Hornstein, *Cogent Education* (2017), 4: 1304016
http://dx.doi.org/10.1080/2331186X.2017.1304016

cogent education

CURRICULUM & TEACHING STUDIES | REVIEW ARTICLE

Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance

Henry A. Hornstein[1*]

Abstract: Literature is examined to support the contention that student evaluations of teaching (SET) should not be used for summative evaluation of university faculty. Recommendations for alternatives to SET are provided.

Subjects: Education - Social Sciences; Economics, Finance, Business & Industry; Business, Management and Accounting

Keywords: teaching; university faculty evaluation; student evaluations of teaching; teaching evaluation

Received: 22 November 2016
Accepted: 28 February 2017

*Corresponding author: Henry A. Hornstein, Department of Business and Economics, Algoma University, 1520 Queen St. E., Sault Ste. Marie, Ontario, Canada P6A 2G4
E-mail: henry.hornstein@algomau.ca

Reviewing editor:
Hau Fai Edmond Law, Education University of Hong Kong, Hong Kong

Additional information is available at the end of the article

1. Introduction

As far back as the 1990s, Adams (1997) and others have challenged the validity of student evaluations of teaching (SETs) as legitimate measures of faculty instructional effectiveness/capability (Wright & Jenkins-Guarnieri, 2012). Adams (1997) mentioned many problems with the use of SETs that have since appeared in the literature such as validity, reliability, gender bias, and a number of other related issues (Beecham, 2009; Boring, Ottoboni, & Stark, 2016; Braga, Paccagnella, & Pellizzari, 2014; Hoefer, Yurkiewicz, & Byrne, 2012; Spooren, Brockx, & Mortelmans, 2013; Stark & Freishtat, 2014; Wright, 2006; Yunker & Yunker, 2003). Yet in the almost 20 years that have passed since his paper was published, SETs remain the main tool utilized in the assessment of instructor teaching competence, promotion and tenure (Boring et al., 2016; Kelly, 2012; Spooren et al., 2013) in both the United States and Canada notwithstanding the noteworthy lack of consensus among scholars as to their legitimacy and validity as measurement instruments.

ABOUT THE AUTHORS

Henry A Hornstein, PhD, is a full time associate professor in the Department of Business & Economics at Algoma University in Sault Ste. Marie, Ontario, Canada. His main disciplines of interest are Project Management and Organizational Change Management, Consumer Behavior, and Business Ethics. His recent involvement in the renegotiation of the Collective Agreement between university faculty and the Board of Governors has highlighted the unfairness and inappropriateness of the current evaluation process undertaken by the university to assess whether faculty will get tenure and promotion. This motivated him to survey the current literature to determine whether Student Evaluations of Teaching, in many cases, the sole measure of teaching competence used in tenure and promotion decisions, are legitimate measures of teaching. His paper, recently accepted by *Cogent Education*, is the result.

PUBLIC INTEREST STATEMENT

For a long while, teaching evaluations have been collected by academic institutions across North America ostensibly to be factored into institutional decisions as to the appropriateness of faculty to get tenure and promotion. The problem has long been that these instruments do not legitimately assess teaching competence, which has always been the almost sole criterion on which to base the decision. At best, the instruments speak about student opinions about faculty teaching capability, and student opinions and/or satisfaction are NOT legitimate measures of teaching capability. This paper presents evidence to substantiate the need to end the use of these surveys because they don't measure what they say they measure. Moreover, the tyranny of teaching evaluations should never be the sole determiner of whether faculty deserve tenure and/or promotion.

The use of student ratings of teaching (which are, in essence, student *opinions* of teaching capability) is the most common process employed at universities to evaluate faculty teaching quality/competence (see Cashin, 1999; Clayson, 2009; Davis, 2009; Seldin, 1999; Seldin, Miller, & Seldin, 2010). Interestingly, they define "effective teaching," even though there is no consensus in the literature on what that is. They are popular partly because the measurement is easy, i.e. students simply fill out forms that require little class and faculty time. The biggest cost is to record the data, and even then, electronic evaluations automate that step. Averages of student ratings appear objective simply because they are numerical. And comparing the average rating of any instructor to the average for a department as a whole is simple. However, as will be seen later, many of these reasons are problematic.

In the 1970's, SETs were intended primarily for *formative* purposes, that is, to improve and shape the quality of teaching. However, since the 1970s, they have, in addition, become the primary indicator for summative evaluation, that is, to "sum up" overall performance to decide about promotion, and tenure. In other words, SET has evolved into the dominant, and in many cases, sole indicator of teaching competence (Berk, 2005; Galbraith, Merrill, & Kline, 2012; Spooren et al., 2013). As mentioned above, this evolution appears to have resulted from the ease with which data are collected, presented and interpreted. However, the interpretations are questionable on conceptual and statistical grounds, that is, few articles in the literature challenge the interpretation that student satisfaction ratings (SET) equate to teaching competence. Fewer still note that SET data, being categorical, cannot be evaluated validly using parametric statistics. Nonetheless, Hamermesh and Parker (2005) argue that SET is used for faculty tenure/promotion evaluations and pay determination regardless of whether the evaluations correspond to legitimate measures of underlying teaching quality/competence. Sproule (2002) argues that when student evaluations are used, adjustment is necessary since the factors being assessed are not under the control of faculty. Sproule (2000) points out that while SET encourages students to view themselves as customers/consumers of education, the relationship between student and instructor is not analogous to the customer–client covenant, notwithstanding administrations' attempts to corporatize the university.

As is shown in this paper, the persistent practice of using student evaluations as summative measures to determine decisions for retention, promotion, and pay for faculty members is improper and depending on circumstances could be argued to be illegal.

2. Measurement

There are basic measurement-level issues with how SET are interpreted which render them essentially useless as instruments for the evaluation of faculty promotion and tenure. The data that are used presently at most universities to assess teaching competence are paper-and-pencil, and/or electronic-based surveys which are administered to students attending classes. Students are expected to rate on a Likert scale their instructors using a number of statements ostensibly pertaining to teaching behavior in the classroom. The data are then collected and a set of categories with ordinal values are presented and assessed (Agresti & Finlay, 1997). That is, currently, the specific rating categories used at most institutions in North America by students to assess teaching competence are/resemble the following: "unacceptable," "very poor," "poor," "satisfactory," "good," "very good," and "outstanding." These categories differ in quality, not in quantity or magnitude. In other words, the "interval distance" between the categories is undefined. Usually, numbers are attached to each category to make it appear as an interval level of measurement so that statistical analyses can be applied (e.g. 1. unacceptable, 2. very poor, 3. poor, 4. satisfactory, 5. good, 6. very good, 7. outstanding). However, when this is done, good judgment and understanding must be applied in the interpretation of any statistical analyses. For example, any statistical evaluation of categorical data (like the SETs) should not include measures of central tendency like means or averages that are appropriate only for quantitative data. An average calculated on categorical data is quite meaningless and misleading (Stark & Freishtat, 2014). For instance, saying that categories have an average is analogous to saying that the average of a pen, peach, eraser, and piece of paper is a book. So, in terms of how SETs are currently used at most universities, one can legitimately question that means are reported

to faculty, and/or are used in decision-making around tenure, promotion and hiring. It is not possible to interpret average scores of categories. The categories are not truly ordinal, and in any case, there is no zero. Calculating the means of categories yields results that are uninterpretable—essentially trivial. McKeachie (1996) argues that the SET should not be used as a continuous rating scale, where, for instance, one can say that a rating of 3.5 is more than 3.0 and insignificant decimals determine promotion and tenure. Rather, they should be used as a discrete standard that should be met (within a certain confidence interval and level). Yet, the studies that support the connection between SET and teaching capability/competence (e.g. Cohen, 1981; Marsh, 2007) neither pay attention to the legitimacy of the data used, nor to whether or not appropriate statistical analyses were employed. This renders their conclusions suspect.

Another of the problems with this type of student evaluation of faculty is that few administrators are trained to interpret SET data. It is not uncommon for administrators to examine the scores and assume that those below the mean are bad and those above it are good—never mind that the calculation of means in these situations is simply inappropriate and meaningless. Their reasoning seems to be based on the improbable assumption that all of their faculty members should be above average in all categories (Klajman, 1997), essentially creating a Lake Wobegon effect (Keillor, 1985; Kruger, 1999), which says, "for nearly any subjective and socially desirable dimension ... most people see themselves as better than average" (Myers, 1998, p. 440).

As Abrami, D'Apollonia, and Cohen (1990) point out, "Student ratings are seldom criticized as measures of student satisfaction with instruction ... Student ratings are often criticized as measures of instructional effectiveness" (p. 219).

3. Validity of student assessment

A related issue flowing from the above has to do with whether undergraduate students have the ability to assess instructor teaching competence. The validity of anonymous students' evaluations rests on the assumption that, by attending lectures, students observe the ability of the instructors, and that they report it truthfully. While this view is plausible in some respects, there are also many reasons to question that students are dispassionate evaluators of instructor performance. For example, students' objectives might be and often are different from those of university administration and/or faculty (Braga et al., 2014). Students may simply care about their grades, whereas in most cases, faculty cares about student learning. It is likely that the two are not highly correlated, especially when the same professor is engaged both in teaching and in grading.

Students can reliably speak about *their experience* in a course, including factors that ostensibly affect teaching effectiveness such as audibility of the instructor, legibility of instructor notes, and availability of the instructor for consultation outside of class (Becker, Bosshardt, & Watts, 2012). However, they cannot evaluate *outside their experience*, i.e. how can they assess course pedagogy? By what valid criteria are they able to determine how "knowledgeable" an instructor is about his/her subject area? The criteria that students use through the instruments that they complete are likely to be unrelated to teachers' actual teaching qualities (Boring, 2015). Since student evaluations have yet to be shown to be valid measures of teaching quality (Dunn, Hooks, & Kohlbeck, 2016) but are, nevertheless, often used by administrators to make critical decisions concerning the retention, promotion, and pay of faculty members, it is only reasonable to expect that faculty will do what they can to achieve the highest possible ratings, especially for junior faculty who desire to be tenured. Troy (1995) stated, "Instructors know that in each class some student will say, 'What do I need to do to get an A?' There's no reason the instructor can't say something similar to the students" (p. 1). Moreover, there is no consensus among scholars concerning the definition of "effective teaching" or teaching competence (Spooren et al., 2013), so how is it reasonable to expect students who have little to no content knowledge to be able to evaluate it? Yet university administrators, and tenure and promotion committees act as if the relationship between SET and competent teaching is clear and unequivocal, and that therefore it is reasonable to require faculty to obtain "high" SET scores because that means that their teaching performance is superior. In fact, this relationship is anything

but unequivocal (for example, Langbein, 2008). Students offer only a single perspective on a very complex and multifaceted teaching and learning process that no single source of evidence can reasonably evaluate (Stark & Freishtat, 2014).

4. Response rates and satisfaction

It should be noted that in cases where online evaluation systems are used, Dommeyer, Baum, Hanna, and Chapman (2004) reported average response rates of 70% for in-class surveys and 29% for online surveys. Ling, Phillips, and Weihrich (2012) have found similar results, and these findings represent the pattern seen in the literature, a subject of concern to many faculties because less detailed data are available and there is fear that there will be lower teaching evaluations (Nulty, 2008). Statistically speaking, this suggests that the sample results are questionable in terms of generalizability. Low response rates are a consequence of a number of issues including overall satisfaction with instruction, apathy, absence from class, technical problems, perceived lack of anonymity, lack of importance, inconvenience, inaccessibility, and time for completion (Berk, 2012). However, none of these reasons tend to be considered when particular university tenure and promotion committees interpret the scores. Instead, inevitably, the onus is on the faculty member being evaluated to justify "low scores"—a difficult and in many ways unjustified task since he/she does not have the relevant information on which to base an explanation, and there are significant questions as to the reliability of the instruments used to collect student evaluation information. Moreover, asking faculty to justify any kind of SET score is unnecessary since as has been already mentioned, there is little to suggest that SET say anything about teaching competence (Braga et al., 2014). They do reflect, however, how satisfied those students who completed courses are with the service/teaching they receive (Beecham, 2009), but student satisfaction is a complex phenomenon influenced by a number of variables.

The most important determinants of student satisfaction that are discussed in the literature involve teaching performance and teacher characteristics, career-related issues, programme/course innovativeness and appropriateness, and classroom facilities, among others. Among these determinants, teaching performance, as an important element of education quality, is considered to be the most significant source of student satisfaction (Ginns, Prosser, & Barrie, 2007; Malik, Danish, & Usman, 2010). In particular, factors such as teacher competencies, communication skills, attitudes, likability and appropriate use of humor were found to be positively correlated with student ratings (Duque, 2013). Duque and Weeks (2010) found innovativeness and engagement stimulate student satisfaction. In a similar vein, Lizzio, Wilson, and Simons (2002) identified an association between the student perception of the teaching environment with student satisfaction and the level of academic achievement. Moreover, career preparation and relevance were also found to be significant determinants of student evaluations (Duque & Weeks, 2010; Yeo & Li, 2013). With regard to the teaching environment, researchers have consistently reported that image and tradition (Alves & Raposo, 2007), as well as the availability of adequate facilities, classrooms and resources at post-secondary institutions (Malik et al., 2010; Sakthivel, Rajendran, & Raju, 2005) significantly contribute to overall student satisfaction. These findings confirm the complexity of the student satisfaction variable and suggest that teaching competence is not a component of its assessment.

There is no current literature of which the current author is aware that convincingly argues for the position that SET scores are measures of instructor teaching competence (Beecham, 2009; Braga et al., 2014; Hoefer et al., 2012; Spooren et al., 2013; Stark & Freishtat, 2014; Wright, 2006; Yunker & Yunker, 2003), and no specific evidence of content validity that suggests that the instrument used at many universities measures teaching competence. As a matter of fact, to the current author's knowledge, universities tend not to advocate any clear theory of effective teaching. Moreover, a number of the above authors (Beecham, 2009; Braga et al., 2014; Spooren et al., 2013; Stark & Freishtat, 2014) have suggested that SETs are measures of popularity and liking (utility) rather than bonafide measures of teaching capability, and should be used cautiously, if at all, to evaluate faculty promotion and tenure applications. Moreover, the inadequate and highly variable response rates mentioned earlier render this information statistically questionable.

In a study relating SET scores to academic performance, Braga et al. (2014) conclude that good teachers are those who require their students to exert effort; students dislike having to expend this effort, especially the least able ones, and their evaluations reflect the utility they enjoy from the course. So, in other words, the lower the evaluations, the better that student performance tends to be because the instructor has required students to expend significant effort in order to achieve better grades, and students dislike expending effort. These investigators also have suggested that the best indicator of instructor teaching quality is student academic performance at the end of a semester/term.

Boring et al. (2016) evaluated SET measures in two different universities, in different continents, across a broad range of course topics, and concluded that SETs are biased against female instructors by an amount that is large and statistically significant. They also found that the bias affects how students rate even supposedly objective aspects of teaching, such as how promptly assignments are graded; the bias varies by discipline and by student gender, among other things; it is not possible to adjust for the bias, because it depends on so many factors; *SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness* (italics mine); gender biases can be large enough to cause more effective instructors to get lower SET than less effective instructors. These findings led Boring et al. (2016) to conclude that SET should not be used for personnel decisions, a conclusion that contradicts the practice employed by most North American universities and colleges.

5. Conclusion

Despite 1970s and 1980s research that supported the validity of SET as a measure of instructor teaching competence (e.g. Marsh, 2007), these studies have been beset by questionable conceptual and statistical interpretations that have tended to be overlooked by scholars. Moreover, the research has never been unequivocal, and the persistent statistical and conceptual errors/misinterpretations have rendered the conclusions questionable at best.

If one truly wants to understand how well someone teaches, observation is necessary. In order to know what is going on the classroom, observation is necessary. In order to determine the quality of instructors' materials, observation is necessary. Most of all, if the actual desire is to see improvement in teaching quality, then attention must be paid to the teaching itself, and not to the average of a list of student-reported numbers that bear at best a troubled and murky relationship to actual teaching performance. University faculty benefits most from visiting each other's classrooms and looking at others' teaching materials routinely. Learning can occur from one another, exchanging pedagogical ideas and practices.

Many studies cast doubt on the validity of SET as a measure of teaching effectiveness (e.g. Johnson, 2003). Some studies find that gender and SET are not significantly associated (Bennett, 1982; Centra & Gaubatz, 2000; Elmore & LaPointe, 1974). Those studies generally address a different issue, namely, whether men and women receive similar SET. That, however, does not control for teaching effectiveness, effort, or other variables. The more relevant question is whether women would receive higher scores for doing the same thing had they been male, and whether men would receive lower scores for doing the same thing had they been female. Boring et al.'s (2016) analysis of their US data shows that is true. Their analysis of their French data shows that, on average, less effective male instructors receive higher SET than more effective female instructors.

Nonetheless, it is unlikely that the use of SET for summative purposes will easily disappear. It is valued by students and by university administrators. Students like to be satisfied, and they like higher grades. Administration wants to retain students, and prefers a low-cost system to monitor faculty that looks "objective" (see Becker, 2000; Sproule, 2000) But what about the faculty? Why should they value this system? Given the use of rank order merit pay systems based on SET scores, the median faculty is "above average." Because 50% are above the median raise, it will probably survive a vote. Of course, 50% are below the "average," which is really a median. This group is likely to be

demoralized. The result is driven by rank order evaluations, which are an important component in many merit pay/continuous reward systems. Yet many scholars note problems with relative pay systems. For example, McKeachie (1996), argues that, even if SETs are valid and reliable (which current literature suggests is not the case), when 90% of teachers at a university are rated "excellent", but, with a relative pay system, 50% are still below the median rating, the consequence is de-motivation and demoralization. If this demoralization has observable consequences for quality teaching, it would serve as another warning that extrinsic rewards, especially when they are awarded on a piece-rate basis, can drive out intrinsic motivations, resulting in a costly loss of performance (Frey & Oberholzer-Gee, 1997; Kreps, 1997). Kane and Staiger (2002) note that variances of rating systems are higher when there are small classes or fewer courses, and that a few disgruntled students can have a large and disproportionate impact on averages.

On the basis of all of this evidence, the conservative and more appropriate approach is to question the validity of SET for all summative purposes, and even to avoid its use in the hiring of new faculty. Continuing this at best equivocal practice is akin to advocating that what students say is the same thing as teaching effectiveness, which, as has been amply demonstrated in the literature is nonsensical. Stark and Freishtat (2014) suggest that reliably and routinely measuring teaching effectiveness will never happen because it does not seem possible to effectively define it. In light of the controversy surrounding the utility of SET, they have made the following recommendations (p. 6):

(1) Drop omnibus items about "overall teaching effectiveness" and "value of the course" from teaching evaluations: They are misleading.

(2) Do not average or compare averages of student rating scores: Such averages do not make sense statistically. Instead, report the distribution of scores, along with the number of responders and the response rate.

(3) Pay careful attention to student comments—but understand their scope and limitations. Students are the authorities on their experiences in class, but typically are not well situated to evaluate pedagogy generally.

(4) Use caution extrapolating student evaluations to the entire class. When response rates are low, extrapolation is unreliable.

(5) Avoid comparing teaching in courses of different types, levels, sizes, functions, or disciplines.

(6) Use teaching portfolios as part of the review process.

(7) Use classroom observation as part of milestone reviews.

(8) To improve teaching and evaluate teaching fairly and honestly, spend more time observing the teaching and looking at teaching materials.

Finally, tenure systems at most universities are ostensibly based on merit, but, as many untenured faculty members know, the secrecy surrounding tenure votes means that controversial faculty, no matter how accomplished or meritorious, risk a negative decision. The ideals of academic freedom are therefore attenuated by the process, even before one considers the impact of student teaching evaluations. Student evaluations, with all the biases they embrace, put pressure on faculty to go slow and not rock the boat. In other words, do not push undergraduates to maximize their intellectual potential because that might fuel resentment, and do not confront the dominant political and religious beliefs of your particular subset of late adolescents even when such beliefs are patently false and when confronting them is supposedly part of the education process and is course appropriate. Undergraduates might retaliate on evaluations. Even post-tenure faculty face pressures from students, alumni, administrators, and sometimes even the public and the media, to conform their views to those views popularly held (Wines & Lau, 2006).

Funding
The author received no direct funding for this research.

Author details
Henry A. Hornstein[1]
E-mail: henry.hornstein@algomau.ca
[1] Department of Business and Economics, Algoma University, 1520 Queen St. E., Sault Ste. Marie, Ontario, Canada P6A 2G4.

Citation information
Cite this article as: Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance, Henry A. Hornstein, *Cogent Education* (2017), 4: 1304016.

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*, 219–231. http://dx.doi.org/10.1037/0022-0663.82.2.219

Adams, J. V. (1997). Student evaluations: The ratings game. *Inquiry, 1*, 10–16.

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall Inc.

Alves, H., & Raposo, M. (2007). Conceptual model of student satisfaction in higher education. *Total Quality Management & Business Excellence, 18*, 571–588. http://dx.doi.org/10.1080/14783360601074315

Becker, W. E. (2000). Teaching economics in the 21st century. *Journal of Economic Perspectives, 14*, 109–119. http://dx.doi.org/10.1257/jep.14.1.109

Becker, W. E., Bosshardt, W., & Watts, M. (2012). How departments of economics evaluate teaching. *The Journal of Economic Education, 43*, 325–333. http://dx.doi.org/10.1080/00220485.2012.686826

Beecham, R. (2009). Teaching quality and student satisfaction: Nexus or simulacrum? *London Review of Education, 7*, 135–146. http://dx.doi.org/10.1080/14748460902990336

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology, 74*, 170–179. http://dx.doi.org/10.1037/0022-0663.74.2.170

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*, 48–62.

Berk, R. A. (2012). Top 20 strategies to increase the online response rates of student rating scales. *International Journal of Technology in Teaching and Learning, 8*, 98–107.

Boring, A. (2015). *Gender biases in student evaluations of teachers* (working paper). OFCE-PRESAGE-SCIENCES PO and LEDa-DIAL.

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. Retrieved from Science Open Research. 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*, 71–88. http://dx.doi.org/10.1016/j.econedurev.2014.04.002

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (Ed.), *Current practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 25–44). Bolton, MA: Anker.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*, 17–33.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*, 16–30.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281–309. http://dx.doi.org/10.3102/00346543051003281

Davis, B. G. (2009). *Tools for teaching* (2nd ed.). San Francisco, CA: John Wiley & Sons.

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education, 29*, 611–623. http://dx.doi.org/10.1080/02602930410001689171

Dunn, K. A., Hooks, K. L., & Kohlbeck, M. J. (2016). Preparing future accounting faculty members to teach. *Issues in Accounting Education, 31*, 155–170. http://dx.doi.org/10.2308/iace-50989

Duque, L. C. (2013). A framework for analyzing higher education performance: Students' satisfaction, perceived learning outcomes, and dropout intention. *Total Quality Management and Business Excellence, 25*(1–2), 1–21.

Duque, L. C., & Weeks, J. R. (2010). Towards a model and methodology for assessing student learning outcomes and satisfaction. *Quality Assurance in Education, 18*, 84–105. http://dx.doi.org/10.1108/09684881011035321

Elmore, P. B., & LaPointe, K. A. (1974). Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology, 66*, 386–389. http://dx.doi.org/10.1037/h0036493

Frey, B. S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review, 87*, 746–755.

Galbraith, C., Merrill, G., & Kline, D. (2012). Are student evaluations of teaching effectiveness valid for measuring student outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education, 53*, 353–374.

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education, 32*, 603–615. http://dx.doi.org/10.1080/03075070701573773

Hamermesh, D., & Parker, A. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*, 369–376. http://dx.doi.org/10.1016/j.econedurev.2004.07.013

Hoefer, P., Yurkiewicz, J., & Byrne, J. C. (2012). The association between students' evaluation of teaching and grades. *Decision Sciences Journal of Innovative Education, 10*, 447–459. doi:10.1111/j.1540-4609.2012.00345.x

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York, NY: Springer-Verlag.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *The Journal of Economic Perspectives, 16*, 91–114.

Keillor, G. (1985). *Lake Wobegon days*. New York, NY: Viking Press.

Kelly, M. (2012). *Student evaluations of teaching effectiveness: Considerations for Ontario universities*. Toronto: Council of Ontario Universities (COU #866).

Klajman, G. (1997, February). Nightmares of academic assessment. *ASSESS - Assessment in Higher Education*. Retrieved from ASSESS@LSV,UKY.EDU

Kreps, D. M. (1997). Intrinsic motivation and extrinsic incentives. *American Economic Review, 87*, 359–364.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77*, 221–232. http://dx.doi.org/10.1037/0022-3514.77.2.221

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review, 27*, 417–428. http://dx.doi.org/10.1016/j.econedurev.2006.12.003

Ling, T., Phillips, J., & Weihrich, S. (2012). Online evaluations vs in-class paper teaching evaluations: A paired comparison. *Journal of the Academy of Business Education, 12*, 150–161.

Lizzio, A., Wilson, K., & Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: Implications for theory and practice. *Studies in Higher Education, 27*, 27–52. http://dx.doi.org/10.1080/03075070120099359

Malik, M. E., Danish, R. Q., & Usman, A. (2010). The impact of service quality on students' satisfaction in higher education institutes of Punjab. *Journal of Management Research, 2*(2), 1–11.

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). New York, NY: Springer. http://dx.doi.org/10.1007/1-4020-5742-3

McKeachie, W. (1996). *Student ratings of teaching* (Occasional Paper No. 33). American Council of Learned Societies, University of Michigan. Retrieved from http://archives.acls.org/op/33_Professonal_Evaluation_of_Teaching.htm

Myers, D. G. (1998). *Psychology*. New York, NY: Worth.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education, 33*, 301–314. http://dx.doi.org/10.1080/02602930701293231

Sakthivel, P. B., Rajendran, G., & Raju, R. (2005). TQM implementation and students' satisfaction of academic performance. *The TQM Magazine, 17*, 573–589. http://dx.doi.org/10.1108/09544780510627660

Seldin, P. (1999). Building successful teaching evaluation programs. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 213–242). Boston, MA: Anker.

Seldin, P., Miller, J. E., & Seldin, C. A. (2010). *The teaching portfolio: A practical guide to improved performance and promotion/tenure decisions* (4th ed.). San Francisco, CA: Jossey-Bass.

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*, 598–642. http://dx.doi.org/10.3102/0034654313496870

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives, 8*, 2.

Sproule, R. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review, 21*, 287–294. http://dx.doi.org/10.1016/S0272-7757(01)00025-5

Stark, P. B., & Freishtat, R. (2014). *An evaluation of course evaluations*. doi:10.14293/S2199-1006.1.SQR-EDU.AOFRQA.v1Stark. Retrieved from Science Open: https://www.scienceopen.com/document/vid/42e6aae5-246b-4900-8015-dc99b467b6e4?0

Troy, M. (1995). *Changing the evaluation culture*. Retrieved from http://marsquadra.tamu.edu./TIG/FacultyEvalArticles.ChangingtheEvaluationCulture.ht

Wines, W. A., & Lau, T. J. (2006). Observations on the folly of using student evaluations of college teaching for faculty evaluation, pay and retention decisions and its implications for academic freedom. *William & Mary Journal of Women and the Law, 13*, 167–202.

Wright, R. E. (2006). Student evaluations of faculty: Concerns raised in the literature, and possible solutions. *College Student Journal, 40*, 417–422.

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*, 683–699. http://dx.doi.org/10.1080/02602938.2011.563279

Yeo, R. K., & Li, J. (2013). Beyond SERVQUAL: The competitive forces of higher education in Singapore. *Total Quality Management & Business Excellence, 25*, 95–123.

Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business, 78*, 313–317. http://dx.doi.org/10.1080/08832320309598619