (1).

Because of the highly discrete scale, the normal distribution is clearly not appropriate for Italy. In this context, let A be the discrete variable taking values 1–4, corresponding to the four proficiency levels from lowest to highest. Expression (1) becomes:

$$P_{jj} = \sum_{A=1}^{4} P(A|S = j)P(Y = 1|A, S = j) \tag{12}$$

and counterfactual probability (2):

$$P_{jk} = \sum_{A=1}^{4} P(A|S = j)P(Y = 1|A, S = k) \tag{13}$$

The performance distribution $P(A|S)$ is estimated non-parametrically, given gender and geographical area (*North West*, *North East*, *Center*, *South and Isles*); the transition probability $P(Y = 1|A, S)$ is estimated with binary logit models. Hence, we refer to the term "semi-parametric" to account for the fact that one factor is estimated parametrically and the other one is not.

Note that although in the relevant literature social class is defined with respect to parental occupation [18], we operationalize S with reference to the highest parental educational attainment[18]; the reason is that with the latter definition there seems to be stronger coherence between the different data sources involved in the correction

[17] These ratios vary from 0.93 for high status-high ability students to 1.32 for low-status-low ability students.

[18] Although we do not employ this classification here, the data allow to classify individuals into three social classes as in the simplified British *National Statistics Socio-Economic Classification,* used for example in Jackson et al. [15].

of sample selection.[19] Hence, in what follows, all terms indicating social origins will refer to parental education.

15.5.2 Sample Selection Correction Factors

As we have seen in Sect. 15.4, in order to correct for sample selection, for the evaluation of $P(A|S)$ we need to estimate $P(G = 1|S) = P(S|G = 1)P(G = 1)/P(S)$. The three factors in the right hand side have been obtained by gender as follows[20]:

- $P(S|G = 1)$ has been estimated directly from the graduates survey data;
- $P(G = 1)$ is the marginal graduation probability; it has been computed as the ratio of the number of graduates in 2001[21] to the number of students who passed the lower secondary final exam in 1996 (ISTAT, *Annuario di Statistiche Demografiche*).
- $P(S)$ is the national distribution of the highest parental educational level for the 1982 birth cohort (the 19 years old at the time of graduation), derived from the 2001 Population Census.[22]

Minor inconsistencies were found: the estimates for females with tertiary and upper secondary educated parents slightly exceeded 1 and were then set to unity. This is likely to be related to the employment of different data sources, which can be affected by non-sampling error[23]; moreover for the estimation of $P(S)$ we had to

[19] Note also that for Italy the odds ratio between Y and S when status is measured by social class is much lower than that relative to the highest parental educational level. Moreover, some recent works seem to be going in the same direction (see e.g. [16]).

[20] Correction factors were first computed also by geographical area; however, due to within-country migrations, some inconsistencies between the different data sources arise. Due to this in the end they were computed at the national level.

[21] Data directly obtained from the Education Ministry Statistical Office.

[22] Since the highest parental educational level distributions for the children who have obtained the lower secondary school degree in a given year are not available, as a proxy we calculate $P(S)$ for the all children born 14 years before according to Census data. These populations do not overlap for two reasons: first, some students may graduate earlier or later, due to repetitions; however, if grade failure is roughly stationary, the difference should be negligible. Second, the Census data includes children who have not obtained the lower secondary school degree. We assume that the students failing to pass the lower secondary school examination belong to lower educated families; the assumption is highly reasonable, since, as we have shown above, the great majority of the students passing the exam with the lowest mark *sufficiente* come from the lowest social strata. The number of students of the lowest social strata has been adjusted by subtracting from it the number of students who have not passed the final exam (data provided by the Ministry of Education); the distributions were then evaluated accordingly.

[23] Sampling variability should enter here only via $P(S|G = 1)$, but standard errors of the estimates are very small, and cannot by itself explain these inconsistencies.

Table 15.1 Estimated probabilities $P(G = 1|S)$ of attaining the upper secondary school degree by parental educational level and gender

Parental education	Tertiary	Upper secondary	Lower sec./primary
Males	0.97	0.92	0.50
Females	1.00	1.00	0.59

use parental educational at age 19 instead of age 14. Final sample correction factors are reported in Table 15.1.[24]

15.5.3 Results

Primary and secondary contributions to $\mathrm{P}(Y = 1|S = j)$ are evaluated following the approach outlined in Section 4.1. A sketch of the procedure is shown in Fig. 15.1.

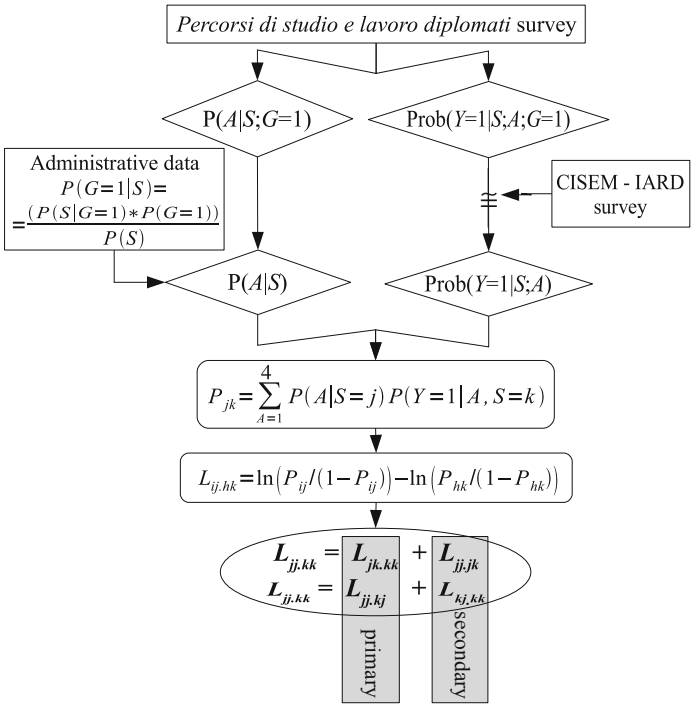


Fig. 15.1 Scheme of the decomposition

[24] Different sets of $P(G = 1|S)$ were applied to check robustness of results: decomposition of expressions in (4) appears to be only slightly affected by mild changes in these percentages.

The final estimates of the ability distribution $P(A|S)$ by gender and geographical area, corrected for sample selection, are reported in Table 15.2. As expected, marks are much higher on average for children from well educated families; moreover we observe that females are better performers than males, and that marks are higher in the South and Islands than in the rest of the country.[25]

Table 15.3 shows the raw observed transition rates to the academic track for all sub-groups. As expected, the propensity to enrol into a *lyceum* is much higher among better performing students and for those from higher status.

$P(Y = 1|A, S)$ is modelled with binary logit regression on A, gender and geographical area for each value of S. Preliminary log-linear analysis highlighted that there are no significant interactions effects, so only the main effects were included in the model. Since dummy coefficient estimates related to the nominal version of A differ by approximately the same amount, in the end we used the model where A is a quantitative variable (taking values 1–4 from *pass* to *excellent*). Results are shown in Table 15.4.

As expected, the propensity to enrol into a liceo is much higher among better performing students. At the same level of demonstrated ability, however, the transition probability is much higher among high status children.[26] Gender differences are less marked. Gender is significant for lower and upper S: females with low educated parents are more likely than males to enter the academic track given their level of ability, while within families with tertiary education transition probabilities are higher for males.[27] Geographical effects are not very clear; since the probabilities P_{ij} resulting from a simplified models without this variable are very similar to those derived from the extended model, the more parsimonious specification was eventually employed in the decomposition.

Estimates of factual and counterfactual probabilities P_{ii} and P_{ij} are reported in Table 15.5. Rows refer to school mark distributions according to parental education, while columns indicate which level of S is used to model the transition probability to the academic track. Thus, the numbers located on the diagonal are the actual estimated transition probabilities $P_{jj} = P(Y = 1|S = j)$. These values are higher for females; stronger gender differences are observed for low S: females with low parental education are almost twice as likely to enrol into a *liceum* than males.

Off diagonal elements P_{jk} are instead counterfactuals, combining lower secondary school marks distributions and conditional transition probabilities for different parental educational levels. For example, $P_{11} = 0.709$ is the (marginal) transition rate of a male with tertiary educated parents. The corresponding

[25] Standard errors of the estimates are not reported; given the complex sampling scheme they could be obtained only with non-standard resampling techniques. Similar arguments also apply to Table 15.3. To give a rough idea of their magnitude with respect to Table 15.2, assuming simple random sampling standard errors would vary between 0.005 and 0.032.

[26] This can be seen from the raw probabilities in Table 15.3 and is reflected in the values of the constant in the logit models in Table 15.4.

[27] Standard errors of the estimates and derived p-values have been computed assuming a simple random sampling, and are thus somewhat underestimated.

Table 15.2 Lower secondary school final mark distribution $P(A/S)$ after sample selection correction, by highest parental educational level, gender and geographical area

	Male					Female				
		Lower sec. school final mark					Lower sec. school final mark			
	Parental educ.	Pass	Good	Very good	Excellent	Parental educ.	Pass	Good	Very good	Excellent
North West	Tertiary	0.17	0.25	0.28	0.29	Tertiary	0.12	0.23	0.22	0.43
	Upper sec	0.33	0.32	0.19	0.17	Upper sec	0.16	0.29	0.29	0.26
	Lower sec/prim	0.69	0.17	0.10	0.04	Lower sec/prim	0.56	0.23	0.13	0.08
North East	Tertiary	0.19	0.33	0.28	0.20	Tertiary	0.06	0.25	0.29	0.40
	Upper sec	0.38	0.30	0.20	0.12	Upper sec	0.15	0.34	0.27	0.23
	Lower sec/prim	0.72	0.15	0.08	0.05	Lower sec/prim	0.58	0.22	0.12	0.08
Centre	Tertiary	0.20	0.28	0.20	0.32	Tertiary	0.10	0.17	0.28	0.45
	Upper sec	0.36	0.30	0.19	0.15	Upper sec	0.20	0.28	0.24	0.28
	Lower sec/prim	0.73	0.15	0.06	0.06	Lower sec/prim	0.59	0.21	0.12	0.08
South Isles	Tertiary	0.16	0.19	0.26	0.39	Tertiary	0.06	0.18	0.23	0.54
	Upper sec	0.34	0.28	0.20	0.18	Upper sec	0.17	0.24	0.23	0.37
	Lower sec/prim	0.69	0.17	0.07	0.06	Lower sec/prim	0.54	0.19	0.13	0.14

Table 15.3 Raw transition rates to the academic track $P(Y = 1|A, S)$ by highest parental educational level, lower secondary school final marks, gender and area

	Male					Female				
		Lower sec. school final mark					Lower sec. school final mark			
	Parental education	Pass	Good	Very good	Excellent	Parental education	Pass	Good	Very good	Excellent
North West	Tertiary	0.37	0.61	0.80	0.90	Tertiary	0.29	0.61	0.82	0.84
	Upper sec	0.13	0.30	0.47	0.77	Upper sec	0.07	0.25	0.50	0.67
	Lower sec/prim	0.04	0.10	0.14	0.46	Lower sec/prim	0.03	0.12	0.28	0.48
North East	Tertiary	0.31	0.58	0.78	0.96	Tertiary	0.21	0.61	0.72	0.91
	Upper sec	0.08	0.20	0.43	0.73	Upper sec	0.08	0.24	0.43	0.64
	Lower sec/prim	0.03	0.07	0.18	0.59	Lower sec/prim	0.03	0.10	0.21	0.56
Centre	Tertiary	0.51	0.68	0.84	0.92	Tertiary	0.32	0.55	0.83	0.93
	Upper sec	0.10	0.26	0.51	0.72	Upper sec	0.15	0.26	0.53	0.74
	Lower sec/prim	0.03	0.09	0.20	0.45	Lower sec/prim	0.06	0.21	0.34	0.50
South and Isles	Tertiary	0.36	0.68	0.69	0.86	Tertiary	0.23	0.42	0.68	0.85
	Upper sec	0.10	0.21	0.39	0.63	Upper sec	0.13	0.22	0.50	0.68
	Lower sec/prim	0.02	0.08	0.28	0.47	Lower sec/prim	0.09	0.13	0.28	0.51

Table 15.4 Logit models for the transition probabilities to academic track

	$S =$ tertiary			$S =$ upper secondary			$S =$ lower sec./primary		
Full model	β	p-val	Exp(β)	β	p-val	Exp(β)	β	p-val	Exp(β)
ind_NorthWest	0.286	0.014	1.33	0.208	0.001	1.23	−0.147	0.101	0.86
ind_NorthEast	0.252	0.053	1.29	−0.034	0.646	0.97	−0.196	0.055	0.82
ind_Center	0.594	0.000	1.81	0.256	0.000	1.29	0.126	0.162	1.14
Gender (female)	−0.290	0.000	0.75	0.069	0.168	1.07	0.385	0.000	1.47
ind_buono	1.030	0.000	2.8	0.955	0.000	2.6	1.050	0.000	2.86
ind_distinto	1.827	0.000	6.22	1.976	0.000	7.22	2.032	0.000	7.63
ind_ottimo	2.765	0.000	15.88	2.895	0.000	18.09	3.088	0.000	21.94
Constant	−0.778	0.000		−2.242	0.000		−3.314	0.000	
Simplified model									
ind_NorthWest	0.289	0.013	1.34	0.210	0.000	1.23	−0.149	0.097	0.86
ind_NorthEast	0.259	0.046	1.3	−0.031	0.669	0.97	−0.197	0.053	0.82
ind_Center	0.595	0.000	1.81	0.257	0.000	1.29	0.125	0.164	1.13
Gender (female)	−0.287	0.000	0.75	0.069	0.169	1.07	0.385	0.000	1.47
A	0.899	0.000	2.46	0.971	0.000	2.64	1.023	0.000	2.78
Constant	−1.614	0.000		−3.213	0.000		−4.327	0.000	
No geographic area									
Gender (female)	−0.288	0.000	0.75	0.072	0.150	1.07	0.382	0.000	1.47
A	0.871	0.000	2.39	0.965	0.000	2.62	1.026	0.000	2.79
Constant	−1.315	0.000		−3.098	0.000		−4.365	0.000	

Table 15.5 Estimates of P_{ij} for Italy

	Male $P(Y = 1\|S; A)$ referring to..			Female $P(Y = 1\|S; A)$ referring to..		
$P(A\|S)$ referring to . . .	Tertiary	Upper sec.	Lower sec./ primary	Tertiary	Upper sec.	Lower sec./ primary
Tertiary	0.709	0.415	0.236	0.726	0.504	0.344
Upper sec.	0.619	0.304	0.153	0.642	0.415	0.271
Lower sec./ primary	0.491	0.185	0.076	0.448	0.248	0.152

probability for an hypothetical child with the ability distribution of the upper class but the (conditional) transition probability of the lower class, is given by $P_{13} = 0.236$; when the ability distribution is that of the lowest class and the transition probability is that of the upper class, $P_{31} = 0.491$.

There is a noticeable tendency to decline faster along rows than along columns, indicating that the differences in family preferences for $Y = 1$ due to S *given* children's marks are more relevant in determining the track choice with respect to school performance differences due to A.

The relative importance of primary and secondary effects is shown in Table 15.6. Both of the formulas in (4) are computed, and produce similar results; average contributions are also reported. The main finding is that *secondary effects* tend to prevail in all contexts, the only exception being that of medium vs. low status females.[28]

It is important to recognise that this result does not imply that class differentials in children's *ability* are weak (see the discussion on measurement error in Sect. 15.3), nor that differentials due to S in children's *school marks* are weak. Results imply instead that *differentials due to S in secondary school choices* are mainly driven by differences in the transition probabilities given previous school performance, while differences in the performance distributions play a weaker role. This may occur either because performance distributions vary little across social status, or because performance does not strongly affect school choices.[29]

The relative importance of secondary effects seems to be stronger for males than for females, and when comparing upper and middle status with respect to middle and low status. With respect to gender, by looking at Table 15.2 we find no clear differences in the social status effect on the performance distribution.[30] Furthermore, from Table 15.4 we derive that the social origin effect on the probability to

[28] Since the estimated quantities are quite complex, the assessment of their standard errors would require a significant effort. Note however that, given the large sample size, we expect them to be reasonably small.

[29] In principle, there could be wide family status differences in the observed level of ability, but if school choices were only weakly affected by performance (choices depending mainly on social status), these differences would not exert a relevant role.

[30] Moreover, by estimating, somewhat improperly, a linear model for performance, we do not find significant interaction effects between gender and status, (i.e. the effect of status on performance does not change with gender).

Table 15.6 Primary and secondary effects decomposition

	Male			Female		
	Tertiary > upper sec.	Tertiary > lower sec./ primary	Upper sec. > lower sec./ primary	Tertiary > upper sec.	Tertiary > lower sec./ primary	Upper sec. > lower sec./ primary
$L_{jj.kk}$	1.720	3.386	1.666	1.316	2.696	1.379
$L_{jk.kk}$	0.406	0.926	0.653	0.392	1.185	0.767
$L_{jj.jk}$	1.314	2.460	1.013	0.924	1.511	0.613
% primary	0.236	0.274	0.392	0.298	0.439	0.556
% secondary	0.764	0.726	0.608	0.702	0.561	0.444
$L_{jj.kj}$	0.484	1.322	0.786	0.359	1.076	0.732
$L_{kj.kk}$	1.237	2.064	0.880	0.958	1.620	0.647
% primary	0.281	0.390	0.472	0.272	0.399	0.531
% secondary	0.719	0.610	0.528	0.728	0.601	0.469
Average % Primary	0.259	0.332	0.432	0.285	0.419	0.543
Average % Secondary	0.741	0.668	0.568	0.715	0.581	0.457

choose the academic track given ability is smaller for females than for males.[31] Given that the effect of ability is very similar across values of *S*, we conclude that the gender difference in the *relative* contributions of primary and secondary effects is due to weaker secondary effects for girls (in *absolute* terms) rather than to stronger primary effects.

Secondary effects are stronger when comparing upper and middle status with respect to middle and low. This difference does not seem to be driven by lower primary effects in the first comparison, since a close inspection of Table 15.2 shows that children ability distributions differ more between the low and medium status than between the medium and high. This finding suggests that high status families attach a strong value to the educational experience per se, regardless of proficiency.

15.6 Conclusions

The results described in Sect. 15.5.3 are particularly interesting when considered within the international context. The most striking finding is that the relative contribution of *primary effects* is *much lower in Italy* than in the other countries for which the analysis has been carried out. Let us review the main results. Primary effects account for about 76% of the total social background effect in UK (Jackson et al. [15], for year 2001), 58% in Stockholm, Sweden (Erikson [7], for 1990), 47% in the German Lander Rhineland (Stocké [20], for 2003), 58% in the Netherlands (Kloosterman et al. [16], for 1999)[32]. The corresponding estimates for Italy are much lower: 29.3% for males and 40.3% for females. Although these values are not fully comparable, because of cross-country institutional differences, definitions of social status[33] and because ability assessments are not always standardized, differences are however large, and it would be of great interest to understand the reasons laying behind them.

We can think of different topics for further work:

(i) In order to interpret the results from a comparative point of view, the *absolute* contributions of primary and secondary effects should be evaluated together with the *relative* ones. This implies recovering comparable estimates of the total effect of social origins on school choices. Note however that cross-country comparisons are even more problematic in this case: employing parental education or social class can give rise to substantials differences within countries.[34]

---

[31] See the constant and the gender coefficient.

[32] The percentage with respect to the high-low status comparison is reported here.

[33] In UK and Sweden father's social class, in Germany mother's social class, in the Netherlands and Italy the highest parental educational level.

[34] We can see this from PISA, for which common alternative definitions are possible. Taking the highest parental educational level the following raw OR between high and low social status are found: Netherlands 4.7, Italy 6.9, Germany 12.9. Taking social class, Netherlands 8.5, Italy 5.8, Germany 8.4.

(ii) The low importance of primary effects in Italy with respect to other countries can have two alternative interpretations: (a) social background differentials in the school performance distributions are relatively weak; (b) the role of ability in educational decisions is weak. Given the difficulties in cross-country comparisons based on national data, evidence from the international assessment carried out on 4th graders, PIRLS (*Progress in International Reading and Literacy Study;* [17]) can help to shed some light on this issue. Simple regression analysis indicate for example that Italy is one of the countries with the *lower* inequality of opportunity with respect to performance scores near the end of primary school.

(iii) The assessment of how specific institutional features – in particular, early tracking – affect equality of opportunity in education is the focus of an interesting body of work [5, 13, 21]: by employing international surveys like PISA, the school design effect is identified by exploiting the cross-country variability. To our knowledge no attempt has been done yet to deepen the understanding of how institutional features promote or discourage primary and secondary effects.[35] In order to put forward educational policies with the aim to reduce educational inequality it would be very useful to try opening the *black box* and separate the effects on school performance from those on choices given performance. At the moment this aim is difficult to accomplish: on one hand it is difficult to harmonise national data to allow for adequate cross-national comparisons, on the other hand international data such as PISA cannot be employed for this purpose, because no measure of ability before school choice is available. This could be an interesting challenge for future research.

References

1. Allmendinger J (1989) Career mobility dynamics: a comparative analysis of the United States, Norway, and West Germany. Max-Planck-Institute fur Bildungsforschung, Berlin
2. Boudon R (1974) Education, opportunity and social inequality. Wiley, New York, NY
3. Breen R, Goldthorpe JH (1997) Explaining educational differentials. Towards a formal rational action theory. Ration Soc 9(3):275–305
4. Breen R, Jonsson JO (2000) A multinomial transition model for analyzing educational careers. Am Sociol Rev 65:754–772
5. Brunello G, Checchi D (2007) Does school tracking affect equality of opportunity? New international evidence. Econ Policy 22(52):781–861
6. Contini D, Scagni A (2010) Equality of opportunity in secondary school enrollment. Comparing Italy, Germany and the Netherlands. Qual Quant, DOI is 10.1007/s11135-009-9307-y

[35] To give an example, why is it that in Italy primary effects are so low? Could it be due to the fact that the compulsory school system is quite highly standardised in Italy? (standardization refers to the degree to which the quality of education meets the same standards nationwide; Allmendinger, 1989). On the other hand, secondary effects are strong. Is this related to the absence of performed-based restrictions to enrolmment into the academic track, at work in other countries (in the Netherlands for example)?

7. Erikson R (2007) Social selection in stockholm schools: primary and secondary effects on the transition to upper secondary education. In: Scherer S, Pollak R, Otte G, Gangl M (eds) From origin to destination. Trends and mechanisms in social stratification research. Campus Verlag, Frankfurt, pp 61–81
8. Erikson R, Goldthorpe JH (1992) The constant flux: a study of class mobility in industrial societies. Clarendon Press, Oxford
9. Erikson R, Goldthorpe JH, Jackson M, Yaish M, Cox DR (2005) On class differentials in educational attainment. Proc Natl Acad Sci 102(27):9730–9733
10. Eurydice, Italian Unit (2006) Eurybase the Information Database on Education Systems in Europe, The education system in Italy 2005/2006, http://www.eurydice.org
11. Goldthorpe JH (1996) Class analysis and the reorientation of class theory: the case of persisting differentials in educational attainment. Br J Sociol 45(3):481–506
12. Hannum E, Buchman C (2003) The consequences of global educational expansion. Occasional paper of the American Academy of Arts and Sciences, Cambridge, MA
13. Hanushek EA, Woessman L (2005) Does educational tracking affect performance and inequality? Difference-in-differences evidence across countries, IZA Discussion Paper n. 1901
14. ISTAT (2004) Percorsi di studio e di lavoro dei diplomati, Roma
15. Jackson M, Erikson R, Goldthorpe JH, Yaish M (2007) Primary and secondary effects in class differentials in educational attainment: the transition to A-level courses in England and Wales. Acta Sociol 50(3):211–229
16. Kloosterman R, de Graaf P, Ruiter S, Kraaykamp G (2007) Parental education and the transition to higher secondary education. A comparison of primary and secondary effects for five Dutch cohorts (1965–1999). In: Proceedings of the RC28 Spring meeting in Brno, Czech Republic
17. Mullis IVS, Martin MO, Gonzalez EJ, Kennedy AM (2003) PIRLS 2001 International report: IEA's Study of reading literacy achievement in primary schools, Boston College
18. OECD (2005) PISA 2003 Technical report
19. Shavit Y, Blossfeld HP (eds) (1993) Persistent inequality: changing educational attainment in thirteen countries. Westview, Boulder, CO
20. Stocké V. (2007) Strength, sources, and temporal development of primary effects of families' social status on secondary school choice, Sonderforschungsbereich 504, WP series, n. 07-60
21. Woessmann L (2007) Fundamental determinants of school efficiency and equity: German states as a microcosm for OECD countries. CESifo working paper 1981

Chapter 16
Labour Market Outcomes for Ph.D. Graduates

Antonella D'Agostino and Giulio Ghellini

16.1 Introduction

In the international framework a high presence of Ph.D. graduates in the labour market has often been identified as a key factor for innovation and for creating technological progress. The Ph.D. graduates are at the same time the most qualified people in terms of educational attainment and those who are trained and most inclined for research careers, therefore they are expected to contribute to the advancement and diffusion of knowledge and technologies. Recently a work at the OECD has raised a number of questions about their education-to-work transition, employment and mobility patterns [4] and one of the aim of the established organization of European Ph.D. students [9] was to improve working and studying conditions for young scientists in order to increase their commitment on European research and to improve the outcomes of European science. Consequentially in Europe it is becoming more and more frequent the adoption of well defined survey design on this population for monitoring their working careers. In spite of that, in Italy the information framework on training and working experience of Ph.D.s. seems to be quite inadequate and fragmented [20], even if Ph.D. studies have been introduced more than 20 years ago. The lack of information on working careers after Education is, unfortunately, a typical trait of the Italian information system on school to work transitions. In any case seems to be rather worrying that also for the restricted sub-population of Ph.D.s. – characterized by the higher level of public investment both in terms of time and of resources – the interest to better know labour market entrance, work experiences and job satisfactions has been so lightly pursued. Unlike Italy, several authors studied the labour market performances or the training of Ph.D.s. in the international view. In the last 20 years, Stephan and Levin [18] studied the adequacy of Ph.D.s. supply; Stephan and Everhart [19] considered the rewards to their education; Martinelli [12] made a descriptive analysis of the labour market outcome of French Ph.D. graduates

A. D'Agostino (✉)
Dipartimento di Statistica e Matematica per la Ricerca Economica, Università di Napoli Parthenope, 80133 Napoli, Italy
e-mail: antonella.dagostino@uniparthenope.it

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_16,

in science and engineering; Nerad and Cerny [14] discussed postdoctoral patterns; Mangematin [13] evaluated the Ph.D. job market from the professional trajectories during the Ph.D., Enders [8] studied Ph.D. experience on the labour market in Germany. Therefore what is really urgent now in Italy is to define a clear perspective for the definition of coherent survey system for monitoring Ph.D. experience from the beginning of their training experience till their labour market entrance and job experience. The availability of current information on that could in fact be useful not only for evaluation of Ph.D. programmes but also to give information to those young students who wish to pursue a career in research. From this perspective a survey project on Ph.D. graduates started at the university of Siena. The main aim of this survey is to reconstruct retrospectively the work history of the Ph.D. graduated with particular attention to the education acquired. In this chapter, using these data, we propose to concentrate the attention on the correlation between earnings 1-year after the end of the Ph.D. and some individual characteristics with the aim to give some interesting signal for policy makers. We are aware that the information available has limits that should be emphasized. A limitation of the data is that the labour market position of Ph.D. graduates is only studied 1-year after the end of Ph.D. and significant divergences in careers could be maybe better observed after 5–10 years of working life; moreover it is very difficult to provide if differences in earnings have a causal interpretations or are simply due to differences in the composition of the groups analyzed. Nevertheless the chapter provides interesting results that help to answer at least in part to this important issues concerning Ph.D assessment for the most unknown and unresolved. The structure of the remainder of the chapter is as follows. Section 16.2 describes the data used for the empirical analysis. Some descriptive statistics on the Ph.D.s. profile and the principal characteristics of their job are provided in Sects. 16.3 and 16.4. Section 16.5 presents the methodology used and the empirical results on the determinants of both the employment status and the earnings 1-year after the end of Ph.D. Section 16.6 concludes the chapter.

16.2 The Data

We draw data from a new cross-sectional survey on Ph.D. graduates conducted in a well known university in Italy, the University of Siena. This survey is representative of all Ph.D. graduates from 2003 to 2006 in such university and it is also one of the few experiences in Italy (for example, we can quote Associazione Dottorato Italiano -ADI, [17]).[1] Data have been collected on four cohorts of Ph.D. graduates identified respectively as XV cohort (enrolled in 1999), XVI cohort (2000), XVII cohort (2001) and XVIII cohort (2002). This population yields 726 observations (for survey design, foreign Ph.D.s have been dropped, they were only 4.47% of the total

[1] At European level it is worth mentioning the recent survey Career Doctorate Holders (CDH) project launched by Eurostat in cooperation with UNESCO and the OECD that involved 23 countries in 2007, unfortunately excluding Italy [5].

population). Data were collected by web methodology and the respondents were asked a variety of questions relating to their Ph.D. training, employment histories and socio-economic background. The collected work history has such a level of detail that it has been possible to derive the employment status 1-year after the Ph.D. thesis and earnings relative to it. Although knowing the limitations of retrospective data this employment condition represents an important source of information on the immediate labour market outcomes of Ph.D. graduates and it is generally used in the research literature on higher education [2, 3]. A satisfactory 78% response rate has been achieved though the complete response rate (regarding only those who have compiled the overall on-line questionnaire) is equal to 68%. In order to account for unit non response bias, cross-sectional weights for persons were derived. Rigorous details of the survey design and the quality of data are discussed in Ghellini and Mulas [10].

16.3 Socio-Biographic Background of Ph.D. Graduates

A considerable proportion of Ph.D. graduates (28.7%), particularly those in Human Sciences (HS) field had not a scholarship during the doctoral training (Table 16.1).[2] About 73% of no beneficiaries of a scholarship were employed, probably starting the doctorate with the hope of career. Of all recipients of a scholarship, those who had funding outside the university are very few (6.0%). Having received a scholarship is one of the main differences between the biographic background. The average age at the university degree is around 26 years in Experimental Sciences (ES), Biomedical and Medical Sciences (BMS), HS fields and 1 year less in Business and Economics Sciences (BES) discipline, while gender differences are quite negligible. This pattern characterizes both sets of Ph.D. graduates. The average age at the beginning of the Ph.D. is generally high, 28 and 27 years respectively for the two groups; consequentially the graduates without scholarship attain their degree on average at

Table 16.1 Funding source of scholarship by cohort and field of study

Funding source of scholarship	Ph.D. cohort				Field of study				Total
	XV	XVI	XVII	XVIII	ES	BMS	HS	BES	
Without	25.7	31.9	30.0	25.9	23.4	23.0	41.4	25.2	28.7
University	74.3	68.1	56.4	64.8	69.7	71.7	53.4	67.3	65.3
Others	–	–	13.6	9.3	6.9	5.3	5.2	7.5	6.0
Total (N = 497)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

ES = Experimental Sciences, BMS = Biomedical and Medical Sciences,
HS = Human sciences, BES = Business and Economics Sciences.

[2] In Italy, since 1999, it is possible to offer enrolment in doctoral studies without offering a scholarship. The main source of funding is the Ministry of Education, University and Research, whose yearly budget for scholarships is divided among the Universities, and hence among Ph.D. programmes.

33 years (34 years in BMS field), while those with the scholarship at 31 years and half. Moreover the waiting time from the graduation and the beginning of Ph.D. on average is higher in BMS and HS fields. The former group (BMS) is principally directed to specialization courses and afterward eventually to a Ph.D.; the latter group (HS), is often characterized by individuals who work in primary and in secondary education and consequentially their waiting time is long. However, a further look at the overall population of Ph.D. graduates shows average waiting time generally longer than 2 years, the only exception is in BES field of study. This evidence prevents a linear pattern in post-graduation careers and may reflect an ineffectiveness in the overall Italian higher education system. In addition, the national mobility is very low. The Ph.D.s. retained by the university of Siena are 50.5%. The Ph.D.s recruited from southern regions are only 5.2% whereas the Ph.D.s recruited from the other geographic areas are equally represented. Furthermore, ES and BMS fields are mainly characterized by low mobility Finally, we discover that a favourable socio-economic background characterize this population: 31% of Ph.D. graduates has parents with an university degree, 9% with a further education level.

16.4 Employment Status and Earnings One-Year After the Ph.D. Thesis

One-year after the doctorate degree about 82% of the Ph.D. graduates is employed and among those who are not employed about 21% is in vocational post-doctorate training (Table 16.2). The risk of not being employed is not widespread across cohorts as well as across disciplines. The rate of the not employed graduates for the last cohort is about two times the one of the first cohort. Graduates in the ES show the highest employment rate and those in HS the lowest. Males are slightly more employed than females. This may reflects the beginning of a slow transformation process that an academic career, due to low earnings and unstable perspectives, may becoming a woman-like job. Table 16.3 lists some descriptive statistics of the variables that characterized the occupation separated by gender. The job is mainly concentrated in Siena even if it is not equally shared by gender: the percentage of women is higher. Only 10.1% of Ph.D. graduates works in other countries and they are generally men. As expected, mobility towards the south of Italy is limited: only 5.2% found an occupation in southern regions or islands. The remaining part of the graduates is equally shared among the other locations. With reference to differences

Table 16.2 Employment status 1-year after the Ph.D. thesis by structural variables

Employment status	Ph.D. cohort				Field of study				Gender		Total
	XV	XVI	XVII	XVIII	ES	BMS	HS	BES	M	F	
Employed	85.8	81.9	84.5	72.8	87.3	79.6	77.4	81.2	82.3	81.0	81.6
Not employed	14.2	18.1	15.5	27.2	12.7	20.4	22.6	18.8	17.7	19.0	18.4
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 16.3 Characteristics of the employment condition by gender

Job location	Male	Female	Total
Siena	36.3	40.2	38.5
Other location in Tuscany	17.5	13.0	15.0
Region of North Italy	16.6	17.9	17.3
Region of Centre Italy	10.2	16.9	13.9
Region of South Italy or Island	6.4	4.2	5.2
Other country	13.1	7.8	10.1
Total (N=318)	100.0	100.0	100.0
Job type			
Temporary job	14.0	16.7	15.5
Permanent job	25.4	11.3	17.8
University research grant	27.0	31.8	29.6
Atypical job	19.6	35.0	27.9
Self-employment	14.1	5.2	9.3
Total (N = 388)	100.0	100.0	100.0
Job sector			
Public University	40.2	52.4	47.0
Private University	2.5	1.3	1.8
Public research institution	9.1	6.0	7.4
Private research institution	3.7	4.2	4.0
Private firm – industrial sector	5.2	9.1	7.4
Private firm – service sector	8.7	4.7	6.5
Public administration	16.4	12.0	14.0
International organisations	0.6	–	0.3
No government organisations	0.8	–	0.3
Other occupations	13.0	10.3	11.5
Total (N = 398)	100.0	100.0	100.0

by job type it is not surprising to note that the Ph.D. graduates in permanent job are very few and they are principally men, whereas many of them have a university grant contract or, worse, an atypical job. Academic post-docs experience can not be considered a clear job status since the great majority of the positions are not regulated by employment contracts and are quite similar to individual grants, and atypical job (short-term contracts tied to a particular funding stream and/or research project) is a very unstable job that is frequently underpaid. In addition females are over represented in this work condition. Finally the Ph.D. graduates in self-employment are only 9.3% and principally they are men.

The public university captures about 47% of Ph.D. graduates and females are over represented whereas in public research institution we find more men than women as well as in public administration, in private firm – service sector and in other occupations. About half of Ph.D.s. graduating each year of these succeed in getting a post-doc position at University or in a Public Research Institute whereas Ph.D. graduates in private sectors are generally few. This reflects the Italian society where opportunities in the private sector for high-level young scientists are quite limited, due to the low-technology base of most of the Italian industries and to the

Table 16.4 Average of earnings and hourly earnings by structural variables

	Monthly earnings			Average weekly worked hours[a]			Hourly earnings		
	Mean	Std	N	Mean	Std	N	Mean	Std	N
Gender									
Male	1,441.9	664.3	151	38.6	14.0	168	9.3	4.2	149
Female	1,177.4	511.4	188	35.2	13.6	207	8.7	5.1	188
Field of study									
ES	1,297.3	463.3	124	40.1	11.2	133	8.0	3.7	122
BMS	1,462.7	723.3	80	39.9	12.5	90	9.0	5.1	79
HS	1,106.7	531.6	94	28.5	15.5	100	10.3	5.3	89
BES	1,361.7	697.1	48	40.1	12.9	52	8.6	5.0	47
Ph.D. cohort									
XV	1,436.4	697.6	74	39.0	14.0	79	9.4	4.9	72
XVI	1,327.7	671.2	89	34.2	13.6	97	9.9	5.4	87
XVII	1,214.8	490.7	105	37.2	13.2	115	8.2	4.2	102
XVIII	1,200.8	506.0	78	37.7	14.8	84	8.4	4.3	76
Total	1,293.9	596.9	346	36.9	13.9	375	9.0	4.7	337

[a] Hours worked are set by the respondent as the time spent in employment.

small size of most enterprises. As a further note we discovered also that the job position, logically computed only for employees, is characterized by the highest percentage of Ph.D. graduates employed as technical clerks followed by teachers and in this latter category females are over represented. Whereas in manager and lecturer positions we found more men than women as well as in manager staff. The overall picture of the average earnings (we refer to the net earnings, i.e. after tax and social security contributions expressed in prices 2005) offers an interesting discussion (Table 16.4). We observe high disparities in monthly earnings between men and women: generally men earn about 22% more than women but women generally work less (or they declare to work less) therefore the resulting average hourly earnings is slightly higher for males than for females (about 7% more). Ph.D. graduates in BMS have the highest salaries but they also show a high value of the average weekly worked hours and therefore they do not have the highest value of the average hourly earnings. On the other hand, even if graduates in HS earn less in comparison to the other fields of study, they get on average the highest hourly are (this result could be explained by the presence of teaching positions). Finally, the most disadvantaged are the Ph.D. graduates of the latest two cohorts: their average monthly earnings and their hourly earnings are the lowest.

16.5 Modelling Earnings of Ph.D. Graduates

16.5.1 Methodological Background

Given our main interest in studying the correlation between earnings of the Ph.D. graduates and some individual characteristics, we have to take into account the

possibility that some unobservable individual factors affecting earnings equation can be also correlated with those driving the propensity to the labor force. In other words the working Ph.D. graduates may not be a random subset in our sample. If there is selection bias, using standard OLS, the earnings regression on observed covariates can seriously imply misleading results. For this reason, these kinds of data are generally fitted using the well known Heckman's model [11] that takes into account the selection mechanism. On the other hand the paper of Copas and Li [6] highlighted many deficiencies of this model and they introduced a locally sensitivity analysis approach in order to discover if the standard inference is really sensitive to the departure from the hypothesis of the ignorability of the selection mechanism. In this chapter, we perform a locally sensitivity analysis using the methodology introduced by Troxel et al. [21] that generalized the approach of Copas and Li [6] and Copas and Eguchi [7]. Let y_i be the observed earnings for the ith-Ph.D. s, assumed linearly related to a vector of explanatory variables x (including the intercept term) through the standard multiple regression model:

$$y_i = x_i'\beta + \epsilon_i, \tag{1}$$

where β is the associated vector of parameters to be estimated and ϵ_i is an error term assumed normally distributed with mean zero and variance τ. We allow the probability of y_i being observed ($z_i = 1$) to depend on the value y_i through the parameter θ as follows:

$$Pr\left[z_i = 1|y_i, x_i\right] = h\left(x_i'\gamma + \theta\ y_i\right) = \frac{\exp(x_i'\gamma + \theta\ y_i)}{1 + \exp(x_i'\gamma + \theta\ y_i)}, \tag{2}$$

where γ is a vector of unknown parameters and $h(.)$ is the logistic link function. The sensitivity analysis is conducted in terms of θ. We evaluate the extent to which an estimate of β for fixed θ, say $\hat{\beta}(\theta)$ depends on the value of θ, where $\theta = 0$ corresponds to the ignorability assumption of the selection process. Troxel et al. [21] proposed the index of sensitivity to nonignorability (ISNI) given as:

$$ISNI = \frac{\partial\hat{\beta}(\theta)}{\partial\beta}_{|\theta=0} = -\hat{\tau}_0(x_s'\ x_s)^{-1}\ x_{ns}'\ h_{ns}, \tag{3}$$

where x_s and x_{ns} are the matrices of predictors for subjects with $z_i = 1$ and $z_i = 0$ respectively; h_{ns} is the vectors of the propensity scores for subjects with $z_i = 0$. Since in Eq. (2) we used the logistic specification, θ is interpreted as the log odds ratio in the observation probability associated with a one-unit change in y. Indeed $\theta = 1$ implies a substantial degree of selectivity, it would mean that the odds for Ph.D. graduates who do not work differ from the average by a factor of around 3.[3]

[3] We are using the logistic selection model, in which case θ is the log odds ratio in the observation probability associated with a one-unit change in y therefore $\theta = 1$ implies a unit increase in y corresponds to an $e^1 = 2.7$ increase in odds of being observed.

In order to make the interpretation of ISNI independent from the scale of y Troxel et al. [21] proposed the sensitivity transformation c defined as follows:

$$c = |\frac{\sigma_y \, SE_y}{ISNI_y}|, \tag{4}$$

where σ_y is the standard deviation *(SD)* of y and SE_y is the standard error of a regression coefficient of interest. The interpretation of c is similar to the parameters of sensitivity to confounding bias in observational studies, as described by Rosenbaum [16] i.e. c is the scale on which the sensitivity is extreme enough that an odds ratio of about 3 corresponds to an effect on β of one standard error. In summary, if c is large, then there is sensitivity only if the nonignorability is extreme, therefore the conclusion based on $\theta = 0$ can be regarded as safe. On the contrary if c is small, say less then 1, there is potential sensitivity even for modest nonignorability.

16.5.2 Sensitivity Analysis

We used the natural log of earnings as dependent variable in the earnings equation. The dependent variable of the selection equation "employment status" is a dichotomous indicator that equals 1 if an individual is employed and 0 if not. The covariates included in the analysis are: the gender, the field of study, the cohort, the age at graduation, the benefit of a scholarship during the Ph.D. training, a dummy variable that indicates if Ph.D.s acquired the graduation at the university of Siena, a dummy variable that indicates whether Ph.D. graduate worked before the end of Ph.D. and finally the parents educational level. Figure 16.1 shows the profile log-likelihood curve for ρ under the hypothesis of the Heckman's model. The parameter ρ indicates the correlation term between the selection equation and the equation

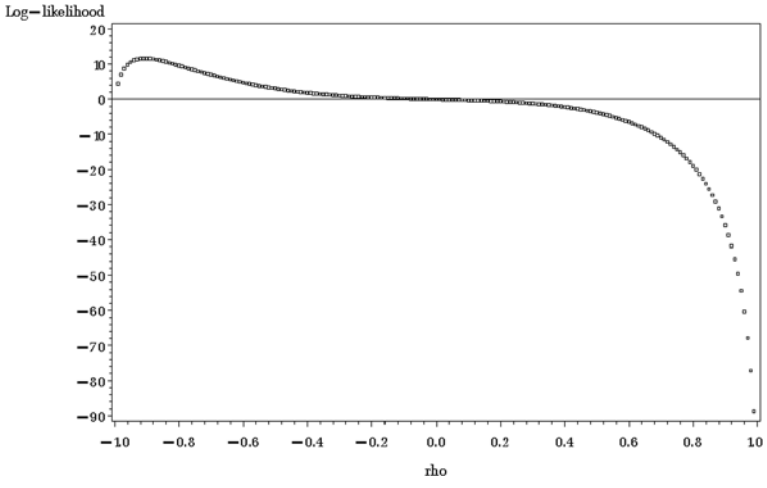


Fig. 16.1 Profile log-likelihood function

Table 16.5 ISNI analysis

Predictors	ISNI	c
Constant	−0.0470	0.95
Gender	0.0007	29.84
Field of study ES	0.0144	1.98
Field of study HS	−0.0117	2.58
Field of study BES	−0.0092	4.19
Cohort XVI	−0.0021	14.44
Cohort XVII	0.0022	13.18
Cohort XVIII	−0.0165	2.02
Location of university degree	−0.0168	1.34
Previous work experience before the end of Ph.D	0.0459	0.52
Scholarship	0.0116	2.28
Family educational level	0.0076	2.97
Age at graduation between 25–28	−0.0154	1.56
Age at graduation over 28	−0.0175	2.01

of interest expressed in (1). We observe that the curve is extremely flat near ρ=0 (data have been scaled so that the profile log-likelihood is equal to zero under the hypothesis of ignorable selection), confirming that the data give almost no information about the selection parameter. The maximum is attained far from zero at about $\rho = -0.91$ and Copas and Li [6] established that the confidence intervals for ρ are very wide and the point of maximum is strongly influenced by the transformation used for data (we use the log transformation). These findings suggest to perform the sensitivity analysis introduced in Sect. 16.5.1. The results of the sensitivity analysis are summarized in Table 16.5. On the whole, the c values are large and suggest only modest sensitivity. For example the β coefficient for "gender" predictor would be insensitive unless a change of 1/29.84 SDs was associated with an odds ratio of about 3 in the observation probability, which means very strong nonignorability. The only predictor that seems to be more sensitive to the ignorability assumption is the "previous work experience before the end of Ph.D." followed by the constant term that however shows a value close to 1. The implication is that in our data the results seem quite robust to the violation of ignorability assumption since only a severe nonignorability and stronger than one expected in practical situations is needed before we could come to a different conclusion than that the conventional inference must be regarded with caution. Therefore we estimate earnings equation under the assumption of ignorable selection process.

16.5.3 Results

Table 16.6 presents the marginal effects from the selection and earnings equations estimated under the ignorability assumption of the selection process. In the earnings equation we add other covariates that account for job characteristics.[4] Since the

[4] In Sect. 16.4 we observed that many variables are affected by non-response item. Even if the percentage in almost all variables is nearly negligible when we put all variables together in the

Table 16.6 Parameter estimates of earnings and selection equations

Variables	Earnings equation $\mu(0) = 1{,}300$	Selection equation $P(z_i^* > 0\|x) = 0.82$
	Marginal effects (s.e.)	Marginal effects (s.e.)
Gender (female)		
– Male	0.12 (0.09)	−0.03 (0.04)
Cohort (XV)		
– XVI	−0.05 (0.06)	−0.02 (0.06)
– XVII	−0.07 (0.06)	0.00 (0.06)
– XVIII	−0.14 (0.06)	−0.15 (0.07)
Age at university graduation (less than 25 years)		
– between 25 and 28	−0.04 (0.05)	−0.13 (0.06)
– over 28	−0.14 (0.08)	−0.13 (0.08)
University degree (not graduated in Siena)		
– Graduated in Siena	−0.02 (0.04)	−0.14 (0.05)
Scholarship (without)		
– With	−0.01 (0.05)	0.07 (0.05)
Family educational level (less than university degree)		
– High	0.06 (0.04)	0.04 (0.04)
Previous work experience before the end of Ph.D. (no)		
– Yes	0.09 (0.05)	0.15 (0.07)
Job location (Italy)		
– Other country	0.48 (0.08)	–
Job type (university research grant)		
– research temporary contract	−0.07 (0.09)	–
– research permanent contract	0.15 (0.09)	–
– research atypic contract	−0.34 (0.06)	–
– other temporary contracts	0.03 (0.08)	–
– other permanent contracts	0.29 (0.07)	–

dependent variable in the outcome equation is the natural logarithm of the earnings, the marginal effect corresponds to a relative change in earnings. For example, if a is the estimated value of the marginal effect, the estimated percentage change in earnings due to a unit increase in x_{ki} is $[exp(a) - 1]100$. The reference individual who has all the dummy variables equal to zero has the following identikit: female, Ph.D. in BMS field, XV cohort, less than 25 years old at graduation, not graduated in Siena, without scholarship during her Ph.D., her parents have a medium/low educational level, with previous work experience, having an university research grant, working in Italy. These characteristics describe a female who earns a log earning of

regression model we lost many observations. In order to avoid eliminating observations from the population, we have decided to impute missing data using the imputation software IVE-ware [15].

about 7 euro that converted back to the real earning is equal to a earning of 1,300 euro. Several marginal effects in both equations are statistically significant at the 5% level and have the expected signs. The field of study influences both equations.

The probability to be employed is equal to 0.82 for the benchmark female. This probability increases by 0.09 if she has a Ph.D. in ES field, we discover no significant differences with the other disciplines. On the other hand her earnings decrease respectively by about 28% and 13% if she has a Ph.D. in HS or in BES. The cohort effect is evident in both equations. Ph.D. graduates belonging to the latest cohort (XVIII) have the lowest probability to be employed and their expected earnings decrease by 14%. On the contrary, we do not discover gender discrimination on the probability to be employed whereas to be male increases the expected earnings by 12% but only at 10% level of significance. The age at graduation has a significant effect even if in the earnings equation only the threshold over 28 years seems to have a significant effect. In summary, the probability of being employed decreases by 0.13 if the either age ranges between 25 and 28 years or it is over 28 years. The expected earnings decrease by 14% if the age is over 28 years. As expected the university location has a significant effect only in the selection equation: the probability to be employed decreases by 0.14 if the Ph.D.s. acquired the graduation at the university of Siena. Previous work experience has a slightly positive effect on earnings (the expected value increases by only 0.09% if she worked before) whereas, as expected, the positive effect in the selection equation is considerable (the probability to be employed increases by 0.15). However this effect has to be considered with caution being exposed to the ignorability assumption. Concerning the effects of the labour market variables, several considerations can be made. Ph.D. graduates in other countries have higher earnings than all those graduates employed in Italy. The expected earnings increase by 48% for Ph.D. graduates working in foreign countries. Those who are employed in research or other sectors with an atypic contract generally earn less than those have a university research grant, the expected earnings decrease by 34% for those who are in research and by 16% for the others. On the other side those who have a permanent contract in public or private sectors out of research earn more. Specifically the estimated expected value of the benchmark female graduate increases by 29%.

16.6 Further Discussion and Conclusions

While higher education systems undergo many transformations every where, little is known about the current developments in the labour outcomes and career paths of doctoral graduates especially in Italy. In this chapter we provided some interesting results in order to fill up at least for a local context the lack of information on this target group. We examined the labor outcomes among the highly educated workers in four discipline (Experimental Sciences, Biomedical and Medical Sciences, Human Sciences and Business and Economics Sciences) that acquired their Ph.D. at the University of Siena. Studying the employment outcomes of Ph.D. graduates

we discovered that this is a highly heterogeneous group of people with a main common denominator of a high level of education. In particular, we find that Ph.D.s acquired, on average, their qualification quite late (32 years old). These results are also in line with the findings obtained from the survey conducted by the university of Milan that we mentioned in Sect. 16.2. In addition, we discovered a very poor mobility concerning the recruitment of Ph.D.s. Moreover, they generally come from families with high educational levels. In sum, as we expected, these descriptive findings reflect the latent static nature of Italy in several sections of the social and economic life. Furthermore from the descriptive analysis, we discovered that 1 year after the Ph.D. graduation the employment rate was about 82%. Although we are aware that the information available does not allow consideration of possible problems of selection, the crude comparison of net monthly earnings between Ph.D.s. and graduates can be however an important starting point for an interesting research issue. The net monthly earnings of Ph.D.s. are on average 1,294 euro. Therefore as expected, wages among Ph.D. graduates are higher than those earned by university graduates that according to Almalaurea [3] are on average only 991 euro 1 year after graduation. However if we take into account the return to Ph.D. we have to compare the 1,294 euro with the wages at 4/5 years after the graduation. According again to Almalaurea that collect data on wages after 5 years, graduates earn about 1,300 euro. Furthermore, in the international comparison, Italian earnings result lower than in nearly all the other European countries. This evidence is really bleak and it would suggest to reconsider the matter taking into consideration the possible causal relationship which has been mentioned before. In this chapter, we studied the correlation between Ph.D.'s earnings and some of their characteristics in order to stimulate a constructive debate that would respond in part to important and urgent issues in light of the big investment that the doctorate is the same. Maybe policy makers should take note for educational policy. From a methodological point of view we followed Troxel et al. [21] by using a sensitivity approach rather than attempting to estimate a full model that takes into account the selection process. In particular, the sensitivity analysis suggested that the empirical results are quite robust to the departure from the ignorability assumption, therefore we applied standard analysis in order to model the earnings function and to determine which factors have a significant and positive influence on the expected earnings. We discovered that several variables affect the expected wages: the gender, the field of study, the cohort, the age at graduation, the work history, the job location and finally the job type. From an interpretative point some more considerations can be made. Even if we stress that our model maybe does not allow to give a proper measurement of the true causal effect of education on earnings, some crucial signals of lack of the higher educational system have been highlighted and they should be studied more thoroughly for the implementation of efficient active politics. Taking into consideration the characteristics of the reference individual described in previous paragraph we discovered that in terms of field of study our findings confirmed our expectation. A Ph.D. in HS or BES fields penalizes the future earnings in comparison with the more scientific fields as experimental or biomedical and medical sciences; this evidence reflects the trend in the university framework in Italy, where the scientific faculties generally give more chances in

terms of job. We also found that job location has a crucial role in the earnings expectation as well as the type of job contract and the job sector. Substantially, we established that working in research sector penalizes in terms of earnings but the same individual working in a foreign country improves its economic situation. Indeed, we found that a Ph.D. graduate working in Italy in a research sector tends to yield much lower wages than those earned by people with the same fixed characteristics working in another country. Considering that the research sector in Italy is generally the natural outcome after the Ph.D. our findings are rather disarming. In addition, we discovered that those who have been locally recruited to Ph.D. have a lower probability to be employed. This may reflects a typical Italian behavior in terms of the recruitment of potential Ph.D. students that often happens at local level without investigating their actual abilities. In sum the overall picture would be still more worrying [1]. Fortunately, on the contrary, individual abilities seem to be safe: the lower is the age at graduation, the higher are both the expected earnings and the probability to be employed. Finally, there are a still few limitations of this study that are worth discussing. First, we use as dependent variable in the outcome equation the monthly earnings (expressed in log scale) even if we have no information about the part-time or the full-time work condition. The choice of monthly instead of hourly earnings can be a matter of opinion but we believe that for the high qualified collective examined seems to be more important the total income despite of the time used to get it. Second in the chapter all has been explored within the assumed parametric model and so our discussion in terms of sensitivity analysis inherits any inadequacies in the model assumption. In spite of that the present chapter is significant to literature because it focuses on the aspects of the employment and placement of Ph.D. graduates in different fields of study which is necessary to be monitored in order to meet the request for information and support decision-making in higher education and science policy.

References

1. Abravanel R (2008) Meritocrazia. Quattro proposte concrete per valorizzare il talento e rendere il nostro paese piú ricco e piú giusto. Garzanti Libri, Milano
2. AlmaLaurea (2006) Employment condition of Graduates-2005 Survey. www.almalaurea.it
3. AlmaLaurea (2007) Employment condition of Graduates-2006 Experimental Survey. www.almalaurea.it
4. Auriol L (2004) Conclusions of workshop (DSTI/EAS/STP/NESTI(2004)28). OECD workshop on user needs for indicators on careers of doctorate holders, September 27, 2004, Paris
5. Auriol L, Felix B, Fernandez-Polcuch E (2007) Mapping careers and mobility of doctorate holders: draft guidelines, model questionnaire and indicators. OECD STI working paper 2007/6
6. Copas JB, Li HG (1997) Inference for non-random samples. J R Stat Soc B 59(1):55–95
7. Copas JB, Eguchi S (2001) Local sensitivity approximations for selectivity bias. J R Stat Soc B 63(4):871–895
8. Enders J (2002) Serving many masters: the Ph.D. on the labor market, the ever-lasting need of inequality, and the premature death of Humboldt. Higher Educ 44:493–517
9. EURODOC (2007) http://www.eurodoc.net

10. Ghellini G, Mulas A (2007) Indagine sui percorsi lavorativi dei Dottori di Ricerca: aspetti di metodo e primi risultati. Paper presented at DIVAGO workshop, 27–29 Sept, University of Palermo
11. Heckman JJ (1978) Dummy endogenous variables in a simultaneous equation system. Econometrica 47(4):931–959
12. Martinelli D (1999) Labor market performance of French Ph.D.s: a statistical analysis. Céreq, Marseille
13. Mangematin V (2000) Ph.D. job market: Professional trajectories and incentives during the Ph.D. Res Policy 29:741–756
14. Nerad M, Cerny J (1999) Postdoctoral patterns, career advancement and problems. Science 285:1533–1535
15. Raghunathan TE, Lepkowski J, Van Voewyk J, Solenberger P (2001) A multivariate technique for imputing missing values using a sequence of regression models. Surv Methodol 27:85–95
16. Rosenbaum PR (1987) Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika 74:13–26
17. Scarabottolo N (2007) Alcune analisi sugli sbocchi occupazionali dei dottori di ricerca dell'Università degli Studi di Milano, Rapporto Finale, Università degli Studi di Milano – Nucleo di Valutazione dell'Ateneo
18. Stephan P, Levin S (1991) Ph.D. Supply. Issues Sci Technol 7:28–29
19. Stephan P, Everhart SS (1998) The changing rewards to science: the case of biotechnology. Small Bus Econ 10:141–51
20. Stirati A, Cesarotto S (1995) The Italian Ph.D. ten years on: educational, scientific and occupational outcomes. Higher Educ 30:37–61
21. Troxel AB, Guoguang Ma, Heitjan DF (2004) An index of local sensitivity to ignorability. Stat Sin 14:1221–1237

Chapter 17
Labour Market Performance of University Graduates: Evidence from Italy

B. d'Hombres, S. Tarantola, and D. Van Nijlen

17.1 Introduction

The massification of tertiary education, increased student mobility through the implementation of the Bologna process, the need for economic rationale behind the allocation of public funds and the demand for higher accountability and transparency have all contributed to the growing interest associated with the publication of rankings of higher education institutions. Multiple indicators, ranging from input indicators to process and outcome indicators, are combined into a single index representing overall university "excellence".[1] While, on the one hand, it could be argued that university mission is multidimensional and that it is not possible to condense the diversified work going on within universities into a single number or ranking, on the other hand, for governance purposes, there is an increasing need to be able to measure "excellence". In addition, academic rankings provide information for comparative assessments which contribute to fostering quality within the academic community. Finally, customers, students and parents can make more informed choices about where to purchase education.

There are currently two international university rankings, the *Academic ranking of World Universities* (ARWU) from Shanghai's Jiao Tong University and the *World University Ranking* (WUR) from the Times Higher Education Supplement (THES) that are widely cited in the media and used by policy makers to judge about university quality.[2] In parallel to these international university rankings, there are

B. d'Hombres (✉)
Econometrics and Applied Statistics Unit, CRELL, Institute for the Protection and Security of the Citizen European Commission, Joint Research Centre, Ispra, Italy
e-mail: beatrice.d'hombres@jrc.it

[1] See Salmi and Alenoush [11] and Usher and Savino [14] for additional information.

[2] In the *ARWU*, universities are ranked in accordance to their research performances using as criteria the number of Nobel laureates, highly cited researchers, articles published in Nature and Science, articles in Science Citation Index (*SCI*)-expanded and Social Science Citation Index (*SSCI*), and a composite indicator of academic performance normalized by the size of the institution. In the 2007 *THES* World University Ranking (*WUR*), the opinion of scientists and international employers plays a crucial role. Around 5,101 researchers and 1471 employers were asked

M. Attanasio, V. Capursi (eds.), *Statistical Methods for the Evaluation of University Systems*, Contributions to Statistics, DOI 10.1007/978-3-7908-2375-2_17,

more than 30 national university rankings that are flourishing around the world. The publication of league tables of institutional performances based on raw indicators is receiving widespread support, and this, despite several criticisms of underlying methodological flaws and limits of this ranking exercise.

In Italy, since 2000, faculty rankings are yearly published by the newspaper *La Repubblica*. These rankings are based on a wide set of performance indicators, yet none of these cover labor market performances of graduate students.

This chapter aims to investigate the determinants of unemployment and earnings of Italian university graduates. The first part of the study indicates that the family background of graduates is not significantly correlated with labour market outcomes 3 years after graduation, once we control for pre-university educational performances. To the extent that several studies report that the family environment significantly influences university enrolment and withdrawal decisions, it appears that family background-related inequalities on the Italian labor market are indirect and result from the effect of the family environment on education before the completion of a university diploma. We also observe strong effects associated with the degree that is studied and, in line with other papers, we find wide regional differences. Note, however, that we have to be careful before interpreting such correlations as causal relationships given that we might still be omitting relevant covariates in the wage and employment equations. In the second part of the chapter, we address the issue of using gross educational performance indicators for assessing university effectiveness. We derive labor market performance based-rankings of Italian faculties of economics. Our analysis underlines that (i) performance indicators need to be adjusted for students and regional's characteristics in order to be able to carry out a fair performance assessment among universities, (ii) the confidence intervals associated with each university position often overlap preventing from clearly differentiating universities' performances. Failure to make such adjustments might lead to misleading conclusions.

The structure of the chapter is as follows. Section 17.2 presents a review of the literature on the labor market outcomes of graduates. Section 17.3 describes the dataset. Section 17.4 presents the empirical results concerning the determinants of labor market performances of Italian graduates. Section 17.5 concludes.

---

to indicate the best universities. This "peer review" counts for 50% in the total score of a university. In addition, the following other criteria are used: research impact in terms of citations per faculty member, staff/student ratio, percentage of students and staff recruited internationally. These international rankings are often criticized because of the sole focus on academic research output, thus ignoring non publicized output and labor market output for the students. The applied research also disadvantages universities that are more oriented towards social sciences and humanities and universities from non-English speaking countries. See http://www.arwu.org/ranking.htm and http://www.thes.co.uk/worldrankings/ for additional information.

17.2 Review of the Literature on the Determinants of Labor Market Outcomes of Graduates

The literature on labor market outcomes of tertiary graduate students mainly consists of studies in Anglo-Saxons countries. For instance, Smith et al. [12], Smith and Naylor [13] and Bratti et al. [3] study the university effect for graduates in *UK*. All these papers illustrate that the labor market position of graduates only depends to a limited extent on the university attended. The main determinants of labor market outcomes turn out to be characteristics that are related to the individuals. Smith and Naylor [13] and Bratti et al. [3] show that it might be extremely misleading to base university rankings on raw data without taking into account students' characteristics and the uncertainty associated with the rank of each university. Bratti et al. [3] find that, for UK university graduates, university marginal effects associated with the probability of being employed or in further study, are significant, relatively to the default case, for only 28 out of 101 institutions, once we adjust for individual characteristics.

To the best of our knowledge, there are two studies on labor market performances of Italian graduate students. Brunello and Cappellari [4] investigate the determinants of earnings and employment prospects of university graduates in Italy. After showing that employment and wage effects widely vary across universities, they examine different hypothesis that could explain those observed university variations. The authors find that attending a private university (conditional on the field of study) has a substantial payoff. The pupil-teacher ratio and the number of students in the college and field of study respectively affect negatively and positively employment weighted earnings.[3] Di Pietro and Cutillo [9] study whether university quality is a significant determinant of labor market outcomes of Italian graduates (measured by earnings and over education). They proxy university quality with the performance indicators published by *La Repubblica*. The results show that individuals graduated from a research oriented university experience better labor market outcomes than those who studied in less research oriented institutions.

The focus in the current analysis differs. As in the two mentioned studies, we start the empirical exercise with an analysis of the determinants of labor market performances but we rely on a more recent survey. In addition we discuss in detail the limits associated with university rankings based on gross indicators. To that end, we derive and compare unadjusted and adjusted labour market performance based-rankings of Italian faculties of economics. We show how important it is to adjust for differences in students' intake characteristics as well as to take into account the uncertainty associated with the position of each institution.

---

[3] Following Card and Krueger [6], Brunello and Cappellari [4] have employed a two-step approach. They have first estimated a wage equation and an employment equation. Both equations are function of individual and regional covariates and of university per field of study dummies. Then in a second step the estimated coefficients associated the university per faculty effects are regressed on field of study dummies, university dummies as well as some measures of university quality.

17.3 Labor Market Determinants of Italian Graduate Students

17.3.1 Data

The data stem from a survey conducted by the Italian National Institute of Statistics (*ISTAT*) in 2004 on *Laurea* holders, i.e., students who graduated in 2001 from university with a *Laurea*. About 26, 006 individuals were interviewed (Computer assisted telephone interview, $C.A.T.I$), which represents about 16% of the cohort of 2001 graduate students. The response rate reached 67.6%. The focus of the survey lies on the labour market experiences of the respondents during the 3 years following their graduation from university.

The graduate students covered by this survey completed a *Laurea* under the "old university system", i.e., before the implementation, in the academic year 2001/2002, of the set of university reforms implied by the Bologna process. Under this period, the duration of studies to complete a *Laurea* varied according to the field: 4 years in scientific and humanistic disciplines, 5 years in engineering.[4]

In the following analysis, an individual is regarded as unemployed if she is not working and simultaneously looking for a job at the time of the interview. Individuals who are not "unemployed" are either inactive (a large portion of them are studying/doing internships) or working at the time of the interview. The individual hourly wage is net of taxes. Seasonal and occasional workers are excluded from the analysis since we do not have information on the salary for this group of workers.[5]

In addition to the information on the working status and the wage of the respondents, the survey contains extensive additional information. The variables that will be used in the empirical analysis can be regrouped into four subsets: (i) individual socio-demographic characteristics, (ii) pre-university education-related variables, (iii) university-related variables and (iv) job characteristics. The first set of variables includes information on the individual's socio-demographic characteristics such as gender, marital status, region of residence, parental education and parental occupation when the respondent was around 14 years old. The second set is made of variables linked to the pre-university educational background of the respondents such as high school grade and type of high school attended. The third set of variables refers to the academic performance of the respondents and includes variables on the type of degree and university attended, grade obtained for the Laurea, whether the

[4] Under the old university system, the university in Italy was mainly based on one level. The university offered one qualification, the *Laurea*. By the end of 1980s, the demand for shorter studies led to the introduction of university diplomas (Diplomi Universitari, DU) whose duration was only 3 years. Students were thus offered the possibility, once graduated from high school, to choose between diploma courses and degree courses. However, the proportion of students enrolled in diploma courses was low (below 8% of students) mainly because university diplomas were offered in a limited number of fields and were hardly recognized when students wanted to continue with a degree course.

[5] The relevant question is: "Is your current job seasonal, occasional, or continuous?" The number of seasonal or occasional workers amounts to 1,933.

respondent graduated "cum laude" and the number of years spent for the completion of the degree. The fourth set of variables includes job-related information such as the number of years of experience in the current job, the type of job and the duration of the contract.

Table 17.1 reports the accurate definition of the variables employed in the empirical exercise while Table 17.2 displays for the year 2004 raw figures on average hourly wage and employment probability by gender, region and field of study. The summary statistics reported in column 1 of Table 17.2 as well as the subsequent empirical analysis on the determinants of employment probabilities are based on a sample of 23,511 individuals of which 10.45% are unemployed.[6] Similarly the sample of wage earners used in the rest of the paper for analysing the determinants of wages amounts to 9,641 observations. The mean hourly wage reaches 8.21 euros for the studied group.[7]

As shown in Table 17.2, we observe significant variations across regions and gender. Males are 50% less likely to be unemployed than females. These gender differences are likely to be in part the result of gender-specific labor supplies. The unemployment rate ranges between 5.01% (North-West) and 20.24% (South) according to the region of residence. Hourly wage differences are of lower magnitude. Males earn on average 5% more than females. Hourly wages are not significantly different across regions.

We also observe significant differences by field of study/faculty: around 2.9% of graduate students in medicine are unemployed while this figure reaches 20.8% for graduates in law. For these two degree subject groupings, the average hourly wage is respectively 11.3 and 7.7 euro. Variations between faculties are noticeable. For instance, while graduates in engineering and literature experience a similar median hourly wage (see Fig. 17.1), hourly wage variations are much higher for graduates in literature than for graduates in engineering. Similarly, employment probabilities for graduates in economics and statistics vary a lot while the lowest variance is observed for graduates in medicine.

---

[6] Out of the 26,006 students who participated in the survey, 2,495 of them were removed from the analysis because of missing information on some variables of interest. Around 17% of excluded graduates do not declare their labour-situation while respectively 34 and 36% of excluded graduates do not report information on the type of degree or university attended and their family background.

[7] A large part of respondents was excluded because of the fact that they were either unemployed at the time of the survey, pursing studies or only had a seasonal or occasional job. For these 3 groups no data on the salary is available. They represent approximately 65% of the graduates removed from this sample. Graduates in Medicine or Laws are significantly more likely to pursue their studies and those living in the North-East of Italy have a lower propensity to have a seasonal or occasional job. We also observe that a significant number of respondents indicate that they perform paid working hours, but do not report a monthly salary (about 10% of excluded observations).

Table 17.1 Definition of the variables

Job characteristics	
Unemployed	Indicator taking on the value 1 if the individual is unemployed at the time of the interview, zero otherwise
Long term contract	Indicator taking on the value 1 if the individual has a long term working contract, 0 if it is a short term contract
Full-time	Indicator taking on the value 1 if the individual is employed full-time, 0 if he is working on part-time
Net hourly wage	Net hourly wages measured by the net monthly wage divided by the number of hours worked during the month
Specific experience	Number of months spent in the current job
Educational background: university related information	
Laurea: more years than expected	Indicator taking on the value 1 if the individual has taken more time than the legal basis to get a university degree, 0 otherwise
Laurea: at lest 2 years more	Indicator taking on the value 1 if the individual has taken at least 2 years more than expected to get a university degree, 0 otherwise
Score: *Laurea*	*Laurea* graduating marks (1–4 point scale, 1 = score below 80, 1 = score between 80 and 89, 2 = score between 90 and 94, 1 = score between 95 and 99)
Cum lode	Indicator taking on the value 1 if the individual got a *Laurea* cum lode, 0 otherwise
Field of study dummies	14 fields of study (Sciences, Chemistry/Pharmacy, Geo-Biology, Medicine, Engineering, Architecture, Agrarian, Economics/Statistics, Political Sciences, Law, Literature, Linguistic, Teaching, Psychology)
University dummies	67 universities

Table 17.1 (continued)

Educational background: pre-university related information	
Technical school	Indicator taking on the value 1 if the individual is graduated from a technical secondary school, 0 otherwise
Vocational school	Indicator taking on the value 1 if the individual is graduated from a vocational secondary school, 0 otherwise
General school	Indicator taking on the value 1 if the individual is graduated from a general secondary school, 0 otherwise
Score: school leaving examination	Higher secondary school diploma marks (36 = lowest score, 60 = highest score)
Students' characteristics	
Sex	Indicator taking on the value 1 if the individual is a male, 0 otherwise
Marital status	Indicator taking on the value 1 if the individual is married or living in couple, 0 otherwise
Regional dummies	
Family background (when the student was 14 years old)	
Father occupied	Indicator taking on the value 1 if the father was working, 0 otherwise
Father education: tertiary	Indicator taking on the value 1 if the father has a university degree, 0 otherwise
Mother education: tertiary	Indicator taking on the value 1 if the mother has a university degree, 0 otherwise
House wife	Indicator taking on the value 1 if the mother was an housewife, 0 otherwise

Data: Percorsi di lavoro e di studio dei Laureati, 2004.

Table 17.2 Summary statistics

	Full sample % of unemployed respondents	Sample: wage earners Hourly wage in euros
	Sample size: 23 511	Sample size: 9 641
Full sample	10.45	8.21
Females	13.03	8.00
Males	7.66	8.41
North-East	6.06	8.09
North-West	5.01	8.35
South	20.24	8.08
Centre	9.10	8.11
Isles	18.22	8.08
Sciences	11.47	9.20
Chemistry-Pharmacy	7.61	8.33
Geo-Biology	16.51	8.33
Medicine	2.91	11.35
Engineer	5.09	8.41
Architecture	9.58	7.41
Agrarian	13.30	7.51
Economics-Statistics	11.04	7.79
Political/Sciences	10.17	7.57
Law	20.83	7.69
Literature	17.55	8.44
Linguistic	16.42	8.09
Teaching	11.19	7.66
Psychology	13.84	8.35

Data: Percorsi di lavoro e di studio dei *Laureati*, 2004.

17.3.2 Empirical Methodology

In order to investigate the labor market determinants, we first model the probability of being unemployed 3 years after graduation as follows:

$$P_i = \alpha_1 + X_i\beta_1 + \sum_f F_i^f \theta_1^f + \sum_u U_i^u \phi_1^u + \varepsilon_i \tag{1}$$

The dependent variable P_i in Eq. (1) is a variable taking on the value 1 if the individual $i = 1 \ldots N$ is unemployed, and 0 otherwise. The set of control variables X_i includes information related to personal characteristics (sex, marital status), past educational experiences (pre-university and university performances, high school type), family background (education and occupational status of students' parents) of the respondent and macro region of residence dummies. The regions are regrouped into 5 macro regions. F_i^f is a set of degree subject (or faculty) dummies with $f = 1 \ldots F$ and U_i^u is a set of university dummies with $u = 1 \ldots U$. Graduates come

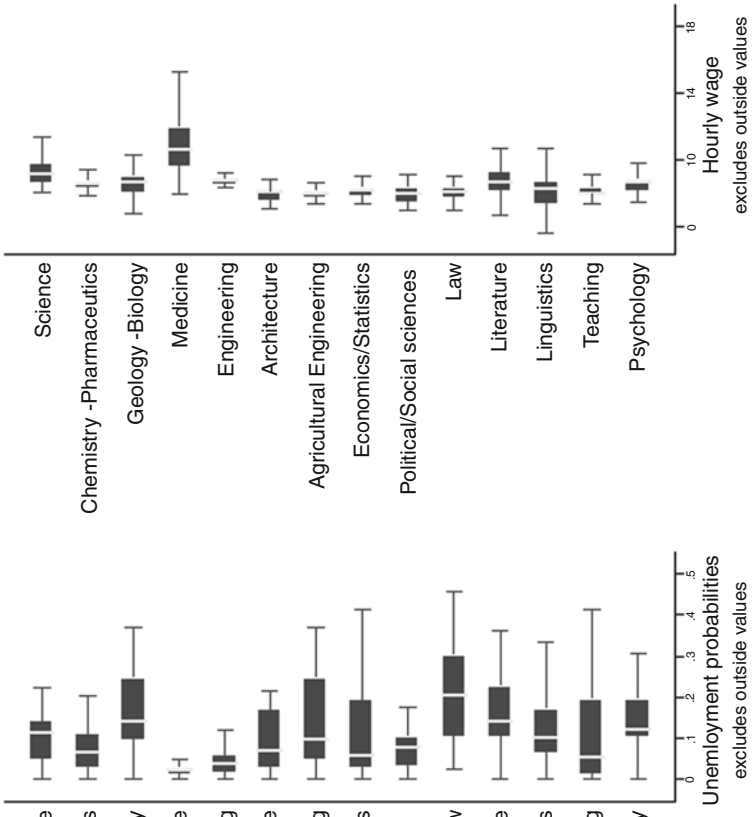


Fig. 17.1 Unemployment probabilities and hourly wages by field of study: within faculty variations. Note: The boxes are bordered at the 25th and 75th percentiles of the x-variable (unemployment probability and hourly wage) with a median line at the 50th percentile. Whiskers extend from the box to the upper and lower adjacent values. Adjacent values are calculated utilizing the interquartile range (IQR) which is the difference between the first and third quartile values. The upper adjacent value is the largest data value that is less than or equal to the third quartile plus 1.5 IQR and the lower adjacent value is the smallest data value that is greater than or equal to the first quartile minus 1.5 IQR

from 14 different faculties and 67 universities In total, respondents are grouped in 484 faculty/university combinations. Finally, ε_i is the disturbance term. Equation (1) will be estimated with both a linear probability model and a logit model.

Then, we estimate the following wage equation:

$$W_i = \alpha_2 + X_i\beta_2 + \sum_f F_i^f \theta_2^f + \sum_u U_i^u \phi_2^u + \xi_i \tag{2}$$

where W_i represents the logarithm of the hourly wage defined as the monthly wage divided by the number of monthly worked hours. In addition to the covariates included in Eq. (1), we add a set of control variables related to job characteristics, such as specific labor market experience and the contract type.

The main focus of the chapter is on the set of estimated coefficients $\widehat{\phi}_1^u$ and $\widehat{\phi}_2^u$. Conditioned on the other control variables included in Eqs. (1) and (2), we interpret $\widehat{\phi}_1^u$ and $\widehat{\phi}_2^u$ as the marginal contribution of each university to graduates' labor market performances.

However, for this interpretation to be correct

$$E(Z_i, \varepsilon_i) = 0 \text{ and } E(Z_i, \xi_i) = 0$$

must be satisfied, with $Z_i = \left\{X_i, F_i^f, U_i^u\right\}$. University differences can be interpreted as a measure of the relative effectiveness of each university if and only if students are randomly assigned to universities once we control for the set of covariates (X_i, F_i^f).

This will not be the case if we fail to control for unobservable variables that are both correlated with university choice and labour market outcomes. In particular, we need to assume that the family background variables and the pre-university education-related variables included in Eqs. (1) and (2) are enough to capture the individual innate ability and well as the individual motivation, two components of individual heterogeneity that are likely to affect the choice of the university as well as the performance on the labour market.[8] In addition we introduce macro regional dummies in Eqs. (1) and (2) to control for differences across regions in labour market conditions, assuming that labour market conditions within macro region are homogenous. In other words, while we do include in Eqs. (1) and (2) a large set of covariates, we might still be omitting relevant variables that could bias estimates of the effect of the university attended.[9]

[8] Moreover while we include in the wage and employment equations the parental education and parental occupational status, we do not have information on the family income. Income constraints could have prevented the respondent from enrolling in a university of her choice but far from her parents' home.

[9] Note that there is also the risk to include covariates that capture the marginal contribution of the faculty. We check the sensitivity of our results to the choice of control variables by estimating more parsimonious specifications with only prior-entry university related variables.

Tables 17.3 and 17.4 report the results of the empirical analysis. Table 17.3 displays the estimated coefficients associated with the variables related to personal characteristics and educational background while Table 17.4 reports those on degree subject and regional location dummies. Columns 1−3 present estimates of Eq. (1). In columns 1 and 2, we rely on a linear probability model while in column 3 we display results obtained with a logit model. We report the marginal effects at the average values of the independent variables in the third column. Note that in column 1, we do not include the covariates related to the students' academic performance (*Laurea* graduating mark, etc). The last two columns of Table 17.3 present linear estimates of Eq. (2). In column 4, the coefficients associated with Z_{ijkr} and academic performances are constrained to be equal to zero. This assumption is relaxed in column 5.

17.3.3 Empirical Results

We discuss the determinants of unemployment and earnings in the next section. We mainly base the discussion on results displayed in columns 2 and 5 of Table 17.3 and Table 17.4 .[10]

Respondents' characteristics: Gender differences are observed on the labor market. Males are around 4.0% more likely to be employed and are paid 6.4% more than females. The marital status also affects labor market performances: the wage premium for being married or living in couple amounts to 3.4% and the unemployment probability significantly decreases by 2.1%. Regarding the wage equation, part of the effect might be explained by differences in tax levels by marital status and household size (recall the wage is net of taxes).

Educational background: Estimates indicate that pre-university qualifications do not affect labor market performances. Indeed the type of high school (technical, vocational or general high school) as well as the high school leaving examination score is not significantly different from zero in both Eqs. (1) and (2) with unemployment probability and the logarithm of hourly pay.[11] This result, similar to the one obtained by Boero et al. [2], is due to the high correlation between pre-university and university performances. Indeed, when we omit to include university performance related variables (i.e., column 1 of Table 17.3), the higher school graduating score is significant and respectively negatively and positively correlated with unemployment probabilities and hourly wages.

The performance at university impacts differently on employment and earnings. As Boero et al. [2], we do not observe any effect of the *Laurea* graduating mark on labor market outcomes. However, when we do not include *Laurea* subject dummies

[10] Note that the results are similar with cluster standard errors at the university level.

[11] Note that the dummy variable "other school", is positive and significant in the wage equation. However, this result is only due to the fact that most of individuals for which this dummy variable takes the value one are employed teachers for who the number of working hours per week does not exceed 18.

Table 17.3 Determinants of labor market performances

	Unemployment equation			Wage equation	
	Linear specification		Logit	Linear specification	
Pre-university educational background					
Score: school leaving examination	–0.0008 (−2.90)	–0.0004 (−1.48)	–0.00039 (−1.59)	0.001 (3.35)	0.0003 (0.63)
Secondary school type	Excluded category: vocational school				
General school	–0.007 (−0.64)	–0.004 (−0.39)	–0.0021 (−0.23)	0.020 (1.11)	0.011 (0.62)
Technical school	–0.012 (−1.01)	–0.010 (−0.90)	–0.0061 (−0.68)	0.010 (0.55)	0.007 (0.40)
Other school	–0.023 (−1.67)	–0.023 (−1.66)	–0.014 (−1.62)	0.11 (4.58)	0.10 (4.47)
Academic performances					
Score: *Laurea*		–0.0034 (−1.22)	–0.0018 (−0.85)		0.002 (0.53)
Cum Lode		–0.0021 (−0.78)	–0.0013 (−0.62)		0.023 (5.22)
Laurea: more years than expected		0.011 (2.02)	0.011 (2.61)		–0.016 (−1.95)
Laurea: at least 2 years more		0.012 (2.34)	0.0087 (2.20)		0.0031 (0.39)
Personal characteristics					
Sex	0.040 (9.37)	0.040 (9.51)	0.032 (9.36)	–0.063 (−9.32)	–0.064 (−9.57)
Marital status	–0.020 (−4.54)	–0.021 (−4.84)	–0.018 (−5.56)	0.032 (4.63)	0.034 (4.88)
Family background					
Mother education: tertiary	–0.0068 (−1.34)	–0.0063 (−1.24)	–0.0048 (−1.15)	0.0003 (0.048)	0.0002 (0.026)
Father education: tertiary	–0.0040 (−0.81)	–0.0038 (−0.77)	–0.0026 (−0.67)	0.010 (1.40)	0.010 (1.43)
Father occupied	0.0028 (0.28)	0.0026 (0.26)	0.0018 (0.25)	0.007 (0.45)	0.007 (0.48)
Housewife	–0.00033 (−0.079)	–0.00013 (−0.031)	0.00078 (0.23)	0.003 (0.41)	0.002 (0.29)
Job characteristics					
Long term working contract					–0.022 (−3.31)
Specific experience					0.0011 (4.90)
University/faculty dummies	*YES*	*YES*	*YES*	*YES*	*YES*
Regional dummies	*YES*	*YES*	*YES*	*YES*	*YES*
F test: joint signif of the Uni. effects (p-value)	0.00	0.00	0.00	0.00	0.00
Observations	23,511	23,511		9,641	9,641

Table 17.4 Determinants of labor market performances

	Unemployment equation			Wage equation	
	Linear specification		Logit		
Field of study	Excluded category: Law				
Chemistry	–0.12	–0.11	–0.049	0.10	0.099
	(–11.2)	(–10.6)	(–12.7)	(6.31)	(5.89)
Biology	–0.041	–0.035	–0.015	0.084	0.068
	(–3.77)	(–3.11)	(–2.43)	(4.57)	(3.61)
Medicine	–0.17	–0.16	–0.087	0.31	0.29
	(–20.4)	(–18.4)	(–26.9)	(13.9)	(12.7)
Engineering	–0.12	–0.12	–0.060	0.086	0.086
	(–13.9)	(–13.6)	(–16.4)	(5.60)	(5.58)
Architecture	–0.098	–0.095	–0.038	–0.043	–0.056
	(–7.73)	(–7.29)	(–6.95)	(–1.75)	(–2.25)
Agrarian	–0.066	–0.059	–0.023	–0.008	–0.021
	(–5.47)	(–4.84)	(–3.68)	(–0.38)	(–0.99)
Economics-Statistics	–0.083	–0.082	–0.036	0.023	0.021
			(–8.99)	(1.55)	(1.43)
Literature	–0.029	–0.024	–0.0098	0.062	0.040
	(–2.79)	(–2.21)	(–1.60)	(3.42)	(2.14)
Linguistic	–0.043	–0.040	–0.017	0.034	0.021
	(–3.44)	(–3.09)	(–2.46)	(1.68)	(1.04)
Teaching	–0.085	–0.077	–0.033	0.0008	–0.022
	(–6.75)	(–5.97)	(–5.80)	(0.043)	(–1.03)
Psychology	–0.039	–0.032	–0.0071	0.078	0.063
	(–2.59)	(–2.12)	(–0.74)	(3.16)	(2.56)
Sciences	–0.072	–0.070	–0.030	0.15	0.14
	(–6.66)	(–6.41)	(–5.85)	(8.27)	(7.94)
Political–Sciences	–0.086	–0.081	–0.034	–0.004	–0.013
	(–7.70)	(–7.21)	(–6.85)	(–0.20)	(–0.69)
Regional dummies	Excluded category: NorthEast				
North-West	–0.025	–0.025	–0.023	0.032	0.033
	(–2.52)	(–2.54)	(–2.69)	(2.27)	(2.34)
Centre	0.023	0.023	0.021	–0.030	–0.031
	(2.16)	(2.15)	(2.03)	(–1.90)	(–1.94)
South	0.100	0.099	0.079	–0.057	–0.055
	(9.09)	(8.97)	(5.48)	(–3.39)	(–3.28)
Isle	0.068	0.068	0.060	–0.014	–0.014
	(4.42)	(4.39)	(3.23)	(–0.57)	(–0.56)
University dummies	*YES*	*YES*	*YES*	*YES*	*YES*
Observations	23,511	23,511	23,511	9,641	9,641

and university dummies (the results are not reported but available upon request), the coefficient associated with the *Laurea* score becomes significant and respectively positive and negative in the unemployment and wage equations. This result is in line with Bagues et al. (2008). Indeed, the authors show the existence of variations in grading standards across faculties that are driven by faculty variations in students'

enrollment more than by teaching quality differences. The graduating mark variable is thus very collinear with faculties dummies and does not reflect (or only partially) variations in students' quality. Finally, individuals having taken more time than the legal duration to complete the *Laurea* are 1.1% more likely to be unemployed while it does not affect hourly pay. Also cum laude *Laurea* graduates experience a significant wage premium.

Family background: None of the family background variables have a significant effect on the labor market outcomes. Indeed, whether we control or not for students' academic performance, having parents with a university degree has no impact on labor market success. Similarly, the occupational status of the father (employed versus unemployed) when the student was 14 years old is found uncorrelated with hourly wages and employment probabilities. We note however that several papers [5, 7, 8] indicate that the family environment has a significant influence on the decision of Italian secondary school graduates to enrol at university and dropout of university. In addition, although results are not reported here, we find that the education of the father has a significant effect on labour market outcomes only when we do not include in the specification variables related to students' pre-university educational performance (high school grade and type of high school attended). This suggests that the family environment makes a difference before graduating from university.

Employment characteristics: With respect to employment characteristics, specific labor market experience is positively and significantly correlated with hourly wages while having a long term working contract is negatively correlated with hourly wages. The inclusion of these variables does not affect the value and significance of the other covariates. Short terms contracts are compensated by higher hourly wages.

Degree subject: Table 17.4 shows the estimated coefficients associated with degree subject dummies. The reference group is law. Based on results reported in column 5, 9 out of 13 degree field dummies are significantly different from zero at the 5% level whereas all field dummies are significant in the unemployment equation (column 2). We observe strong effects associated with the degree that is studied. Graduates in medicine and engineering are, for instance, 18 and 12% less likely to be unemployed than graduates in law. Simultaneously, the hourly wage for those two categories of graduates is 29% and 8.6% higher than the hourly wage by graduates in law. On the other hand, graduates in architecture are 5.6% less paid than graduates in Law. These results highlight differences in returns to education that persist across subjects.

Regional differences: Like Brunello and Cappellari [4] or Bagues et al. [1], we observe important regional differences. Job opportunities and wage levels are significantly lower in the South than in the North-East of Italy. Similarly, unemployment probability is 2.5% lower and hourly wages are 3.3% higher in the North-West than in the North-East. Despite these regional differences, students' mobility remains low in Italy. Brunello and Cappellari [4] test and reject the hypothesis that this might be due to liquidity constraints. Note that if we include 20 regional dummies instead of

5 broad regions of residence, the coefficients associated with the other covariates are not significantly modified. However, this analysis is statistically costly because, the identification of $\widehat{\phi}_{k1}$ and $\widehat{\phi}_{k_2}$ only relies on individuals (i) having graduated in different universities but yet located in the same regions or (ii) who moved to another region after graduation. On the other hand, in the most parsimonious specification reported in Table 17.4, we are assuming that the labor market conditions are homogeneous within each broad region of residence.

University fixed effects: If we include university fixed effects instead of university dummies (which gives equivalent results), while conditioning for graduates' characteristics, courses and regional related variables, the fraction of the variation of W_{ijkr} and P_{ijkr} due to inter-university variations is very low and respectively equal to 0.025 and 0.019 for the unemployment and wage equation However, the $F-test$ reported in the bottom of Table 17.4 implies that there are significant joint faculty effects and as a consequence a faculty random effects model would be inappropriate.

17.4 Faculty-Performance Indicators: The Case of Economics

University rankings are not immune from methodological flaws. Goldstein and Spiegelhalter [10] extensively discuss the methodological problems associated with the publication of league tables and suggest a set of principles that should be fulfilled before publishing HEIs rankings. First, results should always be contextualized. In other words, fair comparisons can only be made if appropriate adjustments are made for external contextual factors. Second, the uncertainty of the results should be displayed. Third, results should be presented for multiple indicators to avoid concentrations on only one aspect of performance.

In the next section, we aim at illustrating the relevance of the first two concerns. To that end, we focus on the faculties of economics and derive "adjusted" rankings based on the estimation of Eqs. (1) and (2). We show that the relative gross performance of each university is in large part the product of contextual factors. In addition, the uncertainty associated with the rank of each institution does not allow for a clear differentiation among them.

17.4.1 Unadjusted Versus Adjusted Ranking Based on Labor Market Outcomes

The first point stressed by Goldstein and Spiegelhalter [10] is the need to adjust league tables of HEIs for the quality of intake students and the local context. Indeed, the underlying assumption behind most of university rankings using output indicators is that HEIs are the only one responsible for these outcomes. However, in our case study, the determinants of labor market outcomes for Italian tertiary graduates

are numerous. In particular, we have found that the pre-university educational backgrounds and the region of residence of graduates are significantly associated with their performance on the labor market.

In Table 17.5, we compare the adjusted and unadjusted labour market-based university rankings, and this only for the faculty of economics. Columns 1 and 3 display rankings based on raw data while those reported in columns 2 and 4 are adjusted for the other covariates included in Eqs. (1) and (2). These two rankings were obtained from the estimation of Eqs. (1) and (2) on the restricted sample of individuals having graduated in economics. Graduates in economics are distributed over 51 universities. For the "unadjusted" ranking, we constrain the coefficients associated with the covariates – but the university dummies – in Eqs. (1) and (2) to be equal to zero while we relax this assumption for the "adjusted" ranking.[12] However, in order to be sure that some of the explanatory variables in Eqs. (1) and (2) are not already capturing university fixed effects, we only control for prior-entry university related variables.

Figures 17.2 and 17.3 compare the "adjusted ranking" with the "unadjusted" one. The correlation between the adjusted hourly wage level-based ranking and unadjusted hourly wage level-based ranking is equal to 0.95 while for the employment probability-based rankings, this correlation reaches 0.96. These high correlations hide significant variations. For instance, the faculty of economics at the university of Bergamo ranks sixth with the unadjusted hourly wage level-based ranking but drops by twelve positions in the adjusted ranking. Similarly, the faculty of economics at Milano Bicocca ranks ninth with the unadjusted hourly wage based ranking, while it drops to the twenty-fifth position with the adjusted ranking. On the other hand, the faculty of economics in Bari is not part of the top 20 with the unadjusted hourly wage based ranking but jumps to the eighth position once we control for students characteristics and the regional context. This last result suggests that (i) the quality of students who enroll at the faculty of Bari is on average lower than in other faculties of economics and/or (ii) labor market characteristics in Puglia are much more unfavorable than in other parts of Italy. In Fig. 17.3, the faculty of economics in Sassari ranks thirtieth according to the unadjusted ranking based on unemployment probability, but it is in the top three Universities when we control for student characteristics and regional context. In general, we note that the regional context plays an important role: the Universities located in northern Italy, where labour market conditions are more favourable, are in a better position with the unadjusted ranking than with the adjusted one. The opposite holds for those located in Southern Italy. These results show that much care is needed before drawing any conclusions from unadjusted rankings. We observe, however, that some universities maintain the same position with the hourly wage-based unadjusted and adjusted rankings: this is the

[12] The faculty dummies are not included in the two specifications given that in the following analysis we only cover graduates in economics.

Table 17.5 Faculties of economics: rankings based on labour market outcomes

	Unemployment based ranking		Logarithm of hourly wage based ranking	
Faculties	Unadjusted ranking	Adjusted ranking	Unadjusted ranking	Adjusted ranking
Ancona	20	11	27	24
Aquila	31	16	5	1
Arcavacata	40	39	40	31
Bari	46	46	22	8
Benevento	41	42	45	43
Bergamo	13	24	6	18
Bologna	1	5	33	35
Brescia	5	21	29	41
Cagliari	33	23	10	6
Campobasso	35	32	47	47
Cassino	36	41	42	38
Catania	47	47	13	23
Chieti	38	34	43	26
Ferrara	2	2	44	46
Firenze	9	6	8	9
Foggia	37	35	14	7
Genova	17	29	25	40
Lecce	44	45	46	44
Macerata	25	18	32	27
Messina	43	40	37	30
Milano Bicocca	7	20	9	25
Modena	11	13	28	34
Napoli	39	36	4	2
Napoli II	42	43	34	22
Napoli Parthenope	34	33	23	11
Padova	26	30	31	36
Palermo	29	10	41	42
Parma	19	19	24	32
Pavia	12	25	12	28
Perugia	8	4	26	29
Pisa	22	17	39	39
RomaIII	16	12	3	4
Roma Sapienza	15	8	20	21
Roma Tor Vergata	28	31	18	14
Salerno	45	44	35	13
Sassari	30	3	30	20
Siena	23	15	11	10
Torino	14	26	38	45
Trento	10	14	7	12
Trieste	21	28	2	5
Udine	18	22	16	19
Urbino	3	1	15	16
Varese	4	9	19	33
Venezia	24	27	17	17
Vercelli	27	38	1	3
Verona	6	7	36	37
Viterbo	32	37	21	15

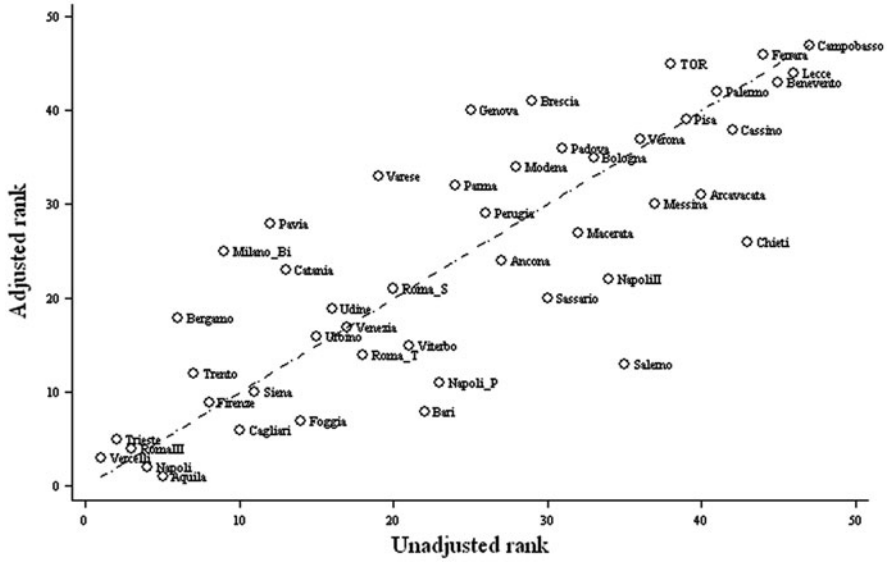


Fig. 17.2 Adjusted *versus* unadjusted ranking based on wage levels

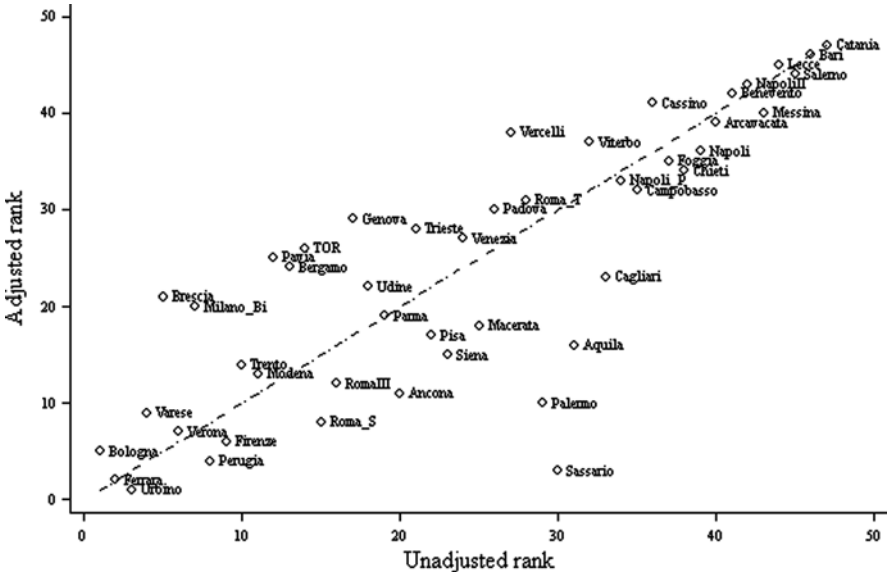


Fig. 17.3 Adjusted *versus* unadjusted ranking based on unemployment probabilities

case, for example, of the faculties of economics of Ferrara, Firenze and Palermo. We observe similar patterns for the ranking based on unemployment probabilities.

17.4.2 Uncertainty Associated with Rankings

The second concern stressed by Goldstein and Spiegelhalter [10] is related to the fact that published league tables never present estimates of the statistical uncertainty associated with the rankings. In other words, one could ask whether there is a real difference in performance between universities that rank 9th and 10th, or whether the uncertainty associated to such ranks makes them undistinguishable in practical terms.

In Figs. 17.4 and 17.5 the estimated position of each faculty based on the hourly wage is displayed with the associated confidence intervals obtained with the estimation process. We note that these intervals overlap quite broadly, therefore the difference in University performance is not significant. Nonetheless, on the one hand, when we do not control for the contextual factors, it is possible to find intervals that do not overlap, concluding that the faculties have different performances. On the other hand, once we control for the contextual factors (see Fig. 17.5), we observe a further increase of the confidence bounds, meaning that the adjustment for student characteristics and regional differences makes the situation more homogeneous across Universities. Similar behaviour occurs in Figs. 17.6 and 17.7 where the ranking is made with respect to the unemployment probability.

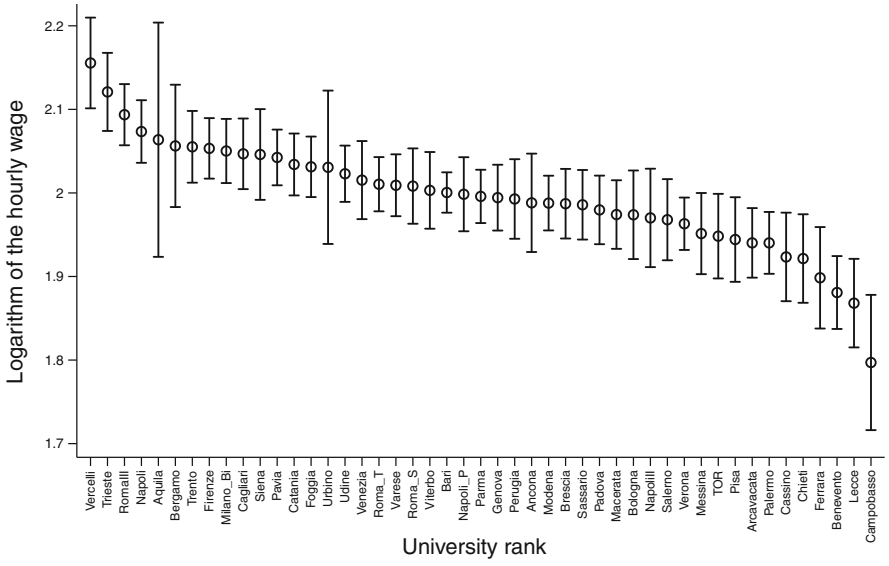


Fig. 17.4 Logarithm of the hourly wage based ranking and confidence bounds: "unadjusted ranking"

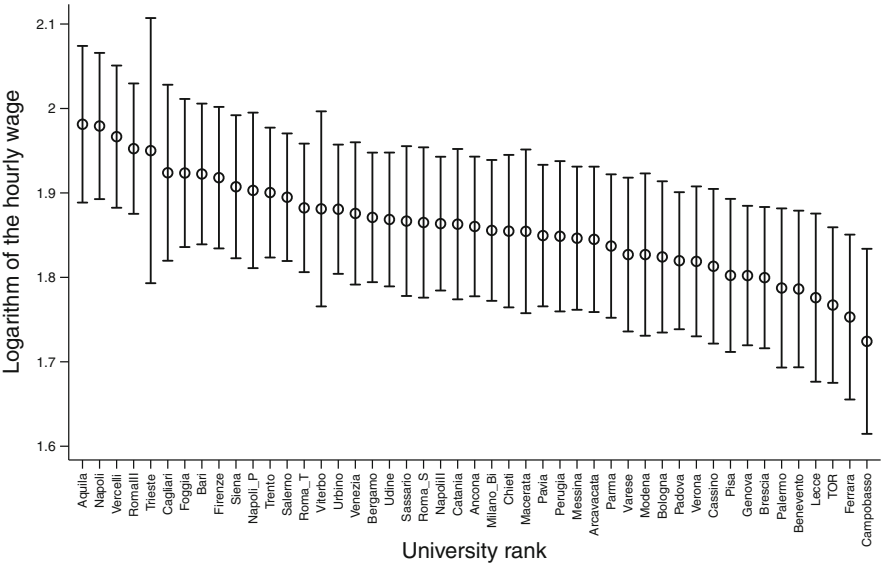


Fig. 17.5 Logarithm of the hourly wage based ranking and confidence bounds: "adjusted ranking"

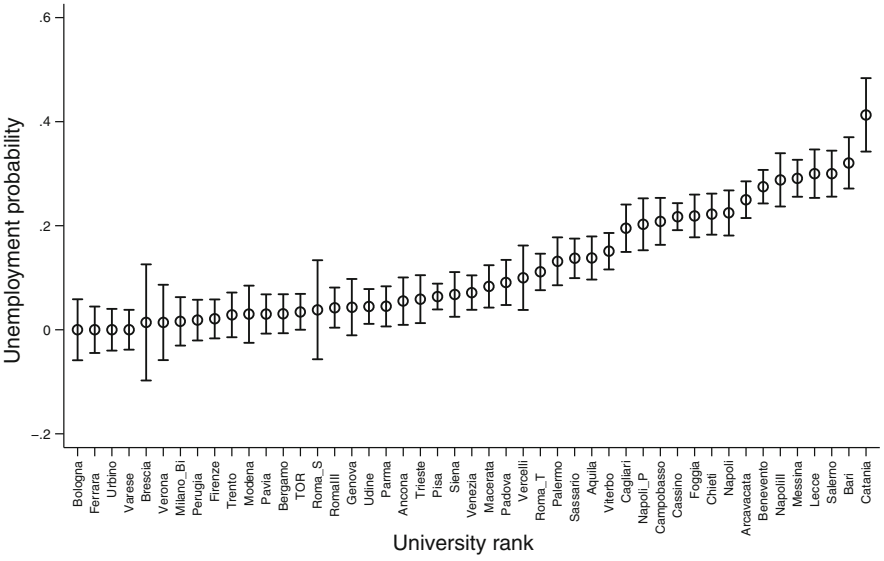


Fig. 17.6 Unemployment probability based ranking and confidence bounds: "unadjusted ranking"

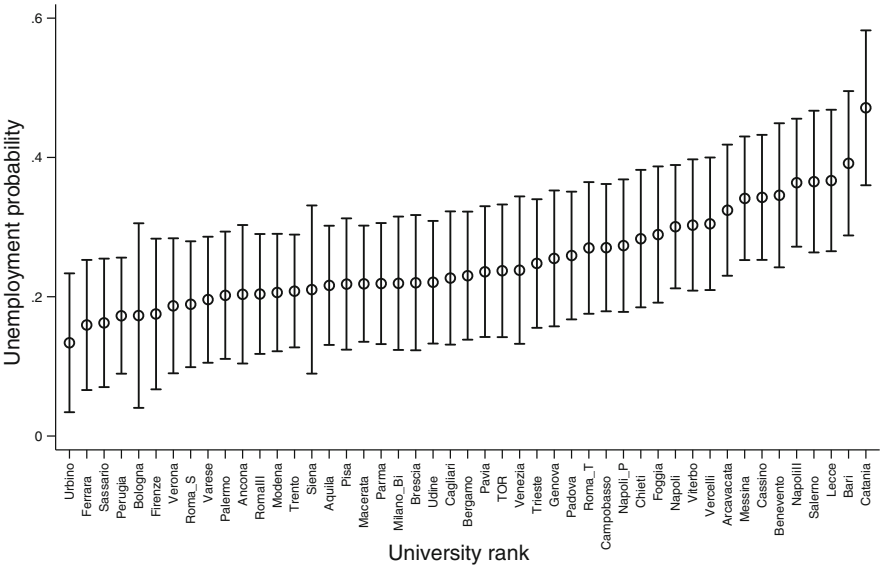


Fig. 17.7 Unemployment probability based ranking and confidence bounds: "adjusted ranking"

17.5 Conclusion

In this chapter, we investigate the determinants of labor market performances of Italian university graduates. Our results show in particular that the family background of graduate is not significantly correlated with hourly wages and employment probabilities 3 years after graduation, once we control, in particular, for pre-university educational performances. It seems that the family environment have some bearing on the labour market essentially through its indirect effects on educational decisions before completion of a university diploma. We also observe strong effects associated with the degree that is studied. Graduates in hard sciences experience, *ceteris paribus*, lower unemployment probabilities and higher hourly wages than graduates in soft sciences. Finally, in line with other studies, we find wide regional differences.

Focusing on the faculty of economics, we then examine in more detail the impact of the faculty attended on employment probabilities and wage levels. Our empirical results illustrate the relevance of Goldstein and Spiegelhalter's [10] principles that should be applied when one presents performance indicators. First, results should always be contextualized. In other words, fair comparisons can only be made if appropriate adjustments are made for external contextual factors. Second, the uncertainty of the results should always be displayed.

We observed that the adjustments for external contextual factors have the effect to increase the confidence bounds for the universities' ranks, making their performances more homogeneous in terms of labour market outcomes. Unadjusted results may cause dangerous inferences about the relative performance of each institution,

especially when they are used for strategic purposes. As Goldstein and Spiegelhalter [10] state "we can use rankings as screening instruments, but not as definitive judgments on individual institutions".

Acknowledgement The authors would like to thank Professor Tuzzi for her review as well as the participants to the 2008 DIVAGO final workshop in Palermo for the useful comments.

References

1. Bagues M, Labini MS, Zinovyeva N (2008) Differential grading standards and university funding: evidence from Italy. Working papers 2008-07, FEDEA
2. Boero B, McKnight A, Naylor R, Smith J (2001) Graduates and graduate labour markets in the UK and Italy. Working paper CRENoS 200111, Centre for North South Economic Research, University of Cagliari and Sassari, Sardinia
3. Bratti M, McKnight A, Naylor R, Smith J (2004) Higher education outcomes, graduate employment and university performance indicators. J R Stat Soc Ser A Stat Soc 167(3):475–496
4. Brunello G, Cappellari L (2008) The labour market effects of alma mater. evidence from Italy. Econ Educ Rev 27:564–574
5. Cappellari L, Lucifora C (2008) The "Bologna Process" and college enrolment decisions. IZA discussion papers 3444, Institute for the Study of Labor (IZA)
6. Card D, Krueger AB (1992) Does school quality matter? Returns to Education and the characteristics of public schools in the United States. J Polit Econ 100(1):1–40, University of Chicago Press
7. Cingano F, Cipollone P (2007) University drop-out. The case of Italy. Economic working papers 626, Bank of Italy, Economic Research Department
8. d'Hombres (2007) The impact of university reforms on dropout rates and students' status: evidence from Italy. JRC Scientific and Technical reports 22733
9. Di Pietro G, Cutillo A (2006) University quality and labour market outcomes in Italy. Lab Rev Lab Econ Ind Relat 20(1):37–62
10. Goldstein H, Spiegelhalter D (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc 159(3):385–443
11. Salmi J, Alenoush S (2007) League tables as policy instruments: uses and misuses. J Higher Educ Manage Policy 19(2):24–62
12. Smith JP, McKnight A, Naylor R (1999) Graduate employment outcomes and university performance measures. Econ J 110(464):F382–F411
13. Smith J, Naylor R (2001) Determinants of degree performance. Oxf Bull Econ Stat 63:29–60
14. Usher A, Savino M (2006) A world of difference: a global survey of university league tables. Educational Policy Institute, Toronto