

TIME AND ENERGY WELL SPENT?  
REVIEWING STUDENT EVALUATIONS OF TEACHING

ERIC R. BJORKLUND

*This paper was completed and submitted in partial fulfillment of the Master Teacher Program, a 2-year faculty professional development program conducted by the Center for Teaching Excellence, United States Military Academy, West Point, N., 2008.*

Although the formal process of students evaluating their professors has existed since the early 1920's, the process continues to stir up debate, controversy, and the need for further research. The University of Washington pioneered this process by first asking their students to fill out questionnaires on their professors almost eighty-five years ago.<sup>1</sup> However, the expansion in the use of student feedback to evaluate teachers has only recently blossomed. In fact evidence shows that only 28% of universities asked for student feedback for teacher evaluations in 1973. The data by 1993, however, showed an increase to 86%.<sup>2</sup> With this clear trend in recognizing the perceived value of such student evaluations of teaching (SETs), the literature on the topic clearly shows that there is no agreement on the scope and limit of that value. This literature review will focus on three areas that are at the core of the current debate. First, the discussion of the possible utility of student feedback on instructors is summarized. Next, the validity argument of these SETs (whether or not they measure what they intend to measure) is discussed. Finally, this review will conclude with some of the main themes that currently find a large amount of scholarly consensus as well as some of the areas where further research is requested.

Within the teaching profession many different tools to measure effectiveness of instruction are used. Student evaluations of teaching are one form of measurement that is commonly used in American universities. However, it is

---

<sup>1</sup> James A. Kulik, "Student Ratings: Validity, Utility, and Controversy," in *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?*, eds. Michael Theall, Philip C. Abrami, Lisa A. Metz (San Francisco: Jossey-Bass Publishers, 2001), 9.

<sup>2</sup> Peter Seldin, "The Use and Abuse of Student Ratings of Professors," *The Chronicle of Higher Education*, July 21, 1993, A40.

important to determine whether or not these SETs are actually helpful in assessing the teacher's work in *and out* of the classroom. Luckily there is both significant testimonial and experimental evidence showing that SETs are useful in both assessing and improving teaching quality.<sup>3</sup> In general SETs are most beneficial when they are supplemented with consultation.<sup>4</sup> Institutional support in the form of taking the time to ensure the ratings are clearly explained, measured to a fixed standard, and compared longitudinally from past evaluations is critical to ensuring the utility of SETs.

While the above evidence supports the effectiveness of SETs, there are several areas in which SETs have more limited utility. A key point that Marsh and Roche make is that multidimensionality must be factored into the evaluations.<sup>5</sup> High enthusiasm is no substitute for a general lack of skill in teaching. However, if these are averaged in a single effectiveness score the results will show an average teacher. Thus, it is important to keep the different variables separate when viewing and consulting on the results. Finally, another interesting limitation to utility was noted in Basow and Silberg's analysis of gender comparisons in SETs. Surprisingly, female teachers scored lower on SETs on a series of comparison studies conducted in the late 1990's.<sup>6</sup> Causation of these results is not definitive, especially since several factors like students' major, students' perception of sex-atypical professions, and students' perception of availability were not or could not be isolated in the

---

<sup>3</sup> Herbert Marsh and Lawrence A. Roche, "Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility," *American Psychologist* 52(11) (Nov 1997): 1195.

<sup>4</sup> *Ibid.*, 1196.

<sup>5</sup> *Ibid.*, 1187.

<sup>6</sup> S.A. Basow and N.I. Silberg, "Student Evaluations of College Professors: Are Male and Female Professors Rated Differently?" *Journal of Educational Psychology* 79(3) (1987): 312.

survey. In term of utility, it is important to consider that some gender bias may be present in student evaluations of teaching.

In addition to determining whether or not SETs are useful in evaluating and improving teaching, much of the literature on the subject focuses on whether or not these tools measure what they intend to measure (validity). The studies related to this subject are typically broken down into two separate groups: validity based on students' opinions about the quality of instruction and validity based on whether they accurately reflect teaching effectiveness.<sup>7</sup> The former is normally accepted as a good way to measure student satisfaction while the latter's validity is typically critiqued as being unable to differentiate between the process and product of teaching.<sup>8</sup> However, determining good ways to measure teaching effectiveness is critical to evaluation and teaching improvement. Thus, much research has been done since to improve validity. McKeachie attempts to summarize some of the critical factors in determining the validity of SETs. Agreeing that SETs are the single most valid source of data on effective teaching, he does acknowledge that ratings and learning are not perfectly correlated.<sup>9</sup> Factors such as large course size and objective tests often influence the precision of teaching evaluations. From this data, McKeachie touches on one of the most important considerations when using SETs for evaluation of teaching and learning: teaching is an interactive art that can never

---

<sup>7</sup> Philip C. Abrami, Sylvia D'Apollonia, and Peter A. Cohen, "Validity of Student Ratings of Instruction: What We Know and What We Do Not," *Journal of Educational Psychology* 82(2) (1990): 219.

<sup>8</sup> Ibid.

<sup>9</sup> William J. McKeachie, "Student Ratings: The Validity of Use," *American Psychologist* 52(11), (Nov 1997): 1219.

be simplified to a listing of “most effective” techniques.<sup>10</sup> What works for one group of students may not be as effective to a second, yet similar group. Great teachers are able to respond to this.

Regardless of the debate over the limits of validity, SETs will remain an important part of the teaching and learning process in higher education. A final consideration on validity is discussed by Ory and Ryan. They accurately point out that there is no single standard to how student evaluations of teaching are administered, collected, or used as a measure of teaching and learning.<sup>11</sup> With this wide diversity, it is very difficult to make clear judgments on validity. Thus, more research with specifics on the context of the environment in which SETs are presented is an important step forward.

Within the literature on student evaluations of teaching, authors have tried to summarize some common themes found through the research as well as propose the areas that need additional focus for further research. One of the key areas that has achieved general consensus among the literature is that SETs generally agree with results from other forms of teaching effectiveness like learning measures, expert observations, and alumni ratings.<sup>12</sup> Another key observation is the consistently strong correlation between high grades and good student evaluations.<sup>13</sup> Finally, all student evaluations should follow some general guidelines to maximize both utility

---

<sup>10</sup> Ibid., 1223.

<sup>11</sup> John C. Ory and Katherine Ryan, “How Do Student Ratings Measure Up to a New Validity Framework,” in *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?*, eds. Michael Theall, Philip C. Abrami, Lisa A Mets (San Francisco: Jossey-Bass Publishers, 2001), 41.

<sup>12</sup> Kulik, 23.

<sup>13</sup> Anthony G. Greenwald and Gerald M. Gilmore, “Grading Leniency is a Removable Contaminant of Student Ratings,” *American Psychologist*, 52(11), (Nov 1997): 1210.

and validity. Institutions of higher education should ensure: evaluations have a stated purpose, produce reports that are easily understood, and continually reassess the evaluation system.<sup>14</sup> However, just as strongly as there is consensus on several aspects of using SETs in the classroom, there are other areas that experts in the field identify as needing additional research. Some of these areas that need a further look are the influence of body language, variety in vocal pitch, and the potential correlation with high ratings and low learning.<sup>15</sup>

Despite the areas that require more research, student evaluations of teaching provide strong and useful information on teaching effectiveness and for ways to improve our teaching. However, in order to maximize the potential from these tools of teaching and learning they must be used carefully, consistently, and supported with the necessary explanation and consultation. Finally, it is critical to understand that SETs are *one* important measurement tool in evaluating and improving teaching that must be combined with other means.

---

<sup>14</sup> Michael Theall and Jennifer Franklin, "Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction?" in *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?*, eds. Michael Theall, Philip C. Abrami, Lisa A Mets (San Francisco: Jossey-Bass Publishers, 2001), 52-54. The authors provide a longer list of guidelines. Include all stakeholders, publicly present clear information about the evaluation criteria and procedures, educate the users of the results to avoid misuse, keep a balance of individual and institutional needs, include resources for improvement, keep formative information confidential, adhere to rigorous measurement principles, establish a legally defensible process, and consider the combination of evaluative data with assessment.

<sup>15</sup> Kulik, 24.

## Annotated Bibliography

Abrami, Philip C., D'Apollonia, Sylvia, and Cohen, Peter A. "Validity of Student Ratings of Instruction: What We Know and What We Do Not." *Journal of Educational Psychology*, 1990, 82(2), 219-231.

An empirical study of the various statistical analyses conducted to determine "validity" in the use of student ratings. The authors conclude that all the statistical surveys have been unable to control for all the variables and that the results are not conclusive. Instead more research should be done as it is clear that quality teaching and learning can be captured through effective construction of and assessment of student feedback on instructors.

Basow, S. A., and Silberg, N. I "Student Evaluations of College Professors: Are Male and Female Professors Rated Differently?" *Journal of Educational Psychology*, 1987, 79(3), 308-314.

An empirical study that compares evaluations of male and female professors. Female instructors scored worse on most measures, especially when evaluated by male students. Leading statistical factors for this were a strong correlation between enthusiasm and ability. Finally, a longitudinal comparison to similar studies today would yield a useful benchmark to progress since 1987.

Greenwald, Anthony G., and Gerald M. Gillmore. "Grading Leniency is a Removable Contaminant of Student Ratings." *American Psychologist*, Vol 52(11), Nov 1997, 1209-1217.

An empirical study that analyzes whether or not you can control for "grade inflation" in assessing student feedback. This gets at the validity question of student feedback. Although there are repeated studies that show a strong correlation between high feedback scores and expected course grades, this study recommends against abandoning student ratings. Instead, there are still other measures that can be assessed regardless of student expectations on grades.

Marsh, Herbert. W, and Roche, Lawrence A. "Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility" *American Psychologist*. Vol 52(11), Nov 1997, 1187-1197

Looking at various different approaches to determine validity of student evaluations, the authors conclude that much of the controversy over the topic is due to not separating out various components of effective teaching. . However, at the same time the best way to assess teaching is by taking multidimensional approach. This includes more broad construct-validation approach.

McKeachie, W. J. "Student Ratings: The Validity of Use." *American Psychologist*, Volume 52(11), Nov 1997, 1218-1225.

Unlike other statistical surveys on student evaluations of their professors, McKeachie questions the utility of statistical surveys to assess “validity.” He agrees that feedback can be valid, but variables like expected grades have a clear influence. Instead of focusing on statistical analysis to determine the utility, more attention should be on the training of those that use (or abuse) the feedback for evaluating teachers.

Penny, Angela R. and Robert Coe “Effectiveness of Consultation on Student Ratings Feedback: A Meta-Analysis.” *Review of Educational Research*, Jan 2004; vol. 74: pp. 215 - 253.

The authors explore whether “debriefing” faculty on student ratings is an important strategy to support university teachers in learning from student ratings feedback. Through meta-analysis, the authors indicated that the various approaches to consultation are not equally effective. The authors suggest that only certain strategies of presenting feedback to faculty can maximize the benefits of student ratings feedback.

Seldin, Peter. “The Use and Abuse of Student Ratings of Professors.” *The Chronicle of Higher Education*, July 21, 1993, A40

This short summary article captures the main themes in student evaluations of teachers. While written in 1993, it shows the huge increase in use of student ratings. Use has increased from 29% in 1973 to 86% in 1993. The author also focuses on the limits of use. Evaluations must focus on asking the right questions (ones that students are qualified to answer) and should be one of many different sources to evaluate professors for quality teaching. Finally, feedback is only useful in improving teaching if 1) comments are something new to the professor, 2) professor wants improve, and 3) professor knows how to improve.

Theall, Michael, Philip C. Abrami, Lisa A. Mets, eds. *The Student Ratings Debate: Are they Valid? How can we best use them?*, San Francisco, CA: Jossey-Bass Publishers, 2001.

This anthology is a great updated source on student ratings. It is broken down into two parts. The first half of the book is focused on summaries of the current literature on evaluations from the leading authors in the field. These chapters cover the utility of student ratings, the validity of evaluations, and discuss some of the myths on the use and abuse of ratings. The second half of the book focuses on some of the best methodological research done on student ratings of instructors. Also within this section the various authors propose further research in the field.