Available online at https://rjsh.rsu.ac.th
RJSH Vol. 6, No. 1, January-June 2019, pp. 61-66

ISSN 2286-976X / Online: ISSN 2539-5513

Evaluating the Evaluation: Concerns about Student Evaluation of Teaching

Arlan Parreño

Institute for English Language Education, Assumption University, Bangkok 10240 Thailand
E-mail: arlanparreno@gmail.com

Received 19 March 2019 / Revised 28 April 2019 / Accepted in final form 20 May 2019 / Publish Online 8 June 2019

Abstract

It is believed that students' perception of how they are taught is essential in evaluating teachers for faculty development and personnel decision-making purposes. Thus, student evaluation of teaching, or SET, is a staple in educational institutions, especially in colleges and universities. There are, however, questions about the reliability and fairness of such practice. Many factors are perceived to influence student ratings of their teachers' performance, and grading is a persistent concern. As results of such evaluations are commonly used for administrative decisions, such as for faculty promotions or salary adjustments, teachers are tempted to modify their behavior to obtain favorable ratings. It is therefore suggested that student evaluation of teaching be handled with care in terms of formulation and administration, and be used in conjunction with other methods in order to have a valid and reliable evaluation of teachers' performance.

Keywords**:** *teacher evaluation, student evaluation of teaching, evaluation of teaching performance, SET*

1. Introduction

Studies about school effectiveness have shown that the most influential factor in the achievement of learners is the teaching quality, well ahead of curriculum, evaluation, and educational management (Hargreaves, 2014). No wonder then that student evaluation of teaching, or SET, is a common feature in tertiary institutions (Biggs & Tang, 2007).

As students are considered primary stakeholders in colleges and universities, their evaluation of how they are taught is seen as an important source of information in directing teacher development and decision-making involving teachers. Ideally, SET should be conducted to improve teaching quality. SET can help teachers understand how students perceive their teaching (Hattie, 2009), and that understanding can help them make adjustments in their teaching accordingly. In reality, however, results of SET are most of the time used for personnel decision-making in terms of salary, tenure, contract renewal, promotions, and awards (Biggs & Tang, 2007).

The evaluation usually comes in the form of a survey using a questionnaire administered either in person or online. Usually, SET instruments are written and formatted in a way that can be used in any department in order to compare teachers using a quantitative scale, for example from 1 to 5, with 1 as highly unsatisfactory and 5 as highly satisfactory. This quantitative approach is common because it offers "administrative convenience" (Biggs & Tang, 2007), especially in institutions with large teacher populations. Individual student responses can be manually keyed in computers or answer sheets can be fed in computerized readers. Quantitative results can be then easily computed using readily available statistical programs.

While SET has been helpful for administrators to make personnel decisions, its formative influence on teaching quality still remains to be seen. In Marsh's study (2007) of 195 teachers in over 13 years, results of SET have shown no significant impact on teacher development. It is rather strange, if not sad, that the opinion of those who are the direct recipients of the act of teaching seems to be considered so lightly, if considered at all. Hattie (2009) noted that SET appears useless for teachers because teachers do not learn anything significant from these supposedly valuable evaluations. This situation begs the question: Why not?

The reason for this seeming lack of teacher development as a result of SET is probably due to contentious issues about the reliability and validity of such an evaluation method. It is fair to say that teachers have the right not to follow what they believe is questionable. What can they learn from an evaluation that they think is not valid and reliable?

If you look at the more common reality of the use of SET, i.e. personnel decision-making, teachers' concerns regarding the reliability and validity of this approach put the whole teacher evaluation system in question.

2. Concerns about SET

Several studies have indicated that SET is reliable and valid (De Jong & Westershof, 2001; Greenwald, 1997; Marsh, 2007; Marsh & Roche, 1997;Overall & Marsh, 1980) However, it has not stopped many educators and researchers to question such claims.

According to Stark and Freishtat (2014) and Braga, Paccagnella, and Pellizzari (2014), SET ratings do not truly determine teaching effectiveness as they are significantly related to factors that have nothing to do with teaching, such as race, gender, physical traits, etc. Crumbley and Fliedner (2002) noted that SET questionnaires focus on how students perceive teachers rather than on student learning or achievement. Biggs and Tang (2007) also claimed that SET seems to assess teacher "charisma" as they are not designed to measure teaching performance based on changes in student learning but rather evaluate teacher characteristics. Kornell and Hausman (2016) also posited that students may not be fair judges of teaching effectiveness as they have the tendency to focus on teachers' traits. Crumbley and Fliedner (2002) contended that such evaluation design can be prone to distortion as teachers may alter their behavior so as to influence student perceptions. Pounder (2007) believed that teachers may be tempted to influence SET in different ways since the results can have significant impacts on their tenure, salary, promotions, among others.

Biggs and Tang (2007) also pointed out that SET questionnaires contain items that assume that the mode of teaching of all courses is lecture, with items like "speak clearly" or "hands out clear lecture notes". This assumption puts teachers who use different modes of teaching at a disadvantage.

Furthermore, there are factors that have been identified to have an influence on SET results. Koh and Tan (1997) noted that a smaller class size, a large number of evaluation answers, conducting the evaluation at the later part of the week, and higher-level subjects are all related to good SET ratings. The investigation of Badri, Abdualla, Kamali, and Dodeen (2006) revealed that expected and actual grades, course level, class size, course timing, student gender, and course subject have significant effects on SET results. Heckert, Latier, Ringwald, and Silvey's (2006) study also shows students' interest in course content, expected grades, satisfaction with the time of day (of the class) and gender of the teacher significantly relate to teaching performance. Boring, Ottoboni, and Stark (2016) found that SET ratings are influenced by students' grade expectations and instructors' gender, i.e. male instructors are rated more highly than female ones. Aleamoni and Hexner (1980) on the other hand contended that SET scores can be affected by instructions on the evaluation. They noted that students gave high ratings to teachers when told that the evaluation would be for personnel purposes, e.g. tenure or salary decisions. However, when told that the evaluation would be for teachers' personal use, e.g. course improvement, the students gave less favorable ratings.

2.1 SET and Grades

The most consistent concern raised by educators, which is supported by research, is the relationship between SET results and student grades. Although Marsh and Roche (1997, 2000) assured that grades do not significantly influence students' evaluations of their teachers, a number of studies claim otherwise. Greenwald and Gilmore (1997) asserted that the positive relationship between SET scores and expected grades has been proven by experimental studies that involved grade manipulation. The investigations of Abrami, Dickens, Perry, and Leventhal (1980); Addison, Best, and Warrington (2006); Badri et al. (2006); Ducette and Kenney (1982); Griffin, Hilton, Plummer, and Barret 2014); Marsh and Roche (1997); McKeachi (1997) show that grades have a significant positive relationship with SET ratings, i.e. teachers who give high grades receive good ratings.

When Marsh and Roche (1997) investigated student evaluation and their grades, they found that students obtained better grades from those considered as "good" teachers than from those "poor" ones. They posited that the high grades could indicate that students really learned well, or that those grades were reflective of lenient grading. D'Apollonia and Abrami (1997) shared the same view: students may believe that a teacher is effective when they do well in exams, or they may give high evaluations as a reward to teachers who give them good grades.

Marsh (1983); Marsh and Roche (1997) offered three explanations for SET ratings and grade relationship. First, grades really reflect what students learn, i.e. good grades mean that students have learned well. Second, the relationship can be deceiving due to the influence of some factors such as poor interest in a course, previous learning experience, class size and level of the course. Finally, it is the leniency, not the grades, that affects SET ratings, which means teachers who grade students more than they deserve are appreciated by students with better-than-deserved SET ratings.

Greenwald and Gilmore (1997) theorized the following to explain the relationship between grades and SET ratings: good grades and high ratings are results of effective teaching; the over-all learning motivation of students affects their grades and their evaluation of their teachers; the motivation of students towards specific courses affect their grades and their evaluation; students believe that their grades reflect the quality of the course and their own academic capability; and, students' good ratings are an expression of gratitude for good grades. They also put forward the attribution theory to explain why some teachers receive low SET ratings: when students get high grades, they credit it as their personal achievement, but they put the blame on their teachers when they receive poor grades. They contend, however, that students' evaluation of their teachers is mostly based on not what they have achieved or learned from them.

Griffin et al. (2014) seemed to back up Marsh and Roche's (1997, 2000) contention. Their research findings indicate that students' grades are moderately correlated to SET ratings. However, the correlation had a large variance and was not applicable to individual teachers and courses.

However, several studies point to the influence of grading on SET ratings. Ducette and Kenney (1982) found that teachers who give good grades receive higher SET ratings than those who give lower grades. Moreover, they noted that expected grades are significantly related to course difficulty, course effectiveness, and teacher effectiveness. These findings indicate that students give better SET scores to teachers from whom they expect to get good grades with the belief that those teachers are more effective than others. On the other hand, students who believe that they would get low grades tend to think that the course is too hard and too demanding. But instead of taking responsibility, they attribute their expectation of poor grades to the course and the teacher, which means evaluating the course and the respective teacher poorly, too.

Selsby and Sterle (2015) found that students' perceived grades and their perceived variance between what they expected at the beginning and at the end of a course influence their evaluation of their teachers significantly. Those who expected high grades and perceived little or no difference in their final grades rate their teachers better than those who perceived their final grades to be much lower than their expected grades at the onset of the course.

Abrami et al.'s (1980) experiments show that teachers may obtain varying evaluation scores if they use varying standards in giving grades.

Such correlational studies indicating the influence of grades on SET ratings are further supported by surveys. Al-Issa and Sulieman (2007) noted that many of the students they surveyed admitted that they rated their teachers based on their expected grades in those teachers' courses. Crumbley, Henry, and Kratchman's (2001) survey also revealed that about 4 out of 10 students tried to find out the grading behavior of teachers to help them decide which teacher to study with, i.e. choosing the lenient ones.

2.2 SET and Teacher Behavior

Because of the fear that students will give them poor SET ratings, teachers can be tempted to alter their teaching and grading methods to gain favorable evaluations (Crumbley et al., 2001). Such action is, of course, understandable due to the high stakes of SETs in terms of promotion, tenure, or salary adjustment.

Redding (1998) noted that many teachers provide easy courses and administer easy tests so as to be rated favorably by their students. Thus, Redding believes that there are teachers who inflate grades to get good SET ratings and also to avoid student complaints about grades.

Crumbley and Fliedner (2002) believed that teachers have the power to influence students' evaluation of their performance. In their survey, about 40% of school administrators responded that they were aware of teachers who behaved in such a way that would get them high SET ratings. The administrators knew that those teachers graded leniently and gave easy coursework.

To influence student evaluation, teachers actually may not have to give the final grade. They can simply make students believe that they will get high grades. Expecting high grades can cause students to give their teachers high SET ratings (McPherson, 2006).

3. So What Should Be Done?

SET can be a valuable tool in teacher evaluation, but because of various biasing factors, they have to be used with care. Abrami, D'Apollonia, and Cohen (1990) ; Stark and Freishtat (2014) warned school administrators to be cautious in interpreting and generalizing SET results. Biggs and Tang (2007) did not warn only about SET result interpretation but also about the process of administering and formulating SET questionnaires. They believed that SET should be organized by departments and not by faculty or school administration. Moreover, they recommended improvements in questionnaires by making sure that they are constructed in a way that supports "constructive alignment", for example,

- whether students are clear about the intended learning outcomes,
- what standards they have to reach to attain the various grades, and
- that the teaching-learning activities in their experience really help them to achieve their intended learning outcomes.

Teacher evaluation, whether for the purpose of faculty development or personnel decision-making, will always be questioned if it is founded only on SET. To make the evaluation more reliable and fair, it should not be based solely on SET. Emery, Kramer, and Tian (2003); Koh and Tan (2007) ; Kornell and Hausman (2016) ; Stark and Freishtat (2014); Toch (2008) recommended that evaluators gather data for teacher evaluation from various sources. Data from such sources can then be triangulated to obtain a clearer picture of a teacher's performance.

Biggs and Tang (2007) believed that teachers should be given an opportunity to address the criteria set for teacher evaluation. This can be done by compiling a portfolio that should contain a teacher's teaching philosophy with a discussion on how such philosophy is implemented in teaching-learning situations. Lesson plans, materials, samples of student work, and evaluations in respective courses should be included in the portfolio to serve as proof. Evaluators can then analyze the portfolio to determine a teacher's performance based on those teacher evaluation criteria.

Aside from teaching portfolios, Emery et al (2003) recommended the use of students' achievements and peer evaluations. Centra (1987); Mckeachie (1997); Stevens (1987) suggested the following as sources of teacher evaluation information: class observation, portfolios, student interviews, and videotaping of classes. Mckeachie (1997) further suggested the use of written comments from students and teachers' self-evaluation. Koh and Tan (2007) proposed supplemental teacher evaluation tools such as class observation by internal and/or external evaluators, portfolio analysis, teacher self-evaluation, graduate and peer evaluation, and assessment of student achievement. Kornell and Hausman (2016) also believed that evaluations by "expert teachers" and an objective assessment of what students have really learned after finishing a particular subject should be employed together with SET to truly determine teaching effectiveness.

4. Conclusion

Student evaluation of teaching is considered an integral and vital part of the teacher evaluation process in tertiary education. SET rating results may provide useful information for teacher development and for personnel decision. However, empirical studies have continued to question its value that Spooren, Brockx, and Mortelmans (2013) meta-evaluation concludes SET is "fragile".

Questions about the credibility of SET results obviously devalue SET in terms of teacher development. How can teachers be guided on what and how to improve if the evaluation is dubious? What's more, how can teachers be encouraged to improve if they do not trust the evaluation?

Undeniably, personnel decision-making based on SET is definitely impacted by SET's fragility as well. Perceptions of bias and incompetence on the part of administrators are inevitable. Furthermore, the culture of SET as a popularity contest can surely tempt even the most idealistic teachers to work on their popularity as well in order to get favorable personnel decisions.

Questions and concerns about SET definitely point to one thing: SET should not be the sole basis for faculty development and personel decision-making. Yes, SET is a valuable tool in teacher evaluation. However, school administrators should exercise caution in using SET ratings when making decisions regarding tenure, promotions, and salary, as well as when implementing faculty development actions. Moreover, they should not depend only on SET; they should make use of other methods in collecting information about teachers' performance. Class observations, alumni and peer evaluation, video-recording of classes, self-evaluation, student interviews and written comments, student achievement, and portfolios can provide more data about teaching performance. Triangulation of data from such sources, including SET, can produce fair and reliable teacher evaluation.

5. References

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, *82*(2), 219-231.

Abrami, P. C., Dickens, W. J., Perry, R. P., & Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology*, *72*(1), 107.

Addison, W. E., Best, J., & Warrington, J. D. (2006). Students' Perceptions of Course Difficulty and Their Ratings of the Instructor. *College Student Journal*, *40*(2), 409-416.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science*, *9*(1), 67-84.

Al-Issa, A., & Sulieman, H. (2007). Student evaluations of teaching: perceptions and biasing factors, *Quality Assurance in Education*, *15*(3), 302-317.

Badri, M. A., Abdulla, M., Kamali, M. A., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management*, *20*(1), 43-59.

Biggs, J., & Tang, C. (2007). *Teaching for Quality Learning at University* (3rd ed.). New York: Society for Research into Higher Education & Open University Press (McGraw-Hill).

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, Retrieved from https://www.scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, *41*, 71-78. Retrieved from https://doi.org/10.1016/j.econedurev.2014.04.002

Centra, J. A. (1987). Formative and summative evaluation: Parody or paradox? *New Directions for Teaching and Learning*, (31), 47-55.

Crumbley, L. D., & Fliedner, E. (2002). Accounting administrators' perceptions of student evaluation of teaching (SET) information. *Quality Assurance in Education*, *10*(4), 213-222.

Crumbley, L., Henry, B. K., & Kratchman, S. H. (2001). Students' perceptions of the evaluation of college teaching. *Quality Assurance in Education*, *9*(4), 197-207.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*(11), 1198-1208.

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behavior. *Learning Environments Research*, *4*(1), 51-85.

DuCette, J., & Kenney, J. (1982). Do grading standards affect student evaluations of teaching? Some new evidence on an old question. *Journal of Educational Psychology*, *74*(3), 308-314.

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality assurance in Education*, *11*(1), 37-46.

Greenwald, A. (1997). Validity Concerns and Usefulness of Student Ratings of Instruction. *The American psychologist*, *52*(11), 1182-1186.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*(11), 1209-1217.

Griffin, T. J., Hilton Lii, J., Plummer, K., & Barret, D. (2014). Correlation between grade point averages and student evaluation of teaching scores: taking a closer look. *Assessment and evaluation in higher education*, *39*(3), 339-348.

Hargreaves, A. (2014). Forword: Six Sources of Change in Professional Development. In L. E. Martin, S. Kragler, D. J. Quatroche & K. L. Bauserman (Eds). *Handbook of Professional Development in Education: Successful models andpractice, PreK12* (pp. x-xix). NY: Gulliford.

Hattie, J. (2009). *Visible Learning: A Synthesis of over 800 meta-analyses relating to achievement*. NY: Routledge.

Heckert, T., Latier, A., Ringwald, A., & Silvey, B. (2006). Relation of course, instructor, and student characteristics to dimensions of student ratings of teaching effectiveness. *College Student Journal, 40*(1), 195-203.

Koh, C. H., & Tan, M. T. (1997). Empirical investigation of the factors affecting SET results. *International Journal of Educational Management*, *11*(4), 170-178.

Kornell, N., & Hausman, H. (2016). Do the Best Teachers Get the Best Ratings? *Frontiers in Psychology*. Retrieved from https://doi.org/10.3389/fpsyg.2016.00570

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology*, *75*(1), 150-166.

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*(4), 775-790.

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

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92*(1), 202-228.

Marsh, H. W., Hau, K. T., Chung, C. M., & Siu, T. L. (1997). Students' evaluations of university teaching: Chinese version of the Students' Evaluations of Educational Quality instrument. *Journal of Educational Psychology*, *89*(3), 568-572.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225.

McPherson, M. A. (2006). Determinants of How Students Evaluate Teachers. *Journal of Economic Education*, *37*(1), 3-20.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72*(3), 321-325.

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.

Redding, R. E. (1998). Students' evaluations of teaching fuel grade inflation. *American Psychologist, 53*(11), 1227-1228.

Selsby, J., & Sterle, J. (2015).Student perception of achievement influences student evaluation of teaching. *The FASEB Journal*, *29*(1_supplement), 541-37

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.

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*. Retrieved from https://www.scienceopen.com/document?vid=42e6aae5-246b-4900-8015-dc99b467b6e4

Stevens, J. J. (1987). Using student ratings to improve instruction. *New Directions for Teaching and Learning*, (31), 33-38.

Toch, T. (2008). Fixing Teacher Evaluation**.** *Educational Leadership*, *66*(2), 32-37.