Issue nº SS05-006.
7 March 2022

Can student evaluations be trusted?

Karl Kilbo Edlund

Karl Kilbo Edlund is currently a PhD student in environmental epidemiology at the University of Gothenburg. He has previously been president of the University of Gothenburg Student Unions and vice president of the EUTOPIA Student Council.

Paper published by EUTOPIA Student Think Tank (eustt.org) in it's aim of giving voice to european students' shared knowledge.

Abtract

Although nearly ubiquitous in the modern higher education sector, student evaluations of teaching, or course evaluations, have been frequently criticised in scientific literature. From a student's vantage point, these feedback systems represent one of the most commendable ideals of modern pedagogy: the involvement of student voices in course design and development. Therefore, while they are most often (and rightly so) at the periphery of a student's mind, student evaluations become a cornerstone of students' influence over their education. Thus, any doubts of their efficacy, accuracy, or justness should be of great concern. On the one hand, if these doubts were true, course evaluations would represent a major flaw in quality assurance and improvement systems, as well as a false promise of student influence. On the other hand, if these doubts were unfounded, they would nonetheless undermine the value of student input to potential improvements in education.

Keywords

Student evaluation, course evaluation, quality assurance, higher education

Introduction

Student evaluations of teaching have a long history in higher education, one of the earliest being implemented almost a century ago at the University of Washington in 1925 (Guthrie, 1953). This history can be traced in parallel to the rapid expansion of higher education over the last century, but also to the increased societal demand and interest in systematic quality assurance and improvement in general. For instance, in Sweden, course evaluations in higher education are formally recommended to universities since 1970 (Högskoleutredningen, 1992) and mandated by law since 2000 (Svensk författningssamling, 2000).

As course evaluations and their likes have assumed important roles in the inner workings of universities, they have also become an active subject of scientific study. One of the earlier publications in the field succinctly outlined the motivations for performing student evaluations in its opening words: “Students agree quite well on what they believe are important features of good teaching and their judgments provide a valuable measurement of teaching effectiveness” (Guthrie, 1953).

However, evaluating the validity of the results obtained through course evaluations is surprisingly difficult. There is no “golden standard” of course quality to which the course evaluations can be compared—for the obvious reason that quality is situational, often subjective, and inherently multi-dimensional—, nor are all course evaluations created with the same purpose in mind. Instead, different purposes raise entirely different questions of validity and potential biases. Five possible main purposes of course evaluations can be identified (Högskoleutredningen, 1992):

1. To provide some form of grade of the course as information for prospective students
2. To assess teacher competence for processes like pay determination, formal pedagogical merits, division of tasks, or skill development
3. To allocate resources between courses
4. To create an opportunity for students to reflect on their own learning process
5. To self-evaluate the course design in order to inform future course development

When mandated by legal or university regulations, course evaluations may be performed routinely without much care for their specific purpose. Yet the purpose of an evaluation largely determines the kind of responses it receives. It affects everything from how the questions are asked to the way in which respondents consider their answer, and purposes are therefore considered to be largely mutually exclusive (Högskoleutredningen, 1992).

In this text, the focus will be primarily on course evaluations conducted for the last of the purposes listed above: course development. Reviewing available evidence, the text will explore whether input received from student evaluations can be considered a valid reflection of inherent educational quality, and thus if it should be considered a trustworthy basis, among others, for course development.

The relationship between student learning and satisfaction

In the absence of a "golden standard", one way of approaching the question of the validity of student evaluations as a measure of educational quality or teacher ability has been to measure the stability of student evaluation results over time and across student groups. In this regard, it has been noted that sufficiently large groups of students (i.e. more than approximately ten students) tend to give fairly stable average scores to courses taught by the same instructor over time (Richardson, 2005). The same study also noted that students performing better on end-of-course exams tend to give higher ratings. These, however, were not at all (or only weakly) biased because of extraneous factors, such as prior interest, expected grade, course workload, and reasons for taking the course. A similar study (Liaw and Goh, 2003) conducted at University of Malaya found only one source of bias in course evaluation: class size, which, if seen from a student point of view, may well be a quality in itself.

Another prominent way of assessing the validity of student evaluations in scientific literature has been to study the correlations between course ratings and average student achievements, under the premise that the students' average academic performance, with or without adjustments for previous knowledge, is the primary marker of educational quality. The

flaw in this premise is, however, obvious: if the results of student evaluations were only a proxy for students' academic results, conducting them in the first place would not make much sense, as they can hardly replace ordinary exams. The results of these studies have been inconsistent and have also led to various different interpretations. While strong correlations between test performance and course quality have often been interpreted as high validity (i.e., students perform well on exams because of high educational quality), others have instead interpreted strong correlations as a significant bias from students favouring courses with more lenient grading (i.e., students rating courses highly if they are evaluated through comparably easy exams).

The latter is specifically one of the fears frequently raised in literature: that course evaluation favours courses with lenient grading. However, results of the research on this point have been inconclusive. One review concluded that the effect of lenient grading has been found to be small and not always positive (D'Apollonia and Abrami, 1997). Another author commented that "in many of these studies, the researchers made no effort to determine whether the students had earned the higher grades", thus noting that high-quality courses would also be likely to result in the combination of high grades and positive student evaluations (Barth, 2008). These have been dubbed the "leniency hypothesis" and the "validity hypothesis", respectively (Wang and Williamson, 2020).

Factors underlying student responses

The multidimensionality of course evaluations makes it very difficult to estimate their validity from an average of students' overall impression of a course or instructor. One way to overcome this is factor analysis, which has been suggested and applied to evaluation results by another line of research. Factor analysis is a statistical method that aims to disentangle multidimensional data into 'principal components', i.e., underlying dynamics that shape students' responses.

For example, a large-scale factor analysis performed on student evaluations from more than 30,000 students at the University of Southern California between 1976 and 1980 indicated that the nine core questions of the standardised SEEQ evaluation tool corresponded to nine separate factors, which were both stable across courses and course instances, and

had little overlap between each other (Marsh and Hocevar, 1984). In other words, these results indicated that the students were in fact able to consistently discriminate between the nine different aspects evaluated, and that their opinions on each aspect had little influence on their responses to questions regarding other aspects.

Another factor analysis, performed on over 14,000 individual student evaluations (n.b., using an entirely different questionnaire) from the Jordan Valley Regional College, indicated the presence of two main factors (Cohen, 2005). In this case, whereas the first factor primarily consisted of questions about course content, bibliography, clarity of the syllabus, and the intellectual tools for independent analysis that students believed that they had gained, the second factor looked primarily at questions on teacher preparedness, their relationship with students, and their response to students' questions. The factors overlapped in questions about course organisation, clarity of presentation, general assessments of course quality and, to a certain degree, their general assessment of teacher ability. This indicates that students could make a clear distinction between the teacher's ability and the qualities of the course and identify areas that they considered to be affected by both. While this indicates that students see teacher and course design as entangled aspects, it makes the interpretation of such summary questions difficult, which are indeed the basis of most studies using correlation to assess the validity of student evaluations.

A third factor analysis study, this time using student evaluation results from the Georgia Southern University, identified five different factors with otherwise similar results to the two former studies (Barth, 2008). Interestingly though, this study explored the effect of expected and actual grades, which were included in the factor analysis, in addition to the twelve original questions. The results clearly placed grade as a separate factor. However, it showed a positive interaction between grades and how well students found that examinations reflected the course content (i.e., receiving a higher grade was associated with finding that examinations reflected course content to a higher degree), and negative interactions with course difficulty and with how intellectually challenged students found that they had been by the course. Lastly, a multiple regression analysis showed that students' grades had an effect on their overall impression of the teacher's abilities, but that this effect was much smaller than the perceived quality of instruction.

Conclusion

Both primarily different strands of research discussed above indicate that students are remarkably able to distinguish between different factors of quality, and that the validity of student evaluations, when used appropriately, may be high. Of course, there are more aspects to consider, for example equity bias (e.g., containing gender or ethnic dimensions) or sampling bias (comprising the intensively researched question of response rates), which are of great importance to the implementation, design, and interpretation of course evaluations.

As the studies referenced above have underlined, the questions used determine their answers—i.e., in responding to a question, the student will consider whatever they perceive to be relevant. Student evaluations thus show the students' perception of the aspects they believe they should consider. In this way such questionnaires reflect the interplay between the institution's and the students' definitions of quality. The fact that these sometimes differ does not necessarily signify that the instrument of evaluation lacks validity. Specifically, questions asking for overall teacher ability will, naturally, include situational factors that affect how teachers are able to put their pedagogical abilities to use, for instance class size, perceived utility of the subject matter to the students or its relationship to previous and upcoming courses.

Whether the correlation between student satisfaction and their grades, as reported by some studies, is indicative of a positive effect of lenient grading on student satisfaction or on the achievement of highly skilled teachers held in high esteem by their students is a question impossible to answer. It remains unanswered because we lack a measure of student learning that is independent of grading. Moreover, this association is presumably highly dependent on context as well as individual factors and is likely to be bi-directional. The studies reviewed in this text suggest that the association between final grades and course perception is relatively weak, and even more so in questions that prompt the student to consider specific factors of the quality of their education rather than their overall perceptions. To discontinue the use of student evaluations would thus mean losing important and valid input that could contribute to quality improvement. Instead, what we might need is a greater awareness of the importance of both a good course evaluation design and the limitations inherent in all surveys—as well as maybe a little more trust in students' abilities to value the quality of the education they partake in.

Bibliography

Barth, M. M. (2008) 'Deciphering Student Evaluations of Teaching: A Factor Analysis Approach', *Journal of Education for Business*, 84(1), pp. 40–46. doi: 10.3200/JOEB.84.1.40-46.

Cohen, E. H. (2005) 'Student evaluations of course and teacher: Factor analysis and SSA approaches', *Assessment and Evaluation in Higher Education*, 30(2), pp. 123–136. doi: 10.1080/0260293042000264235.

D'Apollonia, S. and Abrami, P. C. (1997) 'Navigating student ratings of instruction', *American Psychologist*, 52(11), pp. 1198–1208. doi: 10.1037/0003-066X.52.11.1198.

Guthrie, E. R. (1953) 'The evaluation of teaching', *Assessment & Evaluation in Higher Education*, 53(2), pp. 220–221. doi: 10.1080/0260293840090201.

Högskoleutredningen (1992) *Frihet, ansvar, kompetens: Grundutbildningens villkor i högskolan (SOU 1992:1)*. Stockholm: Allmänna förlaget.

Liaw, S. H. and Goh, K. L. (2003) 'Evidence and control of biases in student evaluations of teaching', *International Journal of Educational Management*, 17(1), pp. 37–43. doi: 10.1108/09513540310456383.

Marsh, H. W. and Hocevar, D. (1984) 'The Factorial Invariance of Student Evaluations of College Teaching', *American Educational Research Journal*, 21(2), pp. 341–366. doi: 10.3102/00028312021002341.

Richardson, J. T. E. (2005) 'Instruments for obtaining student feedback: a review of the literature', Assessment & Evaluation in Higher Education. Informa UK Limited, 30(4), pp. 387–415. doi: 10.1080/02602930500099193.

Svensk författningssamling (2000) Förordning om ändring i högskoleförordningen (1993:100).

Wang, G. and Williamson, A. (2020) 'Course evaluation scores: valid measures for teaching effectiveness or rewards for lenient grading?', Teaching in Higher Education. Taylor & Francis, 0(0), pp. 1–22. doi: 10.1080/13562517.2020.1722992.