26

The Validity of Student Course Evaluations: An Eternal Debate?

Pamela Gravestock
University of Toronto

Emily Greenleaf
University of Toronto

Andrew M. Boggs
Higher Education Quality Council of Ontario

Student evaluation of courses and teaching at universities remains a highly contentious and divisive topic. Emotions and anecdotal evidence can overrule conclusions drawn from research on the validity and design of course evaluations. However, even amongst researchers, there is significant disagreement on the efficacy of course and teaching evaluations. This paper explores this ongoing dialogue through the medium of a parliamentary debate drawing from the breadth of current research on course evaluations.

Introduction

Student evaluation of courses and teaching is a contentious issue in higher education. Recently, Côté and Allahar (2007) went as far as to assert that professorial fear of student evaluations is a major contributing factor to rampant grade inflation across North America. Controversy centres on the perceived validity of student course/teaching evaluations: are students capable of providing accurate assessments of teaching ability and course content?

The answer to this question has practical implications. For faculty, student evaluations can influence promotion and tenure decisions. For students, evaluations may influence course selection and are often the only opportunity they have to provide feedback on the quality of instruction. Furthermore, these evaluations may be growing in importance in a public policy context increasingly concerned with the 'quality' of higher education.

This paper, based on a session given at the 2008 Society for Teaching and Learning in Higher

Education (STLHE) conference at the University of Windsor, provides an overview of research on student course/teaching evaluation validity, including information about instrument development, interpretation and factors often understood to influence evaluation results. The session presentation, and this paper, are both drawn from a larger research project undertaken on behalf and with the support of the Higher Education Quality Council of Ontario (HEQCO).[1]

The Great Debate

Since the assessment of teaching effectiveness is a contentious issue, it is not surprising that research in this area is equally divided. Consequently, we decided that our STLHE session would explore current research on this topic through the oppositional format of a parliamentary debate. We debated the resolution that: student course evaluations are a valid and reliable measure of teaching effectiveness for the purposes of summative evaluation. We invited session participants to consider the arguments and evidence presented, offer their own thoughts and experiences through 'speeches from the floor,' and vote for the argument they felt was more compelling through 'division of the house.' The modified format of our session may be found in Table 1.

We have reproduced both the Prime Minister/Government's and Leader of the Opposition's speeches below. We do not suggest that there is a clear 'winner' in this debate (although the result of the vote during our conference session was against the resolution), but do point out that there is significant evidence and

Table 1
Format of the STLHE Session

Government's opening speech	Introduce resolution to be debated, outline government's argument and begin building its case.	5min
Opposition's speech	Response to resolution. Outline opposition's argument, respond to government's case and begin building opposition's case.	5min
Speeches from the floor	Opportunity for the honourable members of the assembled House to respond to the government and/or opposition's cases and/or put questions to either side.	10min
Opposition's closing remarks	Response to speeches from the floor and summary of opposition's case.	5min
Government's closing remarks	Response to speeches from the floor and summary of government's case.	5min
Division of the House	A simple call of 'yeah' or 'nay' will be used to measure the opinion of the House.	5min
Committee of the Whole	The speaker/chair is removed to allow for more unstructured discussion – a conventional question and answer session.	10min

[1] The complete research paper, *Student Course Evaluations: Research, Models and Trends* (Gravestock & Gregor-Greenleaf 2008), is available through HEQCO at http://www.heqco.ca/SiteCollectionDocuments/Student%20Course%20Evaluations.pdf

compelling argumentation on both sides of this issue.

GOVERNMENT (opening remarks)

Be it resolved that student course evaluations are a valid and reliable measure of teaching effectiveness for the purposes of summative evaluation.

Mr. Speaker, this resolution must stand.

There is general and long-standing agreement in the research that course evaluation instruments can be, and most often are, reliable tools for measuring instructional ability in that they provide consistent and stable measures for specific items (e.g., an instructor's organizational skills or relative workload). This is particularly true when the tool is carefully constructed and psychometrically tested before use (for examples, see Abrami, 2001; Theall & Franklin, 2001; Wachtel, 1998; Goldschmid, 1978; Marsh & Roche, 1997; and McKeachie, 1997).

Since the 1970s, scholars have been seeking to identify characteristics that bias student evaluation ratings – studies have focused on administrative conditions, course, instructor, and student characteristics. However, in 40 years of research, nothing has been identified that significantly impacts ratings. As Greenwald (1997) notes in his review of the research, the majority of publications produced between 1975 and 1995 favoured validity. McKeachie (1997) argues that student course evaluations are the "single most valid source on teaching effectiveness" (p. 1218). Those who found course evaluations to be valid have shown that ratings data can be correlated to other evidence of teaching effectiveness such as evaluations from colleagues or trained faculty development personnel.

Issues such as class time, discipline, instructor rank and experience, student motivation, course level, and instructor enthusiasm do have a small, but measurable impact on evaluation ratings. However, this impact does not reflect bias but rather indicates valid shifts in teaching effectiveness. Moreover, they can be considered when ratings are interpreted.

The research does show that there is a positive correlation between grades and student ratings. Some instructors interpret this to mean that lenient grading practices can produce inflated ratings. However, Wachtel (1998), Marsh and Dunkin (1992), Murray (1987), and others argue that this positive correlation is simply evidence of student learning: students rate faculty more positively when they have had a positive classroom experience.

Anecdotal evidence also suggests that faculty who assign more course work are penalized by students with low ratings. However, a study by Heckert, Latier, Ringwald-Burton, and Drazen (2006) found that higher evaluations were given to courses in which the difficulty level was viewed as appropriate but were also positive when students indicated they had expended more effort than anticipated. Overall, this study concludes that more demanding instructors received higher evaluations and therefore refutes the grading leniency hypothesis, and the notion that faculty could 'buy' better evaluations with higher grades.

Several decades of research destroy these and countless other myths and misperceptions regarding the validity of student course evaluations. For example, many call into question the ability of students to accurately evaluate teaching effectiveness, arguing that they are not reliable assessors. Studies dating back to the 1970s consistently demonstrate this to be false and show that students are reliable and effective at evaluating teaching behaviours (e.g., presentation, clarity, organization, and active learning techniques), the amount they have learned, the ease or difficulty of their learning experience in the course, the workload in the course, and the validity and value of the assessment used in the course (Nasser & Fresko, 2002; Theall & Franklin, 2001; Ory & Ryan, 2001; Wachtel, 1998; Wagenaar, 1995). Scriven (1997) argues that students are "in a unique position to rate their own increased knowledge and comprehension as well as changed motivation toward the subject taught. As students, they are also in a good position to judge such matters as whether tests covered all the material of the course" (p. 2).

Another persistent myth suggests that ratings reflect instructor popularity or personality. The now famous "Dr. Fox" study from the 1970s, which concludes that an instructor's enthusiasm or personality can impact evaluations, is widely refuted and discounted on methodological grounds. Ory (2001) argues that "personality" may actually measure teach-

ing behaviours, such as enthusiasm, that may in fact influence teaching effectiveness.

Mr. Speaker, let me mention one final myth, not supported by the research: the majority of faculty object to the use of student course evaluations. Studies demonstrate that this is not the case; rather, a high percentage of faculty possess positive attitudes toward this tool.

OPPOSITION (opening arguments and rebuttal)

Mr. Speaker, let me clearly state that I concede all of the government's points. I agree that course evaluation instruments offer reliable data and valid measurements of the questions on the forms.

I do not, however, concede the resolution. Rather, I argue that the government has not presented a sufficient perspective of validity. As the government has proven, evaluation forms can and have been developed that adequately pre-empt the influence of any external, biasing factors. However, this internal validity is meaningless if the forms are improperly constructed or used – if student ratings have insufficient construct and consequential validity. I will argue that current course evaluation practice does not provide these types of validity, and that, consequently, student course evaluations do not provide a valid measure of teaching effectiveness for the purposes of summative evaluation.

Ory and Ryan (2001) note that "to make valid inferences about student ratings of instruction, the rating items must be relevant to and representative of the processes, strategies, and knowledge domain of teaching quality" (p. 32). For course evaluations to be valid measures of teaching effectiveness, not only must the questions reflect those aspects of teaching identified as effective, but the very definition of effective teaching must be identified and agreed upon – but, as Ory and Ryan conclude, no "universal set of characteristics of effective teachers and courses that should be used as a target…appears to exist" (p. 32). Furthermore, educational priorities vary by institution, discipline, and even course. By mandating a generic, prescriptive evaluation instrument, we ensure that evaluations are unresponsive to desired and inevitable variations in teaching styles and goals.

We cannot, therefore, develop an instrument that accurately assesses teaching effectiveness because we cannot yet identify universal, comprehensive, and stable measures of effective teaching.

Even if appropriate measures of teaching effectiveness could be identified – though I have just shown this to be impossible – there remains another insurmountable obstacle to course evaluation validity. This is the obstacle of the appropriate interpretation of course evaluation results by faculty and administrators. Menges (2000) argues that "a great many individuals in the assessment area would assert that no matter how valid and reliable the instrument is, consumers can and do misuse the results from it" (p. 8). According to Menges, this misuse, and consequent compromise to validity, can occur for two primary reasons:

1. Administrators frequently receive too much or too little data to properly read the forms. Individual scores on large numbers of questions present an overload of information; conversely, evaluation data is rarely accompanied by information providing a thorough contextualization of the data, including descriptions of course activities and goals.
2. Once they do receive the forms, users of course evaluation data are unclear about the statistical value of evaluation results, often overestimating the significance of, for example, the difference between a rating of 3.5 and one of 3.7 on a 5-point scale. Administrators interpreting the data can not articulate a meaningful distinction between these two scores, and yet are pleased to report that the instructor with a score of 3.7 is a "better" instructor. These statistical challenges are amplified when such comparisons are made across diverse courses or disciplines.

For these insurmountable obstacles to the validity of course evaluations introduced during the construction of evaluation instruments and the interpretation of evaluation data, Mr. Speaker, I must reiterate my assertion that student course evaluations are not valid indicators of teaching effectiveness.

OPPOSITION (closing arguments)

Mr. Speaker, let me once again reiterate that I agree with the Government that teaching evaluations are quite effective at measuring what they seek to measure. I argue, however, that this is a minor, even meaningless determinant of their validity. Until we can agree on a universal set of effective teaching characteristics, or a universally effective way of organizing and presenting course content, we cannot develop evaluation instruments that can effectively capture the infinite varieties of effective teaching and risk, as McKeachie (1997) states, "penaliz[ing] the teacher who is effective despite less than top scores on one or more of the dimensions" (p. 1218) of teaching measured on evaluations.

At the other end of the evaluation process are the threats to validity introduced in the interpretation of evaluation results by users who overestimate the precision of evaluation data and fail to properly contextualize student ratings according to the particular circumstances, characteristics, and intentions of individual courses and instructors. For these reasons, Mr. Speaker, I must restate my strong belief that student course evaluations are not valid measures of teaching effectiveness for the purposes of summative evaluation.

GOVERNMENT (closing arguments)

Mr. Speaker, my esteemed colleague raises many interesting and relevant issues that institutions should bear in mind when developing course evaluation systems; however, let me recall that the most essential issue here is that of bias. As numerous empirical studies have shown, this can be addressed through instrument design, question selection, administration, implementation, and education about interpretation. As Abrami (2001), Franklin (2001), Theall and Franklin (1989, 2001), Kulik (2001) and others note, and we fully agree, education helps to ensure that when data is used for summative purposes, decisions are fair and equitable.

The issues raised by my colleague do not point to any invalidity in the course evaluation instrument itself but rather to issues affecting the role of teaching in the university more generally and particularly for the evaluation of teaching for summative purposes, including tenure and promotion. Moore and Kuol (2005) argue:

> Given that it is an almost universal phenomenon that research activity reaps more individual rewards than those associated with teaching, efforts to measure the teaching related dimensions of [faculty] performance, and to pay attention to those measures in the context of an individual's professional development helps to create more parity of esteem between the teaching and research components of the academic role. (p. 143)

As such, course evaluations are an essential component to ensure the recognition of teaching in higher education. The quantifiability and comparability of course evaluations makes the imprecise art of evaluating teaching more objective and manageable. As Abrami (2001) argues, there is no other option that provides the same sort of quantifiable and comparable data.

All of this only highlights the need for greater attention to this area and the best way to do this is through the continued use of course evaluations.

Conclusion

During the speeches from the floor, many points were raised both criticizing the use of student course evaluations and supporting evaluations' proper use in an academic environment. Although seminar participants' comments were evenly split for and against the debate's resolution, when participants were given the opportunity to vote, the opposition carried the day by a large majority. It is difficult to explain why there was such a clear winner in this debate. It was apparent that some participants were inherently distrustful of student evaluations of courses and teaching and that even researched evidence could not dissuade them from longheld beliefs in popular myths and misperceptions about course evaluations. It is also possible that others may have been swayed by the argument that more work is needed before any teaching assessment tool

can be declared 'valid.' The varied opinions expressed on this topic during our presentation suggest that the debate over student course evaluations is far from being resolved.

References

Abrami, P.C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P.C. Abrami, & L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research 109*, 59-87.

Côté, J.E. & Allahar, A.L. (2007). *Ivory tower blues: A university system in crisis.* Toronto: University of Toronto Press.

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. In K.G. Lewis (Ed.), Techniques and Strategies for Interpreting Student Evaluations. [Special issue]. *New Directions for Teaching and Learning*, *87*, 85-100.

Franklin, J. & Theall, M. (1989). *Who reads ratings: Knowledge, attitude and practice of users of student ratings of instruction*. Paper presented at the American Educational Research Association annual meeting of the AERA, San Francisco.

Goldschmid, M.L. (1978). The evaluation and improvement of teaching in higher education. *Higher Education*, *7*(2), 221-245.

Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends*. Retrieved March 4, 2009, from the Higher Education Quality Council of Ontario Web site: http://www.heqco.ca/SiteCollectionDocuments/Student%20Course%20Evaluations.pdf

Greenwald, A.G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Pyschologist*, *52*(11), 1182-1186.

Heckert, T.M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to "buy" better evaluations through lenient grading? *College Student Journal*, *40*(3), 588-596.

Kulik, J.A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P.C. Abrami, and L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research*, *109*, 9-25.

Marsh, H.W., & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional approach. In J.C. Smart (Ed.), *Higher education: Handbook of theory and research: Vol. 8* (pp. 143-223). New York: Agathon Press.

Marsh, H.W. & Roche, L. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias and utility. *American Psychologist*, *52*(11), 1187-97.

McKeachie, W.J. (1997). Student ratings: The validity of use. *American Psychologist*, *52*(11), 1218-25.

Menges, R.J. (2000). Shortcomings of research on evaluating and improving teaching in higher education. In K.E. Ryan (Ed.), Evaluating teaching in higher education: A vision for the future [Special issue]. *New Directions for Teaching and Learning*, *83*, 5-11.

Moore, S., & Kuol, N. (2005). A punitive tool or a valuable resource? Using student evaluations to enhance your teaching. In G. O'Neill, S. Moore, & B. McMulline (Eds.), *Emerging*

issues in the practice of university learning and teaching (pp. 141-148). Dublin: All Ireland Society for Higher Education.

Murray, H.G. (1987). Acquiring student feedback that improves instruction. In M.G. Weimer, (Ed.), Teaching large classes well. [Special issue]. *New Directions for Teaching and Learning*, *32*, 85-96.

Nasser. F. & Fresko, B. (2002). Faculty view of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, *27*(2), 187-198.

Ory, J.C. (2001). Faculty thoughts and concerns about student ratings. In K.G. Lewis (Ed.), Techniques and strategies for interpreting student evaluations. [Special issue]. *New Directions for Teaching and Learning*, *87*, 3-15.

Ory, J.C. & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P.C. Abrami, & L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research*, *109*, 27-44.

Scriven, M. (1997). Student ratings offer useful input to teacher evaluations. Retrieved April 2, 2008 from *ERIC Digest* Website: http://www.ericdigests.org/1997-1/ratings.html

Theall, M. & Franklin, J. (2001) Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In M. Theall, P.C. Abrami, & L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research*, *109*, 45-56.

Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *29*(2), 191-121.

Wagenaar, T.C. (1995). Student evaluation of teaching: Some cautions and suggestions. *Teaching Sociology*, *23*(1), 64-68.

Biographies

Pamela Gravestock is the Associate Director of the Office of Teaching Advancement at the University of Toronto, Toronto, Ontario. She holds a masters degree in Art History and History and is currently a doctoral candidate in the Higher Education Group at Ontario Institute for Studies in Education (OISE)/University of Toronto. Her dissertation research focuses on the evaluation of teaching for tenure at Canadian universities.

Emily Greenleaf is the Research Associate and Faculty Liaison for the Office of Teaching Advancement and the Humanities and Social Sciences Coordinator for the Teaching Assistants' Training Program at the University of Toronto. She is currently completing her doctorate degree in Higher Education at Ontario Institute for Studies in Education (OISE)/University of Toronto with a dissertation on the history of the Canadian undergraduate curriculum.

Andrew M. Boggs is a past Research Director with the Higher Education Quality Council of Ontario (HEQCO) and holds a masters degree in Higher Education Policy and History from Ontario Institute for Studies in Education (OISE)/University of Toronto. He is currently pursuing a doctorate in Higher Education Policy at the University of Oxford, United Kingdom.