5th International Conference on Higher Education Advances (HEAd'19)
Universitat Politècnica de València, València, 2019
DOI: http://dx.doi.org/10.4995/HEAd19.2019.9108

Is the Italian student survey on teaching reliable? For what purposes?

Enrico Zaninotto
Department of Economics and Management, University of Trento, Italy.

Abstract

Italian universities submit a compulsory survey to students for the evaluation of teaching activities. The questionnaire, designed by the Italian National Agency for University and Research System Evaluation (ANVUR), aims to evaluate four dimensions of teaching quality (course, instructor, personal interest and overall satisfaction) through twelve questions on a four-level scale. This paper addresses first the issue of the questionnaire's reliability in the representation of the four evaluation dimensions. The main result is that the questionnaire do not represent properly the four dimensions of evaluation which it is intended for. Secondly, through a preliminary statistical analysis, it discusses the use of the survey for comparative purposes. A comparative analysis can be adversely affected by several contextual and subjective factors, like the size of the class and the gender of the instructor. The paper concludes by discussing the difficulty of finding proper conditioning, and raises doubts regarding an uncritical comparative use of student evaluations of teaching activities.

Keywords: *Teaching assessment; Students surveys*

1. Introduction

Since the introduction of the national system of quality assurance, Italian universities have been obliged to submit to students an evaluation questionnaire for all teaching activities (courses, lab lesson and other activities run in the classroom). The basic structure of the questionnaire is defined by the National Agency for University and Research System Evaluation (ANVUR) (ANVUR, 2013).

The results of the survey are used differently by different universities. Some choose to make all the results for each teaching activity or each instructor public, while others present their results aggregated by degree only. Moreover, there are cases in which results are used in incentive schemes or are used to evaluate promotion and tenure.

After some years of widespread collection of data, it is consequently of the utmost importance to answer questions such as: 1) What is the value of the information in the current questionnaire? 2) What use could there reasonably be for the results, and in particular, 3) Is it possible to make individual comparisons aimed at orientating students' choices or official board decisions on tenure, promotion or incentives?

To this aim, this paper uses surveys collected from students at the University of Trento who attended between 2013 and 2016. The questionnaire seems only partially able to represent the dimensions of evaluation for which it was designed. Moreover, we raise doubts on their comparative use for the purposes of student orientation and faculty policies. Without a valid set of controls, comparisons between courses activities and teachers can be strongly biased. Instead, the survey can be an invaluable tool for self-assessment and teaching quality enhancement.

The paper proceeds as follows. First, it presents the current survey and the dataset. Then, in Section 3, an analysis of the questionnaire's reliability is conducted. Section 4 offers some hints on the interpretations of the students' assessments of the teaching activities. The concluding section draws upon several consequences for the use of the survey.

2. The Italian survey assessing the quality of teaching activities

Prior to 2013, several Italian universities independently carried out surveys on students' evaluations of teaching activities and instructors. Since then, ANVUR has issued guidelines for quality assurance, including the compulsory introduction of a survey of students' teaching evaluations, as well as a common framework for the questionnaire. ANVUR proposes two questionnaires; one submitted to attending students, and the other to non-attending students. (In Italy attendance at lessons is not compulsory).

The compulsory set of questions proposed by ANVUR is presented in Table 1. Answers are given on a four-grade scale: completely negative, rather negative, fairly positive and completely positive. At the end of the questionnaire, an open field is dedicated to free comments. Universities are free to add further questions. They can also decide how to submit the questionnaire to students—whether to make answering compulsory and whether to collect it in-class or online—and how to publish results, either aggregating them by degree or showing in detail the scores for individual teachers.

Table 1. List of questions (attending students)

Question Code	Question
D01	Was your preliminary knowledge of the course topics sufficient for your understanding of the course?
D02	Is the workload reflected proportionately in the number of credits the course offers?
D03	Is the teaching material (as presented and made available by the teacher) suitable for studying the topic?
D04	Are the assessment methods clearly defined?
D05	To what extent were the class schedules and other learning activities respected?
D06	Is the teacher able to stimulate/motivate interest towards the subject of the course?
D07	Does the teacher explain the subject in a clear manner?
D08	Are the integrative didactic support activities (tutoring, lab sessions and testing) useful to better understand the subject?
D09	Is the course content consistent with the course presentation as described in the syllabus published on the website?
D10	Is it easy to contact the teacher to obtain further explanations and clarifications?
D11	Are you interested in the subjects taught during the course?
D12	On the whole, are you satisfied with this course?

What follows, is a discussion of the use made of the questionnaires collected among attending students between 2013 and 2016 from the University of Trento. At this university, the survey is conducted online through the digital platform for student services. A student can fill out the questionnaires two thirds into the term and before registering for the exam. Then, they state whether they are an attending or non-attending student. Subsequently, the form is opened and the student completes it. Despite being available from two thirds into the term, the vast majority of students fill out the form immediately before registering for the exam.

The dataset employed consists of 274,000 questionnaires collected between 2013 and 2016 in 6,900 instances or combinations of teaching activities, instructors and years. The same teaching activity could indeed be performed by more than one instructor, and the same instructor was usually involved in more than one course (or other teaching activity).

Unfortunately, due to privacy regulations and with the aim of offering the full assurance of the respondents' anonymity, individual identifiers were automatically eliminated. It is thus impossible to link answers to individual characteristics and information on the students' former careers. This strongly limits the scope of the analysis.

3. Questionnaire reliability

ANVUR divides the questionnaire into four groups of items. D01–D04 are about the teaching activities, D05–D10 evaluate the instructor, D11 is about personal interest and D12 is an overall indicator of the student's satisfaction.

From pairwise correlations, it appears that D12 is, in reality, strongly correlated with the answers to questions D06 and D07. Apparently, the overall evaluation of the learning experience, strongly depends on the teacher's skills and competences.

For a first appraisal of the questionnaire, as a tool able to capture the three evaluation dimensions listed by ANVUR (teaching activity, instructor and personal interest and leaving aside overall satisfaction), a principal component analysis (PCA) was conducted. To this aim, D12 (for the reasons just mentioned) and D08, which concerns integrative activities, have been discarded.

Table 2 presents the results of the PCA, performed using a Varimax rotation on the first three components.

Table 2. Principal component analysis of the answers. Rotation: orthogonal Varimax; Rho = 0.6491; Number of obs = 274,127. Variance explained by the first three components.

Component	Variance	Difference	Proportion	Cumulative
Comp1	2.78741	0.615719	0.2787	0.2787
Comp2	2.17169	0.639526	0.2172	0.4959
Comp3	1.53217	.	0.1532	0.6491

The first three components explain approximately 65% of the variance (Table 2). From loading factors, the three retained components are associated to the questions as follows:

- *Component 1*: depends mostly on questions D4, D5, D9 and D10. It can be interpreted as an evaluation of the organisation of the teaching activities (the schedule, the clear definition of the content and the assessment, and the availability of the teacher during office hours)

- *Component 2*: depends on questions D11, D6 and D7, and can be interpreted as the perceived quality of teaching (the clarity of the teacher and the student's interest in the subject)

- *Component 3*: depends on questions D1 and D2. Here, the main point is the relationship between the student and the teaching activity (the adequacy of the background and the perceived workload)

Two conclusions can be drawn from this first part of the analysis. First, the set of questions seems able to give a comprehensive representation of different dimensions of the evaluation. The information added by each question is not negligible (the answers are not strongly correlated). However, the overall appraisal (D12) for a teaching activity evaluation must be used with care, as it is particularly sensitive to a few dimensions of quality. Second, the grouping of questions proposed by ANVUR does not correspond with what emerges from the data. The principal component analysis statistically groups different questions, and the interpretation of components highlights slightly different dimensions of quality evaluation—organisation, teaching and the relative position of the student with respect to the teaching activity—which, according to this analysis, better interpret students' subjective perceptions of quality.

4. At the roots of the students' answers

Students' questionnaires offer useful information to instructors and to the staff in charge of running degree. However, to understand better how to use the results, it is important to determine the sensitivity of the answers to different conditions. This is especially important if students or staff have the goal of comparing teaching performances. To be significantly compared, evaluations must be conditioned to relevant control variables in order to make comparisons between teachers or courses as similar as possible.

Unfortunately, as has already been said, very little information on respondents is available. For this reason, in this section it is possible to give only some tentative assessments of the problems of a comparative use of the observed results.

A first exercise that has been carried out is an analysis of variance. The total variability of answers can be decomposed in two components. The between variance is the variation that can be attributed to the difference among teaching activities, and the within variance is due to the variability of the evaluations among the students attending a given teaching activity.

It is clear that the larger the between component, the more meaningful it is to compare the teaching activities and teachers. On the other hand, if the within variance is large with respect to total variance, students evaluate the same teaching activity very differently.

To this aim, Question D12, which reflects the overall perception of the quality of the teachers and teaching activities, has been considered. The between effect covers only 30% of the standard deviation.

This means that a large part of the difference among the evaluations of teaching activities is due to the ways that participants of the same activity perceive the teaching experience. The underlying heterogeneity may hamper comparability of teaching evaluations (Bertoni et al., 2017). This could be related either to objective factors, such as the class composition in terms of background and former training, or to subjective factors, such as the way that the students interpret the grading scale. Factors such as internal scale, anchoring and recency affect the evaluation and use of grading scales. If there are factors that systematically influence the composition of the classes, and students self-sort in different courses, a comparison between the course evaluations would be extremely problematic.

As already stated, there are no data on respondents that would make this analysis possible. However, as a first step in this direction, it is possible to consider some observable variables within the teaching environment which can influence the perception of quality.

To this aim, we move from the consideration of individual questionnaires to their aggregation by single instances (a combination of a teaching activity, instructor and year). There are almost 6,900 observations of these instances. Each record contains, for each question, the number of answers in the 1–4 scale as well as some additional information.

A linear regression has been run. The dependent variable is the share of the positive answers on the total answers (sum of the fairly positive and absolutely positive answers). The independent variables are: the number of attending students (total of the collected questionnaires); the gender of the instructor; the level of the degree; the distinction between the Bachelor's degree (L), Master's degree (M), and full five-year degree (U); the department; the average grade; and the year.

Table 3 shows the results of the descriptive regression. It can be seen that, as expected, the number of students is negatively associated to the share of positive evaluations. Five-year degrees have a negative impact on the evaluation, while there is not a great difference between the teaching activities' evaluation of the Bachelor's or Master's degrees. The average grade in the class is positively correlated with the evaluation. This is somewhat surprising; students evaluate the activities and teachers before the exam. A positive correlation could result either from the fact that better teaching improves performance, or that the knowledge which students have on the attitudes of the instructor/examiner leads

them to assign higher evaluations to more generous examiners. A similar result is found by Braga et al. (2014:81) which observe that "teachers of classes that are associated with higher grades in their own exam receive better evaluationfrom their students".

Particularly surprising is the effect of gender. Female teachers have significantly worse evaluations than males. Whether this reflects worse performance (due, for instance, to the work overload of women, who are often tasked with family duties more than men) or an effect of stereotyping could be a matter of discussion.

Table 3. Linear regression model for the share of positive evaluations (fairly positive plus absolutely positive answers over total answers

Dependent variable: share of positive answer to Q12	**Coef.**	**Std. Err.**	**t**	**P>|t|**
Ave.Grade	0.001237	0.000356	3.47	0.001
N. Students	-0.00044	5.17E-05	-8.45	0
Instructor's gender (base: F)				
M	0.020165	0.00462	4.37	0
Degree level (base: Master)				
B	0.05049	0.01307	3.86	0
U	0.05462	0.012959	4.21	0
Constant	-5.6775	3.567168	-1.59	0.112

5. Conclusions

This paper has shown that the compulsory introduction of a student evaluation questionnaire, proposed by the Italian evaluation agency ANVUR, could be an important tool for quality enhancement. The questions are informative, but the analysis has suggested a different aggregation of questions than that proposed by the agency, as well as a partially different interpretation of some individual items.

In the second part, the paper presented preliminary evidence on the difficulty of a comparative use of student evaluations of teaching activities. They can derive from several subjective perceptions and context conditions which can influence the way students evaluate the activity and the instructor. Without a precise knowledge of the conditioning factors, the analysis suggests that a simple, unconditioned comparison of evaluations presents several risks (Stark and Freishtat, 2014) . The size of the class, the gender of the

instructor and the average assessment score are among the variables that influence student evaluation.

References

ANVUR (2013). *Autovalutazione, valutazione e accreditamento del Sistema universitario italiano*. ANVUR, Release 9 January 2013.

Bertoni, M., Rettore, E. & Rocco, L. (2017). How Informative Are Students' Evaluations of Teachers?, IRVAPP Seminar, June 2017.

Braga, M., Paccagnella, M. & Pellizzari, M. (2014). Evaluating students' evaluation of professors. *Economics of Education Review*, v. 41, 71-88. doi: 10.1016/j.econedurev.2014.04.002.

Stark, P. & Freishtat, R. (2014). An Evaluation of Course Evaluations. https://www.scienceopen.com/document_file/6233d2b3-269f-455a-ba6b-dc3bccf4b0a8/ScienceOpen/teachEvalSciOpen14.pdf

Torres, A., David, F., Graça, M. (2017). Quality assurance of teaching and learning: validity and usefulness of student ratings. *12th European Assurance Quality Forum*, Riga 23-25 november 2017. https://eua.eu/component/attachments/attachments.html?task=attachment&id=1069