



Students' evaluation of academic courses: An exploratory analysis to an Italian case study

Marco Centoni^a, Antonello Maruotti^{a,b,1,*}

^a Dipartimento GEPLI, Libera Università Maria Ss. Assunta, Rome, Italy

^b Department of Mathematics, University of Bergen, Bergen Norway

ARTICLE INFO

Keywords:

Student evaluation of teaching (SET)
Multiple Correspondence Analysis
Private university
Indicators

ABSTRACT

Students' evaluations of teaching is a common practice in higher education institutions, with the main purpose of improving course quality and effectiveness. In this paper we would like to contribute to the existing literature on course and teaching evaluation by providing an empirical analysis based on questionnaires collected by an Italian private institution, namely the Libera Università Maria Ss. Assunta (LUMSA), for several degrees in Social Sciences. In order to identify the main factors affecting students' satisfaction, we use not only teaching indicators and degree-specific characteristics, but also control for student-specific characteristics. Our analysis is based on a Multiple Correspondence Analysis for categorical variables, which represents a very useful method to study the multidimensional relationship among qualitative variables, along with a hierarchical clustering, in order to better summarize the results. Our findings reveal that student satisfaction relates to teaching and course organization. Moreover, we find some evidence that students typically evaluate their course on the basis of their experience rather than their personal interests.

1. Introduction

The use of student feedback questionnaires to evaluate teaching performance and quality has become almost ubiquitous in higher education institutions. Questionnaires mainly collect information about lecturers, experienced teaching and overall organization of courses and are typically completed close to the end of the semester by all attending students. Such surveys are meant to provide feedback to help lecturers in improving the course in subsequent years remedying revealed areas of weakness, but there are also other reasons why courses might be evaluated. Following Bedggood and Donovan (2012), at least other five distinct reasons for conducting evaluations can be identified as beneficial. First, course evaluation helps in *investigating a known problem*, by identifying the possible reasons of some issue (like, e.g. a high dropout rate) and suggesting an appropriate remedy, without affecting the reputation held by the department and/or university. Second, evaluation is important for *improving programmes*, by identifying both good and bad aspects of a course which may not be readily apparent to the lecturer otherwise. Third, appropriate course evaluations could provide a valid way for *reorganization of material*, in the case where introduced topics do not appear logical and/or well structured to the students. Fourth,

questionnaires may help in *examining the impact of an innovation* on student interests and learning, in situations where a lecturer makes major changes in course content or in the methods of teaching. Fifth, student evaluations of teaching can be taken into account for *supporting applications for tenures*, when applications for new tenures or promotions are being considered by university departments.

Hence, the understanding of the evaluation process and the assessment of the quality of higher education teaching represent important multidimensional factors to take under consideration as part of university governance and quality management (Cadez, Dimovski, & Zaman Groff, 2017; Ferrante, 2017; Goos & Salomons, 2017; Linse, 2017; Marsh, 1987; Misanew & Tadesse, 2014; Spence, 2019; Stehle, Spinath, & Kadmon, 2012; Steinhardt, Schneiderberg, Götz, Baumann, & Krücken, 2017). The increased focus on the student experience in higher education has been accompanied by significant academic research in different countries, including UK (Grimes, Medway, Foos, & Goatman, 2017), Australia (Marsh, Ginns, Morin, Nagengast, & Martin, 2011), Germany (Seyfried & Ansmann, 2018), Italy (Bini & Masserini, 2016; Iezzi, 2005; Raponi, Martella, & Maruotti, 2016), Spain (del Carmen Bas, Tarantola, Carot, & Conchado, 2017; Murias, de Miguel, & Rodriguez, 2008), China (Yin, Wang, & Han, 2016), Saudi Arabia (Al Kuwaiti

* Corresponding author at: Via Pompeo Magno 22, 00192 Rome, Italy.

E-mail addresses: antonello.maruotti@uib.no, a.maruotti@lumsa.it (A. Maruotti).

¹ Address: Realfagbygget, Allégt. 41, 5020 Bergen, Norway.

& Subbarayalu, 2015), among others.

It must be noted that the participation of students in the process of quality assurance (and evaluation in general) has increased a lot in the past few years, starting from the Prague Communiqué in 2001, the “Bologna with student yes” in 2007, and “The National Unions of Students in Europe”, using different modalities from nation to nation and also within universities of the same country. The quality assurance process established in Italy has introduced a self-evaluation, periodic evaluation and accreditation (AVA) method that starting from 2013 has become compulsory for all Italian Universities. A key dimension of this evaluation system is the assessment of students’ satisfaction with respect to students’ expectations. Hence, the results of SET survey became part of the more general process of quality assurance of tertiary educational activities. The AVA system imposes that findings of SET should be periodically discussed by the main committees for quality assurance (e. g. self-evaluation committees, joint committees of professors and students and audit committees) within each degree program and by the main university ruling bodies with the aim of implementing and programming corrective actions (Chies, Graziosi, & Pauli, 2018; Murmura, Casolani, & Bravi, 2016). In this paper we would like to contribute to the existing literature linking satisfaction on course and teaching with some indicators and students’ characteristics by providing an empirical analysis based on a unique Italian dataset. This work is largely characterized by the use of questionnaires, in order to monitor students’ evaluations in higher education, by focusing on what students think about their education experience and the services provided by their university. We perform an empirical analysis using administrative data collected from the LUMSA database. Our novel dataset comprises graduates and undergraduates enrolled in any of the provided eight degrees in Social Sciences in 2015 and, interestingly, collects information on important academic and extra-academic observed variables.

We consider degree description, type of student (regular, repeating, long-stay) and the number of year enrolled in the degree (considered as a categorical variable) as some of the most relevant academic variables. Student’s gender and nationality are instead analysed to describe students’ extra-academic characteristics. This allows us to explore the existing relationships between student- and course-specific characteristics with the satisfaction, as measured by the questionnaire. To validate our analysis, we formulate different research questions, in line with the main research strands currently discussed in the literature. We explore how students’ overall satisfaction is linked to teaching quality (Bassi, Grilli, Paccagnella, Rampichini, & Varriale, 2017; Feistauer & Richter, 2018; Lee, Kim, & Chan, 2015; Nicolaou & Atkinson, 2019; Spooen & Christiaens, 2017; Sutherland, Warwick, & Anderson, 2019), and the role played by degree- (Thomas, 2018) and student-related (de Jager & Gbadamosi, 2013) factors.

To perform our analysis, we apply a multiple Correspondence Analysis (MCA) for categorical variables (Greenacre, 2010). MCA is part of a family of multidimensional descriptive methods revealing patterning in complex datasets when we dispose qualitative variables. Specifically, MCA is used to represent datasets as “clouds” of points in a multidimensional (Euclidean) space. It is distinctive in describing the patterns geometrically by locating each category of analysis as a point in a low-dimensional space. The results are interpreted on the basis of the relative positions of the categories and their distribution along the dimensions; as categories become more similar in distribution, the closer (distance between points) they are represented in space. It can also be a particularly powerful one as it “uncovers” groupings of categories in the dimensional spaces, providing key insights on relationships between categories, without needing to meet assumptions requirements such as those required in other techniques widely used to analyse categorical data. MCA has been used in higher education and other empirical fields (Kienstra & van der Heijden, 2015; van der Heijden, Teunissen, & van Orli, 1997). In this work, we use MCA to explore patterns of response to questionnaire items, i.e. similarities in the answering behaviours, and look for associations with degree and students characteristics, used as

auxiliary variables. To enrich the description of the data, we further provide a visualization based on a clustering approach. This is particularly useful when the number of individuals is very high and variables are structured according to a hierarchy leading to groups and subgroups of variables. This case is frequently encountered with questionnaires structured into topics and subtopics, as the one considered here (Le Dien & Pagès, 2003). Model-based and non-model-based clustering approaches have been widely used in the higher education literature (Belloc, Maruotti, & Petrella, 2011; Centoni, Del Panta, Maruotti, & Raponi, 2019; Giordani, Ferraro, & Martella, 2020; Shavelson, 1979).

To clarify a crucial point of our analysis, we explain in the following the construct of “satisfaction”. This is necessary to answer the underlying question “satisfaction with what? Course results, teaching, education, etc.”. Students’ satisfaction has been defined and measured in different ways (DeShields, Kara, & Kaynak, 2005; Elliott & Shin, 2002; Grace, Weaven, Bodey, Ross, & Weaven, 2012; Harvey, 1995, among others). However, as common practice in empirical research on student satisfaction, we use SET as a proxy for student satisfaction, measuring it by the evaluations they give to questionnaire items (Bennett & Kane, 2014 and references in Section 2). In other words, we consider students satisfaction with courses as encapsulated in the evaluations they give on all the items of the questionnaire. In this sense, our measure of student satisfaction is done through an attribute (multi-item) level of measurement (Szymanski & Henard, 2001). It has been shown that multi-item scales are, in many cases, more accurate and more valid than single item ratings (Marsh, 1987). Spooen, Mortelmans, and Christiaens (2014) have observed that “SET by means of one question (e.g., ‘Overall I was satisfied with the quality of this course’) may not be very helpful for both monitoring teaching quality and/or the improvement of teaching as this practice assumes that quality of instruction can be observed unequivocally.” (Spooen et al., 2014). Hence, what we actually do in our analysis is to explore how each item contributes to the overall satisfaction and to some extent to the quality of courses, since in the literature satisfaction (with a service, for example), and the quality of that service are used interchangeably, because the quality of the service can only be measured in terms of customer’s satisfaction with the service encounter (Huang & Sudhir, 2020). With the higher education sector becoming an increasingly competitive market, students have become to be viewed as customers (Browne, Kaldenberg, Browne, & Brown, 2017) and student satisfaction has become an important component of quality assurance (Tsiligiris & Hill, 2021). The notion of quality in higher education is a well-established topic (Harvey & Williams, 2010a, 2010b). Despite the concept of quality is not unitary (see Elassy, 2015; Harvey & Green, 1993 for an extensive discussion), there is evidence that student evaluation can be a valid indicator of quality (Pickford, 2013).

The plan of the paper is as follows. We review the literature and define the research questions considered in our exploratory analysis in Section 2. We describe the data source and the questionnaire items, along with the sample characteristics and auxiliary variables in Section 3. The descriptive analysis is first performed for all the available data, and then summarized for each degree course separately. Section 4 introduces the proposed methodology, while results are described and widely discussed in Section 5. We discuss some insights in Section 6 and comment on the implications for university management and future research in Section 7.

2. Literature review and research question development

Let us start introducing a very general, and rather difficult to explore, research question. *Q1: How does the perception of teaching quality relate to students’ overall satisfaction?* Teaching quality itself is vague concept and can be measured by the effectiveness of the lecturer or can be thought as the basis for an overall satisfactory experience. Empirical research on student satisfaction has found a strong relationship between overall course satisfaction and survey questions assessing various aspects of

teaching quality as, for example: teaching ability (Hearn, 1985), 'teaching environment' (Krahn & Bowlby, 1997), teaching quality and expertise (Green, Hood, & Neumann, 2015), teaching efficiency (organisation of the teaching activities, learning materials, receiving hours and information about courses) (Bini & Masserini, 2016), 'helpful lectures' (Sutherland et al., 2019), teaching ability and expertise (Douglas, Douglas, & Barnes, 2006), 'good teaching' (Grace et al., 2012). We believe this is the main aspect of university experience as it may strongly affect the performance of the students and make their stay at the university more profitable.

We investigate how student satisfaction, degree systems and course organization are connected, and in which direction students' characteristics are expected to relate to evaluation results. To validate our analysis, we explore the following research questions.

Degree related factors

Some courses are more difficult to teach (and learn) than others. Thus, course content is likely to influence overall evaluation and item responses depending on the type of courses required by the degree system. The more challenging the perception of the course, the lower the course satisfaction might be (see e.g. Cashin, 1990; Dev & Qayyum, 2017; Langbein, 1994; Spooren, Brockx, & Mortelmans, 2013). Empirical evidence shows that students evaluation tends to be more positive in humanitarian fields than scientific disciplines, like e.g. economics and business (Degheri, 2017; Feldman, 1978; Kember & Leung, 2011). Scholars now recognize that the characteristics of the course itself (e.g. level of course, the subject area and course difficulty/workload) can impact teaching ratings (Worthington, 2002). Zabaleta (2007) observed that instructors in higher-level (elective) courses receive higher evaluations than in lower-level required general education courses. Additionally, Uttl and Smibert (2017) found that the class subject (e.g. Math vs. English) is strongly associated with student evaluations ratings whereby professors who teach primarily quantitative courses are at a disadvantage. Put differently, convincing evidence indicates that background characteristics or factors that have nothing to do with the instructor's behaviour or effective teaching in the classroom could bias student ratings (Worthington, 2002).

According to the above references, our second research question is the following. Q2: *How do course structure and organization relate to the overall satisfaction perceived by students?* Appropriate organizational conditions (including supporting material, ease of website accessibility, continuous information updating) and education-supporting skills (like e.g. office hours and extra classes provided by teaching assistants) are expected to be valid and significant factors to improve perceived course satisfaction.

Student related factors

Students' characteristics, such as the students' interest for taking a course, disposition toward the courses, expected course grade, perceptions of the instructor, and to a lesser extent the physical attributes of the students, including gender and age, can and do impact ratings on SETs (Chen & Hoshower, 2003; Clayson, 2009; Titus, 2008). Students' gender has been the focus of many empirical studies on teaching evaluation since the seminal work of Morgan and Ogden (1981). Results are not unanimous on this matter. Several empirical works find evidence supporting the general feeling that females are typically more generous in evaluating lectures and teaching quality than male students (Centra & Gaubatz, 2000; Hancock, Shannon, & Trentham, 1993; Tatro, 1995). Other studies, instead, do not find a significant pattern of gender differences on teaching evaluations (see e.g. Aleamoni, 1999; Liaw & Goh, 2003).

Following this strand of research, we explore a few further research questions. Q3: *How do perceptions on students' overall satisfaction relate to gender?* As mentioned before, determining differences of teaching

quality and student satisfaction based on gender has been the focus of several empirical works, even though there is no agreement on this. This study uses a cross-sectional model analysis to determine the relationship between gender, students' satisfaction and their perception of lecturers' quality.

Q4: *How does the level of academic student performance relate to student overall satisfaction?* Students' academic retention might be strong influencers of student satisfaction. In this context, information on student activity status (*regular, repeating, long duration*) could represent a good indicator of academic students' career. A long stay in the university system could be a symptom of a slow or difficult career, which would unavoidably affect student satisfaction negatively.

In addition, a number of studies postulate a relationship between prior subject interest and teaching ratings (Wachtel, 1998) indicating that students with a greater interest in the subject area (as indicated by higher past grades) tend to give more favourable teacher ratings (Worthington, 2002). Although the results of most studies are cautionary (Feistauer & Richter, 2018), the findings suggest that courses taught by individuals, with a number of unique or specific characteristics, can impact student ratings.

Q5: *Is personal interest in the subject related to student overall satisfaction?* Students' personal interests in selecting a specific field of study are some of the key determinants for academic success. Student satisfaction with course quality should be therefore interpreted with regard to their interests for enrolling in a particular field of study. We would expect students who expressed high personal interest in the taught course to report higher level of satisfaction compared to other (less motivated) students.

3. Data

The empirical analysis is based on data from different cohorts of students from LUMSA University, an Italian private university offering degree programs in economics, marketing, psychology, law and many more, mainly in social sciences. We use the latest available data provided by the university statistical unit, which refer to the cohorts of students currently enrolled in 2015. Questionnaires were compiled from January 2016 to February 2017 by an online system.

As a private university, LUMSA is based on a direct and ongoing relationship between students and staff members and it boasts of humanistic traditions, technological innovations and Catholic roots as the very heart of the institution. Teaching is distributed across three main departments: the Department of Law, Economics, Politics and Modern languages (Rome campus), the Department of Law (Palermo campus) and the Department of Social Sciences, Communication, Education and Psychology (Rome campus). With around 300 internationally recognized lecturers, the University aims at providing both qualified education for young people and professional development programs for mature students.

For our analysis we will focus on the core courses in social sciences provided by LUMSA, i.e. those collecting the highest number of students: Economics, Marketing, Psychology, Law, Social Work, and Education. All the programs are designed to deliver specific high-level technical and cultural skills. Students in these programs were required to take a fixed sequence of compulsory courses that span over the first two years, a good part of their third year and, in a few cases, also their last year. The number of students enrolled in these programs are 3892 that represented 64.78% of the entire student body. Table 1 reports a description of the sample.

We collect data from compulsory questionnaires which are typically compiled by attending students before taking the final exam. To undertake the final exam, students are obliged to fill in all items in the questionnaire. Questionnaires collect students' opinion on different aspects of the attended course, such as study load, course organization, validity of lecturers (either in terms of teaching, availability and flexibility), course material and, at last, the overall personal interest and

Table 1
Sample information.

Variable	All degrees (n = 3892)	Economics (n = 463)	Marketing (n = 410)	Psychology (n = 343)
Male	1320	308	179	44
On time	3214	336	326	319
First year courses	787	117	126	149
Second year courses	928	128	117	194
Third year courses	935	218	167	–
Fourth year courses	430	–	–	–
Fifth year courses	812	–	–	–
Italian	3711	452	404	332

Variable	Law (Palermo) (n = 832)	Law (Rome) (n = 868)	Social Work (n = 339)	Education (n = 637)
Male	320	389	49	31
On time	709	578	315	631
First year courses	121	79	114	81
Second year courses	136	101	117	146
Third year courses	175	93	119	163
Fourth year courses	147	154	–	129
Fifth year courses	253	441	–	118
Italian	832	860	331	500

satisfaction. [Table 2](#) reports a detailed description of questionnaire items. This version of the questionnaire was established by ANVUR – Italian National Agency for the Evaluation of Universities and Research Institutes – in 2013, and apart from minor changes allowed at local level, is adopted by all Italian universities in order to allow comparisons at national level (advantages and disadvantages of standardised module evaluation are discussed in [Wiley, 2019](#); the National Evaluation Committee of the University System, see [CNVSU, 2010](#), discussed its construction and validation before making it publicly available). Compared with international practice, the questionnaire does not provide for the evaluation of various aspects of student experience, i.e. learning community, assessment and feedback, engagement, that it can find in well known SET questionnaire such as the National Student Survey (NSS) and Postgraduate Taught Experience Survey (PTES) for the UK, Student Evaluation of Educational Quality (SEQ) and National Survey of Student Engagement (NSSE) in North America, or the Course Experience Questionnaire (CEQ) in Australia. Currently, the questionnaire used in the present study is under revision by ANVUR to align it to the best worldwide practice and “to consolidate the system and objectives before their adoptions”. Students were asked to express their level of satisfaction for each item on a scale of 4 balanced modalities (two positive and two negative), labelled as follows: definitely yes (DY), partially yes (PY), partially not (PN) and definitely not (DN). The items address the multidimensionality of the satisfaction process. Each of the considered dimensions are measured by one or more items. Students’ satisfaction on the study load and the general organization (schedule, calendar examinations, etc.) are recorded, along with that of the definition of the

Table 2
Questionnaire items.

Item	Definition
I1	Is the background knowledge of the student sufficient to understand all the topics of the course?
I2	Is the overall study load acceptable compared to the assigned credits?
I3	Is the supporting material adequate?
I4	Is the exam structure clearly exposed by the lecturer?
D5	Is the planned timetable respected?
D6	Does the lecturer stimulate interest about the subject?
D7	Does the lecturer clearly explain the subject?
D8	Are teaching assistant’s activities in line with the overall goal of the course?
D9	Is the course program in line with the syllabus?
D10	Is the lecturer available to give further explanations after the lecture?
D11	Are course information easily accessible from the university website?
I11	Are you interested in the topics of the subject?

modalities and the rules for exam, the real availability of the teachers to meet the students to give the explanations. The satisfaction of the didactic material and the integrative didactic activities, as well as on the sustainability of the study load is investigated too; the organization of the lessons, in terms of adequacy of the classrooms for the lessons, on the exercises and the seminars and on the equipments used for the carrying out of the exercises are judged by the students. At last, the overall interest of the course is collected.

Preliminary descriptive statistics are presented in [Table 3](#), where we report the relative frequencies of each of the four modalities (DY, PY, PN, DN) in all the questionnaire items. The information is first summarized for all the collected questionnaires (*all degrees*) and then grouped by degree courses (*Economics, Marketing, Psychology, Law (Rome), Law (Palermo), Social Work, Education*). The evaluation mode is *DY* (thus very positive) for all the items and in all the courses, with the only exception of Economics and Marketing, where the modality *PY* is sometimes predominant, especially due to reasons mainly related to low-level of student’s background (item I1) or due to the not easy accessibility to university website (item D11). Law and Psychology programs seem to have the best performance, even though frequencies are on average quite consistent across items and degree courses. In any case, as a general comment, for all the items the percentage of students positively satisfied (*DY + PY*) exceeds the 80%, showing an evident skewness in items’ distribution. In this situation, one could reasonably ask whether this evidence is due to self-selection, meaning that only students who enjoyed the course (and the lecturer) were willing to give feedback. However, remember that questionnaires are compulsory for all the students. Moreover, to avoid any sample selection bias, we focus only on those students who attended more than 75% of total lectures.²

We collect evaluations from the administrative archive on 243 courses (35 of which are shared by two or more degrees), grouped as follows: 35 in Economics, 48 in Marketing, 21 in Psychology, 41 and 61 in Law at Rome and Palermo respectively, 29 in Social Work and 21 in Education; all first-cycle degree courses. Courses sizes is varying and as such is the number of questionnaire collected per course, whose range goes from 1 to 447 (mean = 72.5; SD = 68.2; median = 59). Even though questionnaires guarantee students’ anonymity, for our analysis we are able to create a unique key, which allows us to link students’ feedback to several student-specific characteristics. This represents one of the most powerful advantage of our data, since it could reveal a potential pattern existing among items, degree courses and students’ characteristics. This is of course of valuable use not only to improve teaching quality and performance, but also to raise other external (not course-specific) issues to be monitored by university governance.

The auxiliary variables that we use for our analysis are described in detail in the following, where we distinguish between student-specific and degree related factors.

Several students’ characteristics are linked to course evaluations in different ways and with different intensities. Characteristics of the students themselves and their biases in perception and expectations would include any references to: students’ subject interests; students’ disposition toward the courses; students’ gender and age. All aspects widely discussed above, and in the literature. The list of the selected variables together with the main summary statistics are provided in [Table 4](#), where we report the frequencies of characteristics grouped by degree courses. In our sample female students represent the majority in almost all courses, except for the Economics program. Following [Pounder \(2007\)](#), we also collect students’ academic status and maturity (from first- to fifth-year students in [Table 4](#)) as potential breakdown variables to account for differences in response patterns. The general tendency is that more experienced students are typically more lenient in giving evaluations than freshmen (younger) students. Similarly, we would

² As a general information, most of the courses provided by LUMSA require mandatory attendance, though some exceptions are allowed.

Table 3
Course specific descriptive statistics on item responses.

Item	All degrees (n = 17610)				Economics (n = 1910)				Marketing (n = 2111)				Psychology (n = 1260)			
	DY	PY	PN	DN	DY	PY	PN	DN	DY	PY	PN	DN	DY	PY	PN	DN
I1	0.52	0.37	0.08	0.03	0.34	0.49	0.13	0.04	0.42	0.44	0.12	0.03	0.53	0.38	0.07	0.02
I2	0.54	0.36	0.07	0.03	0.38	0.50	0.09	0.03	0.44	0.43	0.10	0.03	0.56	0.33	0.08	0.03
I3	0.57	0.35	0.06	0.02	0.40	0.48	0.10	0.02	0.47	0.41	0.09	0.03	0.59	0.34	0.05	0.02
I4	0.60	0.31	0.06	0.02	0.46	0.43	0.09	0.01	0.52	0.36	0.10	0.02	0.65	0.28	0.05	0.02
D5	0.64	0.31	0.04	0.01	0.53	0.42	0.05	0.01	0.57	0.36	0.05	0.01	0.69	0.37	0.03	0.01
D6	0.59	0.32	0.07	0.02	0.42	0.46	0.10	0.02	0.49	0.38	0.10	0.03	0.64	0.29	0.05	0.02
D7	0.60	0.32	0.06	0.02	0.42	0.46	0.10	0.02	0.51	0.37	0.09	0.03	0.64	0.30	0.04	0.02
D8	0.61	0.30	0.06	0.02	0.51	0.38	0.08	0.02	0.55	0.36	0.07	0.02	0.70	0.25	0.03	0.02
D9	0.56	0.37	0.05	0.02	0.40	0.51	0.07	0.01	0.45	0.47	0.06	0.02	0.63	0.33	0.03	0.01
D10	0.61	0.32	0.05	0.02	0.47	0.45	0.06	0.02	0.53	0.39	0.06	0.01	0.65	0.30	0.04	0.01
D11	0.53	0.39	0.06	0.02	0.36	0.53	0.09	0.02	0.39	0.49	0.10	0.02	0.60	0.33	0.05	0.02
I11	0.60	0.33	0.06	0.02	0.44	0.46	0.08	0.02	0.51	0.39	0.08	0.02	0.67	0.27	0.05	0.01

Item	Law (Palermo) (n = 4700)				Law (Rome) (n = 2086)				Social Work (n = 2082)				Education (n = 3461)			
	DY	PY	PN	DN	DY	PY	PN	DN	DY	PY	PN	DN	DY	PY	PN	DN
I1	0.62	0.30	0.05	0.02	0.52	0.39	0.07	0.02	0.48	0.40	0.09	0.03	0.54	0.35	0.08	0.03
I2	0.63	0.29	0.05	0.03	0.57	0.36	0.05	0.02	0.49	0.40	0.08	0.03	0.58	0.33	0.07	0.03
I3	0.64	0.28	0.05	0.02	0.60	0.33	0.05	0.02	0.52	0.39	0.07	0.02	0.61	0.30	0.06	0.03
I4	0.67	0.25	0.05	0.02	0.62	0.32	0.04	0.02	0.52	0.38	0.07	0.03	0.64	0.28	0.05	0.03
D5	0.71	0.24	0.03	0.01	0.65	0.31	0.05	0.01	0.54	0.38	0.06	0.02	0.68	0.26	0.04	0.02
D6	0.67	0.26	0.05	0.02	0.63	0.31	0.05	0.01	0.55	0.36	0.06	0.03	0.63	0.27	0.07	0.03
D7	0.68	0.25	0.05	0.02	0.63	0.32	0.04	0.01	0.54	0.37	0.06	0.03	0.63	0.27	0.07	0.03
D8	0.64	0.29	0.05	0.02	0.62	0.32	0.04	0.02	0.59	0.32	0.06	0.02	0.66	0.25	0.06	0.03
D9	0.67	0.28	0.04	0.02	0.60	0.36	0.03	0.01	0.48	0.43	0.06	0.03	0.59	0.34	0.05	0.02
D10	0.71	0.25	0.03	0.01	0.64	0.32	0.03	0.01	0.55	0.37	0.06	0.03	0.63	0.30	0.05	0.02
D11	0.64	0.31	0.04	0.01	0.55	0.39	0.05	0.01	0.44	0.44	0.09	0.03	0.55	0.37	0.06	0.02
I11	0.67	0.27	0.04	0.02	0.61	0.33	0.05	0.02	0.57	0.35	0.06	0.02	0.63	0.30	0.05	0.02

Table 4
Students characteristics related to the number of questionnaires collected.

Variable	All degrees (n = 17610)	Economics (n = 1910)	Marketing (n = 2111)	Psychology (n = 1260)
Male	4876	1137	728	105
On time	17475	1890	2088	1260
First year courses	5642	638	956	1214
Second year courses	4293	695	778	46
Third year courses	4117	577	377	-
Fourth year courses	2138	-	-	-
Fifth year courses	1420	-	-	-
Italian	15811	1820	1854	1197

Variable	Law (Palermo) (n = 4700)	Law (Rome) (n = 2086)	Social Work (n = 2082)	Education (n = 3461)
Male	1679	857	204	166
On time	4671	2058	2054	3454
First year courses	825	565	856	588
Second year courses	906	355	552	961
Third year courses	1352	258	672	881
Fourth year courses	1042	401	-	693
Fifth year courses	575	507	-	338
Italian	4687	1904	1995	2354

expect more favorable evaluations from regular students (on time students in Table 4) than repeating or long-stay students.

Finally, given the international nature of LUMSA university, it could be possible to observe significant differences in the responses between Italian and foreign students (we report the frequency of Italian students in Table 4).

Characteristics of the course itself and other environmental factors would include any references to course electivity; level of course; subject area; course difficulty; course expectations; sensitivity to students; and

more. It is crucial, therefore, to ensure that evaluation results control for the different nature of degree programs.

4. Methodology

In this section we introduce the descriptive methods we use for our analysis, which allow us to describe, visualize and synthesize data features in the presence of multivariate/multidimensional data. The first part of our application is based on Multiple Correspondence Analysis (MCA) for categorical variables (Greenacre, 2010), which represents a multidimensional (and graphical) method useful to study and analyze the existing relationship among qualitative variables. The second step of our analysis consists in grouping observations according to their similarities (or differences) with respect to all the considered variables, by performing a hierarchical clustering. MCA was chosen in order to explore the data matrix, identifying latent synthetic variables which account the most relevant information in the original data. In simple words, all info about questionnaire's items are synthesized in a smaller number of interpretable latent variables, summarizing the main features of the original data matrix. Component scores can be arranged into a new data matrix, which this time has a reduced number of latent variables. As a result, we get information which allows exploring the structure of examined items. In the second step of the sequential approach, students are classified by employing well-known hierarchical clustering algorithms, e.g. considering Euclidean distance and the Ward method, applied on the component scores to find stable clusters which identify students with similar values of the synthetic variables, i.e. similar patterns in answering questionnaire's items. In the following we describe the main ideas and advantages of the proposed methodology.

MCA can be viewed as a weighted principal component analysis process applied to a set of the binary variables but with the χ^2 -metric on row/column profiles, instead of the usual Euclidean metric. MCA represents data as clouds in a multidimensional Euclidean space, describing the patterns geometrically by locating each category of analysis as a

point in a low-dimensional space, often called dimension. Each dimension is a combination of the original variables and they are constructed so that the first explains as much of the difference between the observed and expected frequencies as possible. Subsequent dimensions are constructed, in order, to explain the maximum amount of the difference between the observed actual and expected values that remain (after the previous dimensions have been fit).

The results are interpreted on the basis of the relative positions of the categories and their distribution along the dimensions; as categories become more similar in distribution, the closer (distance between points) they are represented in space. Although it is mainly used as an exploratory technique, it can be a particularly powerful one as it uncovers groupings of categories in the dimensional spaces, providing key insights on relationships between categories, making the use of MCA particularly relevant in studies where a large amount of qualitative data is collected.

The main features of MCA, with pros and cons, can be summarized as follows:

- MCA is best suited for exploratory research and is not appropriate for hypothesis testing and its correspondence graphs allow spotting the strongest relationships in a set multiway crosstabs;
- the distance between categories is based on a χ^2 metric;
- categories which are closer together have higher χ^2 if analysed in a conventional cross-tabular format;
- MCA is very sensitive to outliers;
- the number of dimensions to be retained in the solution is based on dimensions with inertia;
- the contributions, the test values and the squared cosines help in the interpretation of the results.

In practice, we study students, variables and categories. Two students are close to each other if they answered the questions the same way. We will not be so interested in single students but rather in populations: are there groups of individuals? Moreover, we want to see the relationship between variables and the associations between categories. Two categories are close to each other if they are often taken together. We are also interested in looking for one or several continuous synthetic variables to summarize categorical ones. At last, we want to characterize groups of students by categories. Formally, from a data table, say X , of students \times variables it is therefore natural to construct an indicator (dummy) matrix with individuals in the rows and all of the categories for every variable in the columns. The element x_{ik} of this table has a value of 1 if student i carries category k , and 0 if it does not. This table has $I \times K$ dimensions (with $K = \sum_{j=1}^J K_j$, where J represents the number of variables) and is composed entirely of 0 and 1. The distances between students can be calculated by adding the differences between categories, $(x_{ik} - x_{i'k})^2$ and counterbalanced using a function inversely proportional to I_k (with I_k the number of students carrying category k). This distance (squared) is expressed as

$$C \sum_{k=1}^K \frac{(x_{ik} - x_{i'k})^2}{I_k}$$

where C is a constant. The distance between two categories k and k' is calculated by counting the individuals which carry either category k or category k' (such as $I_{k \neq k'}$), and counterbalancing using a function inversely proportional to I_k and $I_{k'}$. This distance can therefore be expressed as

$$C' \frac{I_{k \neq k'}}{I_k I_{k'}}$$

where C' is a constant. However, according to the encoding ($x_{ik} = 0$ or 1), the number of individuals carrying only one of the two categories is equal to $I_{k \neq k'} = \sum_{i=1}^I (x_{ik} - x_{ik'})^2$. Thus the distance above can be

rewritten as

$$C' \frac{1}{I_k I_{k'}} \sum_{i=1}^I (x_{ik} - x_{ik'})^2.$$

As in all principal component methods, the study of the inertia of the dimensions enables us to see whether or not the data is structured, and, on the other hand, to determine the number of dimensions to interpret. In practice, MCA has been computed from the Burt matrix and variables as principal components have been used in graphical representations.

We then combine the MCA method with hierarchical clustering. Only the first dimensions can be retained to stabilize the clustering by deleting the noise from the data. Performing a clustering onto the first principal components of MCA is a very usual practice especially for questionnaires. Hierarchical trees are considered (James, Witten, Hastie, & Tibshirani, 2013). A hierarchical tree can be viewed as a sequence of nested partitions from the one in which each individual is a cluster to the one in which all the individuals belong in the same cluster. To obtain a partition, in this paper, we use the Ward's criterion. This criterion is based on the Huygens theorem which allows to decompose the total inertia (total variance) in between and within-group variance. The Ward's method consists in aggregating two clusters such that the growth of within-inertia is minimum (in other words minimizing the reduction of the between-inertia) at each step of the algorithm. The within inertia characterizes the homogeneous of a cluster. The hierarchy is represented by a dendrogram which is indexed by the gain of within-inertia. The hierarchical clustering is performed onto the dimensions scores.

All the computations, estimations and graphical representations are implemented in R 4.0.3 (R Core Team, 2019) using the psych (Revelle & Revelle, 2015) and FactoMineR (Lê, Josse, & Husson, 2008) packages.

5. Statistical results

In this section we apply the methodology described in Section 4. We first consider the degree courses all together. Then, a degree-specific analysis is conducted to better appreciate differences between disciplines.

We start by performing MCA on all the available questionnaires. The inertia of the first two dimensions gives information on the relationships among variables and would suggest the appropriate number of dimensions to be selected. Bearing in mind that the inertia of a data table is the (weighted) sum of the variances of its columns. In our case, the first two dimensions express the 78.30% of the total inertia, which means that the 78.30% of the total variability is explained by the plane. Inertia values quantify the amount of variation accounted for by the corresponding dimensions. Since this value is strongly greater than the reference value of 8.75%³, we can conclude that the variability explained by this plane is highly significant. Therefore, we restrict the rest of our analysis on the first two axis. These synthetic variables are often used as mathematical tools to reduce the dimensionality problem, but may also have a physical meaning and help in the description of the results. We summarize their physical meaning in Table 5 and discuss it throughout the following text.

In the rest of the paper, we display the results referring to the first two dimensions only, but provide comments on the others which significantly contain some information. We further investigate what factors are the most separated on the plane, that is which variables are the most appropriate in discriminating among response patterns. According to the Wilks test p-value, we identify *degree* as the best qualitative variable able to capture the distance among observations,

³ The reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 100 data tables of equivalent size on the basis of a uniform distribution.

Table 5
Definition of the principal components.

Degree	Definition of the First Dimension	Definition of the Second Dimension
All	Course organization	Lecturer behavior
Economic	Course organization	Lecturer quality
Marketing	Course organization	Lecturer behavior
Psychology	Course expectation	Lecturer performance
Law (Rome)	Lecturer behavior	Course organization
Law (Palermo)	Lecturer quality	Lecturer performance
Social Work	Course organization	Lecturer performance
Education	Course organization	Lecturer quality

followed by *year of attendance*, *maturity* and *Italian*. Specifically, we find that degrees in Law (either in Rome and Palermo), Education and Psychology are the closest to the cloud identifying *definitely positive* evaluations. Economics and Marketing are mainly *partially positive*-evaluated by the students, hence with room left for improvements. Degree in Social Work lies instead in the quadrant characterized by negative evaluations, even though partially positive evaluations are also widely acknowledged.

Focusing on students' characteristics, fourth- and fifth-year students are the most likely to evaluate course positively, while Italian students are on average less generous in their evaluations compared to non-Italian students. Interestingly, we did not find any evidence supporting potential differences based on gender. The main results are displayed graphically in Fig. 1.

To interpret the dimensions, it is useful to look at the contribution of the variables to each of the axis (see Fig. 1). The first dimension is mainly explained by items referring to course organization (items D9, D11, I3, I2), where the *DP* evaluation identifies the negative pole of the axis. The second dimension is strictly related to lecturer's behaviour (items D7, D6, I3, D10), where the positive pole of the axis is characterized by the *DN* modality.

A similar analysis has been also conducted for each degree separately, as degree-specific characteristics may reasonably affect students' responses differently. In the following we summarize the main results characterizing each degree course.

Degree in economics. Our estimates suggest to restrict the analysis on the first two dimensions which, together, explain the 74.24% of the total inertia. According to the Wilks test p-value, the academic year in which a student is enrolled is the variable that discriminates the most among students along the first dimension: long-stay students are typically more satisfied than repeating students, who might give worse evaluations. Course organization (items D9, I3 and D11) and lecturer's quality contribute the most in interpreting such dimension.

Gender helps in discriminating students along the second dimension, where female students provide a more generous evaluation than males. In other words, female are more likely to provide a good evaluation as long as they judge positively the lecturer's clarity of exposition and ability in stimulating interest. We explore the third dimension too, as it provides further interesting insights. It is well-differentiated by students' nationality, where foreign students tend to give more positive evaluations than Italian students. Quality of teaching is very crucial in these two dimensions. Indeed, unsatisfied evaluations are strongly related with low-level lecturers (who might teach unstructured material or do not stimulate students' interest) and lack in course organization. Thus, many aspects are shared with the previous analysis on all the degree courses, with the role of lecturer being even more central (see Fig. 2). This, of course, may have implications in terms of recruitment and promotions.

Using the MCA results, we then apply the hierarchical clustering procedure, as described in Section 4. We identify three main clusters of students, which are essentially classified according to their evaluations: those who evaluate all courses *definitely positive*, those who give a *partially positive* evaluation, and those who are dissatisfied with both quality of courses and lecturers' level.

Degree in marketing. As in the previous case, we select the first two axis, which together explain almost the 72.88% of the total inertia. The

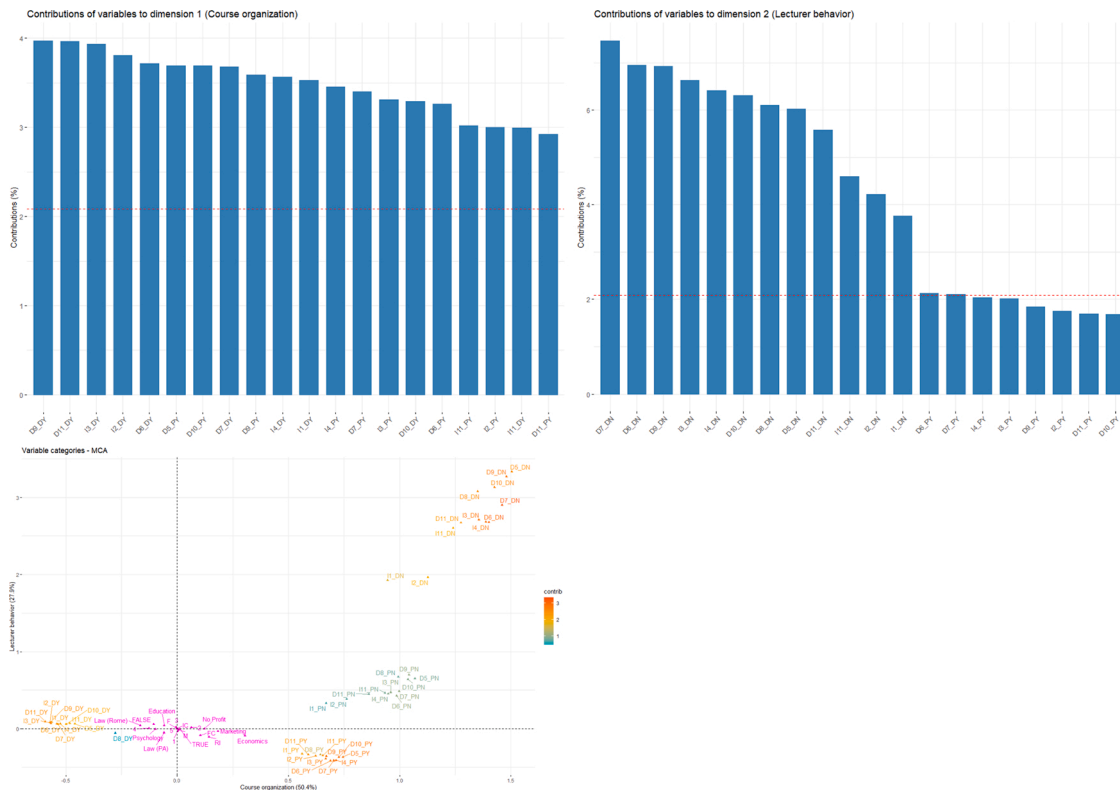


Fig. 1. MCA results. Top panel: Contributions to dimensions. Bottom panel: Superimposed representation on the first plane. All data.

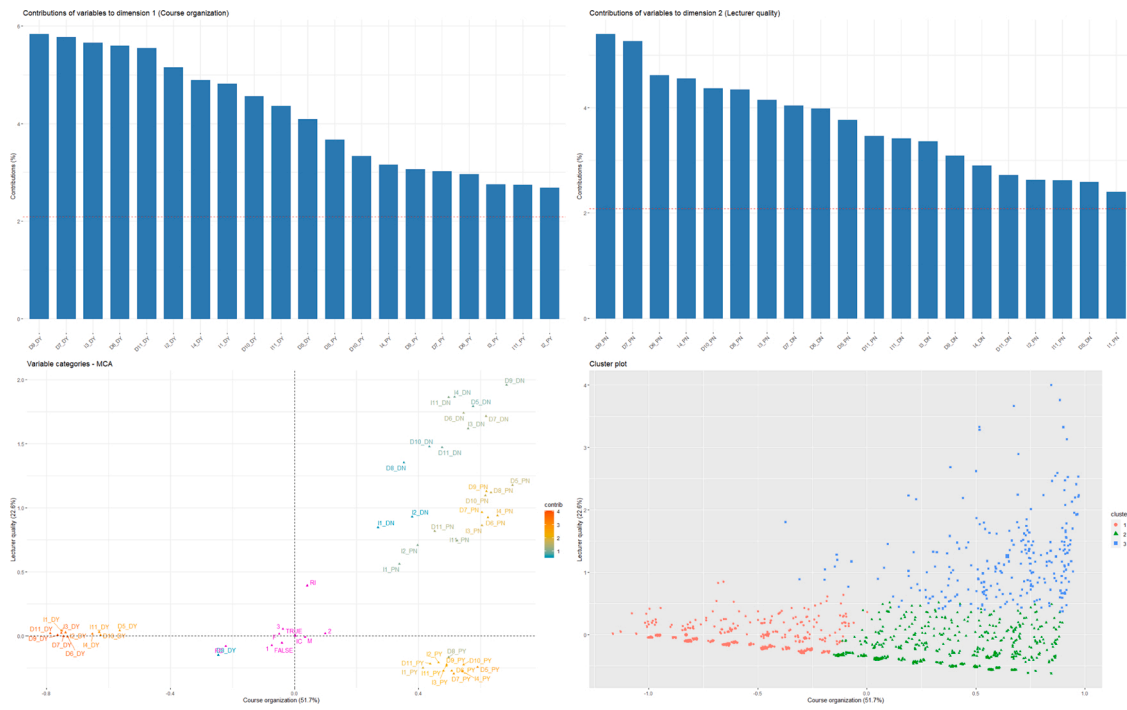


Fig. 2. Graphical representation of the results: Degree in Economics. Top panel: Contributions to dimensions. Bottom left panel: Superimposed representation on the first plane. Bottom right panel: Hierarchical clustering.

MCA results are almost perfectly in line with the profile outlined in the pooled analysis. Therefore, for sake of brevity, we prefer not to go too much into details on this part. We just want to emphasize that negative evaluations mainly come from long-stay students (see Fig. 3) and that nationality and academic year are the most significant variables to discriminate students on the axis. Three groups are identified by the hierarchical clustering. Their interpretation is close to the one discussed for the Economic degree; and this is not surprising. However, the

proportion of generally satisfied students, i.e. those clustered in cluster 1, is a bit higher (48% vs. 40%) than those in the Economic degree.

Degree in psychology. As a preliminary analysis, we start by noticing that all students are enrolled as *regular* and more than the 90% of them are females. Also in this case a 2-dimensional model is chosen, representing the 76.07% of the total variability. The first dimension refers to *partially positive* evaluations on course’s expectations (items D9 and D5) and on lecturer’s characteristics (his/her availability to give further

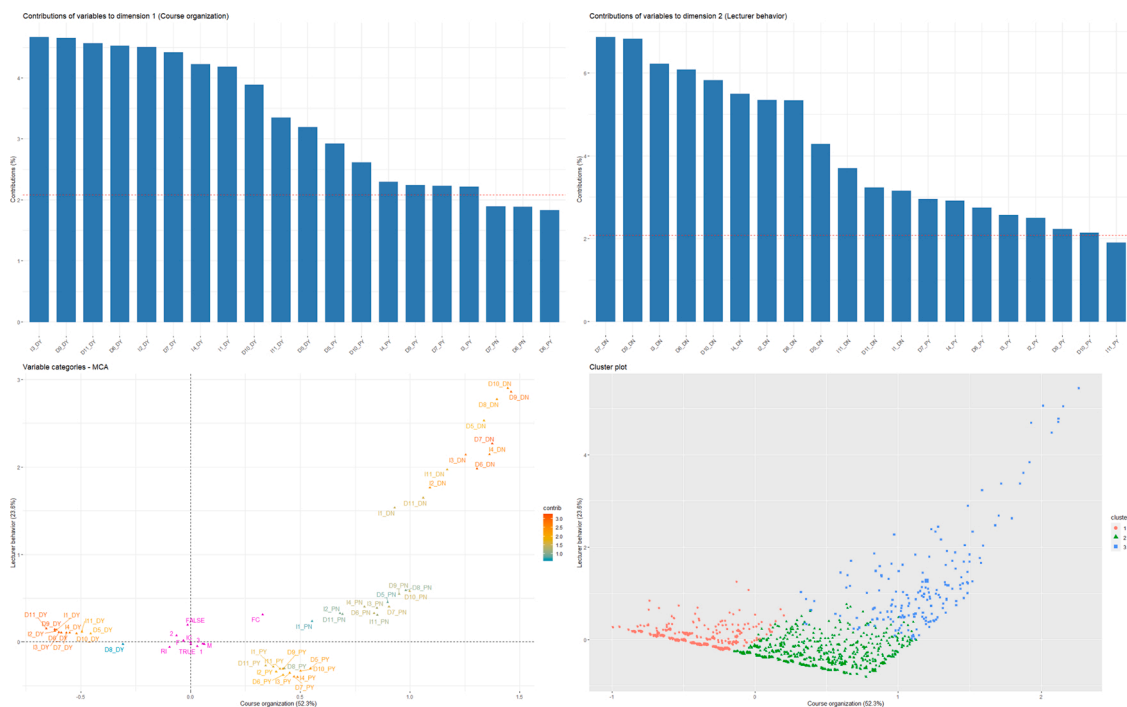


Fig. 3. Graphical representation of the results: Degree in Marketing. Top panel: Contributions to dimensions. Bottom left panel: Superimposed representation on the first plane. Bottom right panel: Hierarchical clustering.

explanations after the class and his/her clarity in explaining the subject). The second dimension is associated, instead, with very poor performance of both lecturer (items D10, D9, D6) and teaching assistant (item D8) which lead to *definitely negative* evaluations. Finally, the third dimension (with its 7% explained inertia) also contains interesting features and refers not only to lecturer's quality (item D10), but also to characteristics related to course's organization, with a general dissatisfaction being in the positive pole of the axis. Students in this field pay a lot of attention on lecturer's ability, which represents one of the first criterion for students to evaluate the attended course (Fig. 4). The clustering structure inferred by the ascending hierarchical method identifies three clusters, in line with the results of the previous analysis.

Degree in law (Rome and Palermo Campus). Rome campus collects the highest number of students' questionnaires. As in all the previous cases, MCA technique selects two main axis, together explaining more than 80% of the total inertia. We identify student's academic maturity as the most discriminating variable among students. Males, first-year and foreign students are the most satisfied with lecturer's quality and course organization, while repeating and long-stay students give on average only partially positive feedback. The first dimension is strictly related to positive evaluations of lecturer's quality (items D10, D6 and D7). Unsatisfactory assessments on course accessibility and supporting material identify the second dimension, while low-level assessment of lecturer's quality and ability characterizes the third dimension. Finally, the classification technique identifies three main clusters, classifying students in definitely satisfied, partially satisfied and dissatisfied categories (Fig. 5).

Several differences can be highlighted when analysing the same degree course in Palermo. We find that student's maturity and nationality are the most discriminating variables among students. Items D10 and I4 are the most related to the first dimension, hence identifying a strict connection with both lecturer's availability and clarity. Inappropriate supporting material (item I3) and inadequate lecturer's ability in stimulating interest are the most representative factors in the second dimension, while the third axis relates also to the general interest in the topics of the subject (item I11), see Fig. 6.

As a further difference, four clusters have been identified in Palermo, with a clear polarization of students across the four evaluation

modalities available in the questionnaires.

Degree in social work. Also this degree course shares a 2-dimensional representation, identifying nationality and student's maturity as the most discriminating variables on the plane. Course organization (in terms of accessibility, appropriate timetable and study load) can summarize the first dimension, which is mainly related to items D11, D9 and I2. Low-level quality and bad performances of both lecturer and teaching assistant strongly explain the second dimension (Fig. 7).

Degree in education. The two identified factors relate to course organization (first dimension), and lecturer's quality (second dimension). The clustering structure is also similar to the one identified in most of the previous degree courses (Fig. 8).

6. Linking the statistical results to research questions: a discussion

The statistical results described in Section 5 give us more details on the story underlying the data at hand. Data exploration is rather useful as a first of any analyses. Then policy implications should be discussed to make data exploration effective, linking the statistical results to the developed research questions, identified from the review of the literature (see Section 2). To increase the readability, we report again the main research questions (Q1,...,Q5) and the respective answers (A1,..., A5).

Q1: *How does the perception of teaching quality relate to students' overall satisfaction?*

A1: This is a very general aspect of the analysis. Of course, different definitions of teaching quality may lead to different considerations. These definitions are heterogeneous and often vague. Let us split teaching quality into two parts, one referring to the course planning and implementation, and another one related to lecturer's characteristics (from his/her relationship with the students to his/her performance). Under this framework, we can conclude that teaching quality, so defined, is strongly linked with students' satisfaction. Indeed, as discussed in the main text and also shown in Table 5, the satisfaction is strongly driven by

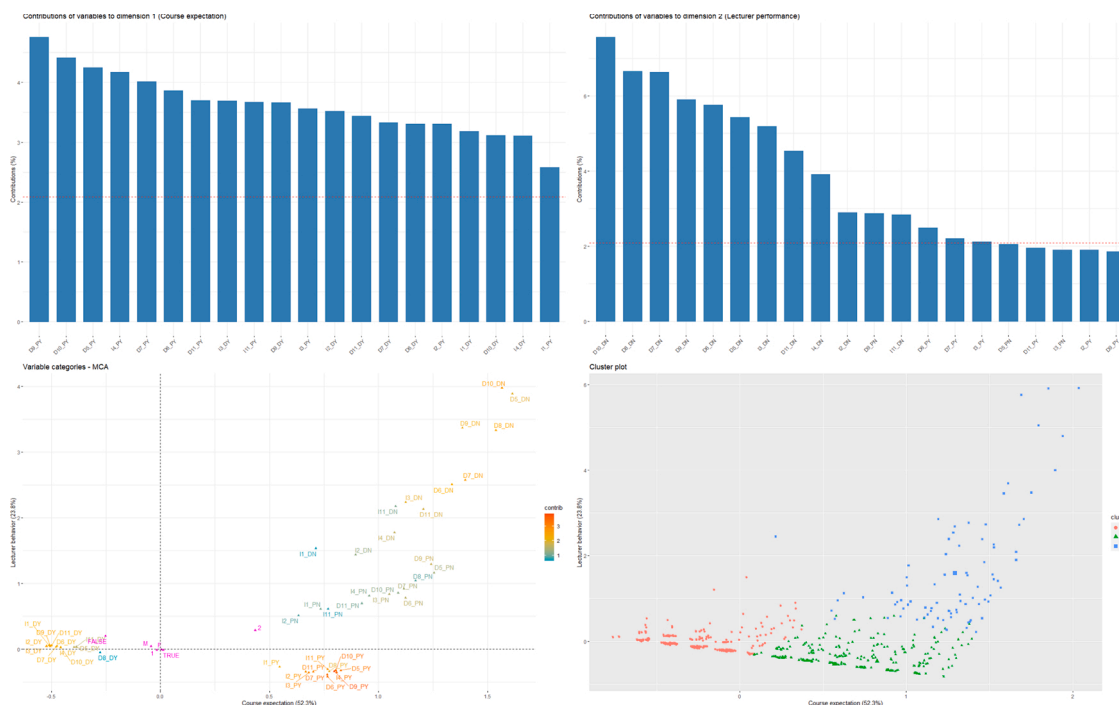


Fig. 4. Graphical representation of the results: Degree in Psychology. Top panel: Contributions to dimensions. Bottom left panel: Superimposed representation on the first plane. Bottom right panel: Hierarchical clustering.



Fig. 5. Graphical representation of the results: Degree in Law (Rome Campus). Top panel: Contributions to dimensions. Bottom left panel: Superimposed representation on the first plane. Bottom right panel: Hierarchical clustering.



Fig. 6. Graphical representation of the results: Degree in Law (Palermo Campus). Top panel: Contributions to dimensions. Bottom left panel: Superimposed representation on the first plane. Bottom right panel: Hierarchical clustering.

course- and lecture-related factors as synthesized by the principal components. This is true for both types of analyses, on the pooled and the degree-specific data.

Q2: *How do course structure and organization relate to the overall satisfaction perceived by students?*

A2: Course organization drives students' satisfaction. Accordingly it attracts the main attention from the students when they have to

evaluate a course. Two aspects are carefully taken into consideration by students: the syllabus, which must be in line with the course program, and the availability of adequate material to support the students during the course and the preparation of the exam. This is true for all degree, with the exception of the Law degree, where the interpersonal relationships with the lecturer have higher importance than the course organization.

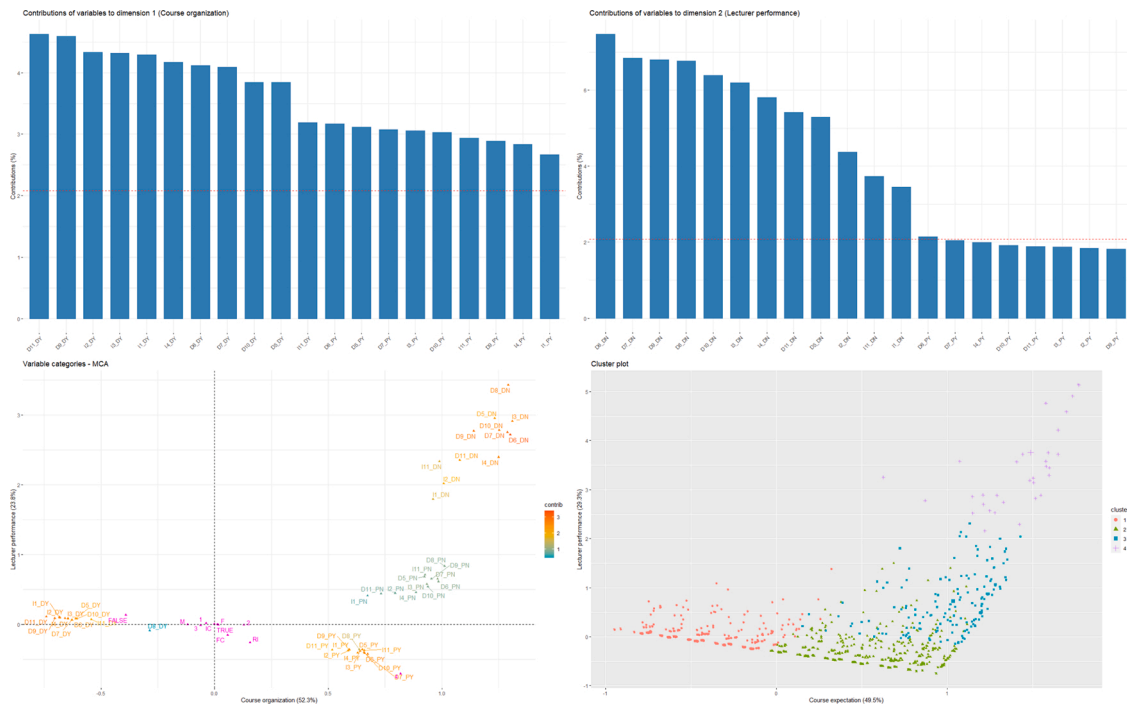


Fig. 7. Graphical representation of the results: Degree in Social Work. Top panel: Contributions to dimensions. Bottom left panel: Superimposed representation on the first plane. Bottom right panel: Hierarchical clustering.



Fig. 8. Graphical representation of the results: Degree in Education. Top panel: Contributions to dimensions. Bottom left panel: Superimposed representation on the first plane. Bottom right panel: Hierarchical clustering.

Q3: *How do perceptions on students' overall satisfaction relate to gender?*
 A3: There is a wide debate in the literature, as reported in Section 2, on this aspect. Our analysis is not conclusive in general, but some comments can be drawn on the Social Sciences degrees so far analysed. We do not observe any difference in the response patterns, and thus in the satisfaction, between males and females. If

any gender effect exists, it may related to other discipline-specific characteristics and degrees.
 Q4: *How does the level of academic student performance relate to student overall satisfaction?*
 A4: The answer to this question is clearly degree-specific, according to our results. High level of academic student performance is linked with Economics, Marketing and Law students, where long-

stay students tend to give more negative evaluations compared to regular students. Instead for the students in Education we find evidence of the opposite tendency, which is however counterintuitive and deserves more investigation.

Q5: *Is personal interest in the subject related to student overall satisfaction?*

A5: Up to our knowledge, the literature does not provide any hints so far. Our analysis suggests that I11, i.e. the item referring to personal interests in the subject, follows the same pattern of all other items. However, it is not possible to identify a cause-effect relationship, which deserves a different statistical methodology to be investigated. Speculating a bit on the MCA results, personal interests drive the choice of a specific degree. This result arises when data are analysed as a pool, as difference across degrees are depicted. The role played by personal interests in evaluating single courses is instead more subtle, as *courses are of interest by default* as students made a degree choice on the basis of taught course. We further notice that I11 contributes to explain dimensions related to the lecturer's quality. This may indicate that the interest in the subject is the effect of a good lecturer rather than the cause of a positive/negative evaluation of the lecture him/herself.

7. Conclusions

This paper provides a novel empirical analysis on course evaluation focusing not only on teaching indicators and degree-specific characteristics, but also controlling for both academic and extra-academic student-specific characteristics. With the aim of guiding scholars to explore similar data with appropriate statistical methods and driven by education-related research questions, we apply a multiple correspondence analysis for categorical variables, which represents a method useful to study and analyse the existing relationship among qualitative variables, and hierarchical clustering, in order to get more info on the heterogeneity among students. The results are of interest for statisticians and researchers in higher education. From a statistical perspective, the analysis leads to the identification of two (or three, for some degrees) latent dimensions, summarizing the evaluation process. These are related to the evaluated course and to the lecturer's characteristics. Hence, dimensionality reduction methods can be used to explore the data in a multivariate setting, making the story underlying the data easy to be interpreted. Moreover, we provide evidence that data are heterogeneous, i.e. clusters of students arise. Though this is not a surprising result, we believe it is important to remark that heterogeneity arise in data on the evaluation of academic courses. If ignored, heterogeneity may strongly bias results based on e.g. regression modelling or other statistical approaches. We link the statistical results to open research questions on the relationships between students satisfaction and several external aspects, widely discussed in the higher education literature. In this respect, results are rather consistent with the existing literature and remark the importance of a good course organization as well as having excellent (with different capabilities) lecturers. When the results in the literature are mixed, we confirm the specificity of some results, which may be valid for some degrees, but not for others. Of course, this is an exploratory analysis and, as such, it should be used to summarize available information to identify strength and weaknesses of courses and degrees, but not to comment on cause-effect relationships. Indeed, more complex methods can be further adopted, but the proposed one could be used for reporting and monitoring the evaluation of courses.

7.1. Implications for university management and teaching staff

This version of the questionnaire is not intended to evaluate lecturers, but rather as a monitoring tool of teaching activities, promptly identifying critical issues. Thus, though there is a debate on using the questionnaire as an assessment tool for evaluating staff performance

(Hornstein, 2017), the university management should use it to improve the teaching environment, rather than for selection and promoting the staff. However, our results have some policy implications that should be considered by those responsible for educational practices and policies at the university. Extremely relevant aspects to be monitored are represented by the ease of access to course information from the university website and the adequacy of course material. Both aspects play a crucial role in the evaluation process and are concrete issues which can be easily addressed, once identified. With regard to teaching, important determinants of student satisfaction are the clarity of exposition and ability in stimulating interest. These features are lecturer-specific and it is more difficult to plan interventions. Nevertheless, refresher courses for staff are constantly organized. This is of particular importance these days, where lecturing online leads to a different lecturing approach than before.

7.2. Limitations and further research

Our analysis has some limitations. The official questionnaire Italian universities use to evaluate academic courses differs from those used elsewhere. This complicates comparisons with other countries, where more extensive questionnaires are considered. Nevertheless, the items considered in this work are strictly related with widely discussed topics in the literature and provide further insights on those research questions. To align the tool to most general questionnaires, ANVUR is currently revising the set of items.

Here, we focus on Social Sciences degrees only. The results cannot be generalized to Natural and Engineering Sciences, without further analyses. Of course, the same methodological approaches apply. A further specificity of the proposed analysis is the use of data from a private university. As such, one of the main characteristics of the LUMSA University, shared by other private institutions, is the direct and ongoing relationship between staff members and their students. The sample characteristics, however, do not differ from those observed in other Italian public universities (Belloc et al., 2011; Belloc, Maruotti, & Petrella, 2010), and this allow us to be confident on the generalization of the results to other universities. There is a range of additional factors whose impact could be investigated. For instance, in future studies, we would consider also information on the easiness or difficulty of the topic of the course (e.g. % of retention rate or average mark in the final examinations) and information related to students' educational background (e.g. results in the entrance test, final mark in secondary school, marks in other examinations). This is currently not possible because questionnaires are not linked with students' careers and information collected at the time of enrolment.

Based on this exploring data analysis, we may plan interventions and then assess the impact of a set of policy interventions on evaluation indicators. This assessment will be dealt with not in mere observational terms, but in causal terms. The causal inference approaches could be chosen on the grounds of the possibility to control for possible confounders: they will range from more formalised methods based on the counterfactual approach to comparative methods, like synthetic control models and scenario-based approaches.

Acknowledgments

We express deep thanks to three anonymous referees for their very valuable comments. We thank Vieri Del Panta and Marco Rosati for their support in the data preparation and for the discussions on a draft version of this work.

References

- Al Kuwaiti, A., & Subbarayalu, A. V. (2015). Appraisal of students experience survey (SES) as a measure to manage the quality of higher education in the Kingdom of Saudi Arabia: An institutional study using six sigma model. *Educational Studies*, 41 (4), 430–443.

- Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13(2), 153–166.
- Bassi, F., Grilli, L., Paccagnella, O., Rampichini, C., & Variabile, R. (2017). New insights on student evaluation of teaching in Italy. *Convegno della Società Italiana di Statistica* (pp. 263–274). Cham: Springer.
- Bedgood, R. E., & Donovan, J. D. (2012). University performance evaluations: What are we really measuring? *Studies in Higher Education*, 37(7), 825–842.
- Belloc, F., Maruotti, A., & Petrella, L. (2010). University drop-out: An Italian experience. *Higher Education*, 60, 127–138.
- Belloc, F., Maruotti, A., & Petrella, L. (2011). How individual characteristics affect university students drop-out: A semiparametric mixed-effects model for an Italian case study. *Journal of Applied Statistics*, 38(10), 2225–2239.
- Bennett, R., & Kane, S. (2014). Students' interpretations of the meanings of questionnaire items in the National Student Survey. *Quality in Higher Education*, 20(2), 129–164.
- Bini, M., & Masserini, L. (2016). Students' satisfaction and teaching efficiency of university offer. *Social Indicators Research*, 129(2), 847–862.
- Browne, B. A., Kaldenberg, D. O., Browne, W. G., & Brown, D. J. (2017). Student as customer: Factors affecting satisfaction and assessments of institutional quality. *Journal of Marketing for Higher Education*, 8(3), 1–14.
- del Carmen Bas, M., Tarantola, S., Carot, J. M., & Conchado, A. (2017). Sensitivity analysis: A necessary ingredient for measuring the quality of a teaching activity index. *Social Indicators Research*, 131(3), 931–946.
- Cadez, S., Dimovski, V., & Zaman Groff, M. (2017). Research, teaching and performance evaluation in academia: The salience of quality. *Studies in Higher Education*, 42(8), 1455–1473.
- Cashin, W. E. (1990). Students do rate different academic fields differently. *New Directions for Teaching and Learning*, 1990(43), 113–121.
- Centoni, M., Del Panta, V., Maruotti, A., & Raponi, V. (2019). Concomitant-variable latent-class beta inflated models to assess students' performance: An Italian case study. *Social Indicators Research*, 146(1–2), 7–18.
- Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, 71(1), 17–33.
- Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education*, 28, 71–88.
- Chies, L., Graziosi, G., & Pauli, F. (2018). The impact of the Bologna process on graduation: New evidence from Italy. *Research in Higher Education*. <https://doi.org/10.1007/s11162-018-9512-4>.
- Clayton, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, 31, 16–30.
- CNVSU. (2010). *Progettazione, implementazione e validazione di un questionario per la valutazione della didattica erogata a studenti universitari*, RdR 2/10. <http://www.cnsvu.it>.
- de Jager, J., & Gbadamosi, G. (2013). Predicting students' satisfaction through service quality in higher education. *International Journal of Management Education*, 11(3), 107–118.
- Degheri, T. J. (2017). *An empirical look at the impact of course and faculty characteristics on student evaluations*. University of San Diego. PhD Thesis.
- DeShields, O. W., Kara, A., & Kaynak, E. (2005). Determinants of business student satisfaction and retention in higher education: Applying Herzberg's two-factor theory. *International Journal of Educational Management*, 19(2), 128–139.
- Dev, S. S., & Qayyum, N. (2017). Major factors affecting students' perception towards faculty evaluation of teaching (SET). *Journal of Social Studies Education Research*, 8(3), 149–167.
- Douglas, J., Douglas, A., & Barnes, B. (2006). Measuring student satisfaction at a UK university. *Quality Assurance in Education*, 14(3), 251–267.
- Elassy, N. (2015). The concepts of quality, quality assurance and quality enhancement. *Quality Assurance in Education*, 23(3), 250–261.
- Elliott, K. M., & Shin, D. (2002). Student satisfaction: An alternative approach to assessing this important concept. *Journal of Higher Education Policy and Management*, 24(2), 197–209.
- Feistauer, D., & Richter, T. (2018). Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. *Studies in Educational Evaluation*, 59, 168–178.
- Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*, 9(3), 199–242.
- Ferrante, F. (2017). Assessing quality in higher education: Some caveats. *Social Indicators Research*, 131(2), 727–743.
- Giordani, P., Ferraro, M. B., & Martella, F. (2020). *An introduction to clustering with R*. Singapore: Springer.
- Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: Assessing selection bias in course evaluations. *Research in Higher Education*, 58(4), 341–364.
- Grace, D., Weaven, S., Bodey, K., Ross, M., & Weaven, K. (2012). Putting student evaluations into perspective: The Course Experience Quality and Satisfaction Model (CEQS). *Studies in Educational Evaluation*, 38(2), 35–43.
- Green, H. J., Hood, M., & Neumann, D. L. (2015). Predictors of student satisfaction with university psychology courses: A review. *Psychology Learning & Teaching*, 14(2), 131–146.
- Greenacre, M. J. (2010). Correspondence analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5), 613–619.
- Grimes, A., Medway, D., Foss, A., & Goatman, A. (2017). Impact bias in student evaluations of higher education. *Studies in Higher Education*, 42(6), 945–962.
- Hancock, G. R., Shannon, D. M., & Trentham, L. L. (1993). Student and teacher gender in ratings of university faculty: Results from five colleges of study. *Journal of Personnel Evaluation in Education*, 6(3), 235–248.
- Harvey, L. (1995). Student satisfaction. *New Review of Academic Librarianship*, 1(1), 161–173.
- Harvey, L., & Green, D. (1993). Defining quality. *Assessment & Evaluation in Higher Education*, 18(1), 9–34.
- Harvey, L., & Williams, J. (2010). Fifteen years of quality in higher education. *Quality in Higher Education*, 16(1), 3–36.
- Harvey, L., & Williams, J. (2010). Fifteen years of quality in higher education (Part Two). *Quality in Higher Education*, 16(2), 81–113.
- Hearn, J. C. (1985). Determinants of college students' overall evaluations of their academic programs. *Research in Higher Education*, 23(4), 413–437.
- Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, 4(1), 1304016.
- Huang, G., & Sudhir, K. (2020). The causal effect of service satisfaction on customer loyalty. *Management Science*, 67, 317–341.
- Krahn, H., & Bowlby, J. W. (1997). Good teaching and satisfied university graduates. *Canadian Journal of Higher Education*, 27(2/3), 157–179.
- Iezzi, D. F. (2005). A method to measure the quality on teaching evaluation of the university system: The Italian case. *Social Indicators Research*, 73(3), 459–477.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*, Vol. 112. Springer.
- Kember, D., & Leung, D. Y. (2011). Disciplinary differences in student ratings of teaching quality. *Research in Higher Education*, 52(3), 278–299.
- Kienstra, N. H., & van der Heijden, P. G. (2015). Using correspondence analysis in multiple case studies. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 128(1), 5–22.
- Langbein, L. I. (1994). The validity of student evaluations of teaching. *PS: Political Science & Politics*, 27(3), 545–553.
- Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1–18.
- Le Dien, S., & Pagès, J. (2003). Hierarchical multiple factor analysis: Application to the comparison of sensory profiles. *Food quality and preference*, 14(5–6), 397–403.
- Lee, H. H., Kim, G. M. L., & Chan, L. L. (2015). Good teaching: What matters to university students. *Asia Pacific Journal of Education*, 35(1), 98–110.
- Liaw, S. H., & Goh, K. L. (2003). Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management*, 17(1), 37–43.
- Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, 54, 94–106.
- Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11(3), 253–388.
- Marsh, H. W., Ginns, P., Morin, A. J., Nagengast, B., & Martin, A. J. (2011). Use of student ratings to benchmark universities: Multilevel modeling of responses to the Australian course experience questionnaire (CEQ). *Journal of Educational Psychology*, 103(3), 733.
- Misanew, A., & Tadesse, M. (2014). Determinants of student and staff satisfaction with services at Dilla University, Ethiopia: Application of single and multilevel logistic regression analyses. *Social Indicators Research*, 119(3), 1571–1587.
- Morgan, B. B., & Ogden, G. D. (1981). Non-instructional correlates of student ratings: A brief review. *Applied Psychology*, 30(3), 409–426.
- Murias, P., de Miguel, J. C., & Rodriguez, D. (2008). A composite indicator for university quality assessment: The case of Spanish higher education system. *Social Indicators Research*, 89(1), 129–146.
- Murmura, F., Casolani, N., & Bravi, L. (2016). Seven keys for implementing the self-evaluation, periodic evaluation and accreditation (AVA) method, to improve quality and student satisfaction in the Italian higher education system. *Quality in Higher Education*, 22(2), 167–179.
- Nicolaou, M., & Atkinson, M. (2019). Do student and survey characteristics affect the quality of UK undergraduate medical education course evaluation? A systematic review of the literature. *Studies in Educational Evaluation*, 62, 92–103.
- Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education*, 15(2), 178–191.
- Pickford, R. (2013). Leadership of the National Student Survey for enhancement. *Quality Assurance in Education*, 21(4), 344–358.
- Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Raponi, V., Martella, F., & Maruotti, A. (2016). A biclustering approach to university performances: An Italian case study. *Journal of Applied Statistics*, 43, 31–45.
- Revelle, W., & Revelle, M. W. (2015). Package "psych". *The Comprehensive R Archive Network*, 337, 338.
- Seyfried, M., & Ansmann, M. (2018). Unfreezing higher education institutions? Understanding the introduction of quality management in teaching and learning in Germany. *Higher Education*, 75, 1061–1076.
- Shavelson, R. J. (1979). Applications of cluster analysis in educational research: Looking for a needle in a haystack. *British Educational Research Journal*, 5(1), 45–53.
- Spence, C. (2019). 'Judgement' versus 'metrics' in higher education management. *Higher Education*, 77, 761–775.
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598–642.
- Spooren, P., Mortelmans, D., & Christiaens, W. (2014). Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from

- confirmatory factor analysis and G Theory. *Studies in Educational Evaluation*, 43, 88–94.
- Spooren, P., & Christiaens, W. (2017). I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores. *Studies in Educational Evaluation*, 54, 43–49.
- Stehle, S., Spinath, B., & Kadmon, M. (2012). Measuring teaching effectiveness: Correspondence between students' evaluations of teaching and different measures of student learning. *Research in Higher Education*, 53, 888–904.
- Steinhardt, I., Schneijderberg, C., Götze, N., Baumann, J., & Krücken, G. (2017). Mapping the quality assurance of teaching and learning in higher education: The emergence of a specialty? *Higher Education*, 74(2), 221–237.
- Sutherland, D., Warwick, P., & Anderson, J. (2019). What factors influence student satisfaction with module quality? A comparative analysis in a UK business school context. *International Journal of Management Education*, 17(3), 100312.
- Szymanski, D. M., & Henard, D. H. (2001). Customer satisfaction: A meta-analysis of the empirical evidence. *Journal of the Academy of Marketing Science*, 29, 16–35.
- Tatro, C. N. (1995). Gender effects on student evaluations of faculty. *Journal of Research & Development in Education*.
- Thomas, N. (2018). An analysis of program evaluation course content in CSHSE-accredited human services baccalaureate programs. *Studies in Educational Evaluation*, 59, 187–194.
- Titus, J. J. (2008). Student ratings in a consumerist academy: Leveraging pedagogical control and authority. *Sociological Perspectives*, 51, 397–422.
- Tsiligiris, V., & Hill, C. (2021). A prospective model for aligning educational quality and student experience in international higher education. *Studies in Higher Education*, 46(2), 228–244.
- Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ*, e3299.
- van der Heijden, P. G., Teunissen, J., & van Orlé, C. (1997). Multiple correspondence analysis as a tool for quantification or classification of career data. *Journal of Educational and Behavioral Statistics*, 22(4), 447–477.
- Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, 23, 191–212.
- Wiley, C. (2019). Standardised module evaluation surveys in UK higher education: Establishing students' perspectives. *Studies in Educational Evaluation*, 61, 55–65.
- Worthington, A. C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education. *Assessment & Evaluation in Higher Education*, 27, 49–64.
- Yin, H., Wang, W., & Han, J. (2016). Chinese undergraduates' perceptions of teaching quality and the effects on approaches to studying and course satisfaction. *Higher Education*, 71(1), 39–57.
- Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, 12, 55–76.