ON THE USE OF COURSE EVALUATIONS FOR PURPOSES OF FACULTY PERSONNEL DECISIONS

H. Paul LeBlanc III, The University of Texas at San Antonio
paul.leblanc@utsa.edu

ABSTRACT

Many universities utilize student teaching evaluations even though their validity is hotly contested in the research. This study examines publicly available data of all sections of courses evaluated in a program at a large public research university over three successive semesters to determine if claims of validity can be confirmed. Additionally, this study investigates other characteristics of the data such as the relationship between course average grade and per course global indicators of teaching and course effectiveness. Given the findings, the study offers policy recommendations for the use of the student teaching evaluations in faculty personnel decisions.

INTRODUCTION

For more than 40 years, universities have been utilizing student reported course evaluations for a variety of purposes including personnel decisions. Indeed, many academic professionals place student course evaluations in their portfolios when on the job market as institutions looking to hire require them. In a groundbreaking study by Rodin and Rodin (1971), results indicated a strong negative correlation between student ratings of an instructor and objective measures of student learning. Rodin and Rodin concluded that "good teaching is not validly measured by student evaluations" (p. 1166).

The research on student evaluation of teaching is controversial. The debate over whether such evaluations are valid measures has not been settled with adherents on both sides of the question. Additionally, questions regarding the application of these measures also have not been answered. With trends toward accountability in higher education increasing, and calls from policy-makers to measure effectiveness and learning outcomes, the need for solid answers to the questions related to student evaluations of teaching has become more acute.

REVIEW OF LITERATURE

The most controversial issue related to the use of student evaluations of teaching (SETs) is its validity. According to d'Apollonia and Abrami (1997), student ratings are only moderately valid. With widespread use of SETs in academia, the utility of these measures have become engrained. In fact, Greenwald (1997) argues that scholars generally agree on the importance of the use of student ratings but disagree on aspects of the validity of those ratings. In investigating the research on SETs, Dowell and Neal (1982) argued that the literature does not support the claim that student evaluations are valid measures of teaching effectiveness. In responding to Dowell and Neal (1982), Cohen (1983) argued that results support overall criterion validity of student ratings. Dowell and Neal (1983) replied to Cohen (1983) by arguing that "student ratings predict a very small proportion of the variance (r = .20) in one important criterion of validity,

student learning" (p. 459). Thus the debate surrounding the validity of SETs seems to hinge on definitions of validity, or the types of validity appropriate or relevant to justify their widespread use for course improvement or personnel decisions.

Onwuegbuzie, Daniel, and Collins (2009) argue that strong evidence exists for criterion-validity in studies about student evaluations of teaching measures but that evidence for content and construct validity is weak. The authors suggest that without this evidence the utility of student evaluations is questionable. However, the controversy deepens as scholars disagree on which form of validity should be primary. For example, Marsh and Roche (1997) recommend using a broad construct-validation approach to student evaluations of teaching, as opposed to the narrow criterion-validation approach.

In order to demonstrate validity, some scholars have investigated more directly the relationship between SETs and grades. In one early study, Brown (1976) found that students' grades significantly influence their ratings of faculty. Following the landmark Brown (1976) study, Gillmore and Greenwald (1999) demonstrated that grading leniency influences higher student ratings. The authors suggested that student ratings for large lecture or strict grading courses should be adjusted statistically. However, these claims are not without their detractors. Centra (2003) discovered that expected grades did not generally influence student ratings. As well, Dee (2007) found that high course workload was not correlated with students' overall rating of instructor. However, it is important to note that this study was conducted in engineering courses where students may already expect workload to be high. It could be that student expectations regarding grades and workload have little to do with individual instructors or courses and could be influenced by other factors.

A common claim among supporters of the use of SETs is that they measure overall teaching effectiveness. For example, Remedios and Lieberman (2008) found that the greatest contributing factor to student ratings of teaching was the perceived quality of teaching. However, Steiner, Holley, Gerdes, and Campbell (2006) posited that several variables outside the control of the instructor introduce bias into student evaluations of teaching.

One such variable is student motivation. Spiel and Gossler (2000) found that student interest in the course was highly correlated with student ratings of courses, and that student interest should be statistically eliminated when evaluating courses. Those motivations of students can play out in ways that decrease or increase SETs for a given course or instructor. For example, in comparing student raters to trained raters or faculty raters, Quirk, et al. (2005) found that student raters were more lenient in evaluating instructors than either of the other two groups. As demonstrated above, the previous research is not conclusive about the internal validity of SETs. More problematical is the lack of research addressing external validity of SETs. External validity deals with the generalizability of the findings of the responders to the classroom population as a whole. Provided the sample size approaches the population size, error may be sufficiently small to make conclusions about the class as a whole. However, if one receives a response rate of less than 60%, "serious biases can be introduced owing to the fact that nonrespondents may differ significantly from those who returned the questionnaire (Blalock, 1979, p. 572-573). In the case of many universities, SETs may be administered within a normal classroom setting. In classroom-administered surveys, response rate is a function of attendance on the day of SET administration. However, when SETs are administered online, response rates may vary widely.

Since SETs are used for descriptive and demonstrative purposes, it stands to reason that both internal and external forms of validity must be met in order to justify their use for individual faculty improvement or personnel decisions. Given this, the current study will attempt to answer these questions:

RQ_1 Can forms of internal validity be demonstrated with a single instrument applied across multiple sections, multiple courses, and multiple instructors within a single discipline?

RQ_2 Can external validity be demonstrated with online administration of a single instrument applied across multiple sections, multiple courses, and multiple instructors within a single discipline?

RQ_3 What other factors may contribute to higher or lower outcomes of SETs when utilizing a single instrument applied across multiple sections, multiple courses, and multiple instructors within a single discipline?

METHODS

To answer the research questions, publicly available data from 169 sections of communication courses over a three-semester period from a large, research-extensive public university were analyzed. The data included lower (n = 64, 37.9%) and upper division (n = 96, 56.8%) undergraduate courses, as well as master's level courses (n = 9, 5.3%). Average class size was 32.09 students (*sd* = 13.37), with a grand total of 5,423 students (with duplication). The average class size for each subset is as follows: (a) Lower division undergraduate courses (*M* = 31.06, *sd* = 14.48), (b) upper division undergraduate courses (*M* = 34.59, *sd* = 11.54), and (c) graduate level courses (*M* = 12.67, *sd* = 3.46).

The instrument utilized was developed over a period of three years through input from multiple faculty groups. The items utilized on the instrument were developed through a collection of possible items from faculty throughout the university, and extensive comparison with other instruments including the IDEA. The final instrument used by the university to measure student perceptions of teaching effectiveness utilized fourteen items. The fourteen items were grouped to measure three underlying constructs: a) course management, b) general learning objectives, and c) instructor characteristics, as were utilized by other instruments purported to measure the same constructs (criterion validity).

To answer the first research question, individual student responses to questions on the online survey were analyzed using Factor Analysis. Toland and De Ayala (2005) utilized Exploratory Factor Analysis to determine the construct-validity of multilevel student evaluations of teaching. Their results indicate that such an approach can demonstrate construct validity.

To answer the second research question, summary data from each section was utilized to determine overall response rate, as well as response rate by course type. To answer the third research question, raw data from each section was analyzed for patterns between course type and other variables (see below). Additionally, summary data from each section was utilized to determine if correlations exist between average course grade and SET outcomes.

RESULTS

Based on the raw data for the Department of Communication, the results failed to indicate support for a claim of internal validity. The online SET instrument was highly reliable (Cronbach's $\alpha = .97$). Additionally, the two global scores (global course rating (G1), and global teaching rating (G2)) were highly correlated ($r = .81$, N = 1653, $p < .001$), and reliable (Cronbach's $\alpha = .88$), thus indicating content validity. However, initial analysis revealed all fourteen items were highly correlated (see Burdsal & Harrison 2008) ranging from the lowest strongly correlated pair ($r = .557$, N = 1634, $p < .001$), to the highest very highly correlated pair ($r = .807$, N = 1630, $p < .001$). Due to the high level of inter-correlatedness, an oblique method of rotation was employed.

Principle Axis Factoring and Direct Oblimin rotation with a .40 loading rule revealed evidence discounting the groupings (thus refuting construct validity). Exploratory Factor Analysis demonstrated that the instrument was unidimensional, and that all fourteen items loaded on a single factor that accounted for 67.98% of the total variance. The range of factor loadings for all fourteen items was .717 to .880. Additionally, the correlation between the computed regression factor score for the one factor and the global rating of teaching was $r = .738$, N = 1610, $p < .01$. The correlation between the single factor and the global rating of the course was $r = .768$, N = 1610, $p < .01$.

To answer the second research question, summary data for all sections of courses for all three semesters was examined (N=169). The results indicate a lack of support for external validity (or generalizability). The overall average response rate for all Communication classes regardless of type of class (lower division, upper division, Master's), type of instructor (tenure-track, non-tenure track, or part-time), or whether the course was a Communication core curriculum course, was 29.87% (sd = 16.98%). When courses with less than 10 students enrolled were eliminated (graduate courses, remote-site distance-learning courses, etc.), the response rate decreased slightly to 29.26% (sd = 16.64%). When examining difference by instructor type, Tenured/Tenure-Track faculty had a slight advantage (33.59%, sd = 18.47) over non-Tenure-Track faculty (26.93%, sd = 15.16%). Overall the maximum response rate achieved was 82.86%. However, only 11 courses out of a total of 167 courses (6.59%) had a response rate greater than or equal to 60%. Additionally, only 61 courses (or 36.1%) had a response rate greater than or equal to 30%, demonstrating potentially serious biases in SET outcomes.

When the average response rates were based on type of course, the results did not improve to meet the external validity criterion within each subgroup (See Table 1, below.) Therefore, the results of the online course surveys may be used for descriptive purposes only, and not for purposes of predicting or assessing student perceptions of teaching effectiveness for courses on the whole.

TABLE 1. PER COURSE RESPONSE RATE (AS A PERCENTAGE) FOR SETS IN COMMUNICATION

Response rate grouping	N (students)	N (surveys)	N (courses)	Resp. Rate M	Min.	Max.
Overall	5416	1602	167	29.45%	2.86%	82.86%
Lower division	1981	487	62	23.89%	3.13%	55.17%
Upper division	3321	1079	96	32.63%	2.86%	82.86%
COM Core (LD & UD)	2504	773	79	29.73%	2.86%	82.86%
Master's	114	36	9	33.88%	14.29%	62.50%

Note: Remote-site DL classes with enrollments below 6 students were eliminated.

To answer the third research question, comparisons were made in ratings of the course and instructor by course level (lower division, upper division, and graduate-level) and instructor type (tenure-track vs. non-tenure track), across all sections with enrollments above 5 students (N = 167). Results failed to find any differences attributable to course level. However, small but significant differences were found between instructor type for global ratings of the course (Tenure-track, M = 3.91, sd = 1.06; Nontenure-track, M = 4.07, sd = .95, $t(1375) = -2.98$, $p = .003$, $\omega^2 = .006$) and teaching of the course (Tenure-track, M = 4.06, sd = 1.12; Nontenure-track, M = 4.18, sd = .97, $t(1375) = -2.17$, $p = .030$, $\omega^2 = .003$).

Interestingly, first prep courses (M = 3.35, sd = 1.18) scored significantly lower than non-first prep courses (M = 4.04, sd = 0.99) on the global ratings of the course, $t(1651) = -6.39$, $p < .001$, $\omega^2 = .024$. As well, first prep courses (M = 3.63, sd = 1.28) scored significantly lower than non-first prep courses (M = 4.14, sd = 1.05) on the global ratings of the instructor, $t(1651) = -4.45$, $p < .001$, $\omega^2 = .011$. Additionally, comparison of the overall averaged global measure (ratings of the course plus ratings of the instructor) to be used for merit purposes for first prep course versus non-first prep courses revealed similar results. First prep courses (M = 3.49, sd = 1.16) scored significantly lower than non-first prep courses (M = 4.09, sd = 0.97) on the overall averaged global measure, $t(1651) = -5.66$, $p < .001$, $\omega^2 = .018$.

As well, courses taught in distance learning mode scored significantly lower on the global measures when compared to courses taught face-to-face. For example, courses taught in distance-learning mode (M = 3.47, sd = 1.10, N = 117) scored significantly lower in overall rating of the course than face-to-face lecture courses (M = 4.09, sd = .97, N = 1269), $t(1444) = 7.75$, $p = .001$, $\omega^2 = .039$. Courses taught in distance-learning mode (M = 3.67, sd = 1.17, N = 117) also scored significantly lower in overall rating of the teaching of the course than face-to-face lecture courses (M = 4.19, sd = 1.02, N = 1269), $t(1444) = 6.26$, $p = .001$, $\omega^2 = .026$.

When examining the relationship between grades and course and instructor ratings, some interesting patterns emerged. To test the relationship, overall course or instructor rating was correlated with overall average course grade for all courses separated by course type and instructor type. For all courses (N = 167), a weak but significant positive correlation was found between course ratings and course grades, $r = .162$, $p = .035$. This relationship did not hold for lower division or graduate level courses. However, a more moderate, positive relationship was found between course ratings and course grades was found for upper division courses, $r = .252$, N = 96, $p = .013$. Instructor type also influenced the relationship between course ratings and course grades. For tenure-track faculty, there was a moderate positive correlation, $r = .394$, N = 62, $p = .002$. However, for nontenure-track faculty, the relationship was significant and negative, $r = -.247$, N = 76, $p = .031$.

For all courses (N = 167), no relationship was found between instructor ratings and course grades. These findings held true regardless of course type (lower division, upper division, and graduate level). However, relationships between instructor ratings and course grades mirrored those between course ratings and course grades. For tenure-track faculty, there was a moderate positive correlation, $r = .297$, N = 62, $p = .019$. However, for nontenure-track faculty, the relationship was significant and negative, $r = -.270$, N = 76, $p = .019$.

DISCUSSION

Considerable work has been done to insure that SETs meet some form of internal validity. Yet the issue with internal validity has to do with whether the measure does what is intended. This study demonstrated the difficulty of claiming internal validity of SETs, particularly as it related to construct validity. Another difficulty with equating student course and instructor ratings with teaching effectiveness is the conflagration of student perceptions of teaching effectiveness and objective measures of teaching effectiveness. This study does little to resolve that problem.

Of greater issue with the use of SETs for the purposes of course improvement and personnel decisions is the problem associated with claiming that these measures represent what the class as a whole perceives to be the characteristics of the course and instructor. As was seen in this study, minimum standards of external validity were not reached in the vast majority of sections of courses. Given the high proportion of courses that do not meet the minimum criterion for external validity, care must be taken in interpreting the results and applying the results for either course improvement or personnel decision purposes.

Although results do not show differences in course and instructor ratings by course type, those ratings differed by whether the course was a first prep for an instructor, was distance-learning, or the tenure status of the instructor. Other factors not investigated here might also influence the ratings. The results demonstrated relationships between grades and course and instructor ratings. The most interesting finding was that positive relationships exist between course grades and course and instructor ratings for tenure track faculty, but the relationships between course grades and course and instructor ratings were negative for non-tenure track faculty. The jury may still be out regarding the magnitude of effect various factors may have on the outcome of SETs. The effect sizes of the results of this particular study were relatively small. Regardless of these findings, universities are likely to continue to use SETs for course improvement and personnel decision purposes. Harrison, Douglas and Burdsal (2004) argue that measures of overall evaluation of teaching was correlated with other measures of teaching and has the advantage of being most understood by faculty as a measure of teaching effectiveness used in making personnel decisions. However, McKeachie (1997) argues that the problem with the use of student evaluations of teaching for personnel decisions lies in the lack of sophistication of the personnel committees who use the ratings. More pointedly, Sproule (2002) argues the exclusive use of student evaluation teaching ratings is pseudoscience and undermines the academic mission of universities. For example, although SETs are used for personnel decisions, such as for merit pay increases, they are often framed as providing an opportunity for increasing teaching effectiveness through course improvement as a main purpose.

Additionally, use of SETs may have unintended consequences. According to Fraley (1998), the merit pay process has had the unintended consequence of lower academic standards.

Eiszler (2002) found that results indicate that student ratings of instruction encourage grade inflation. Additionally, Beran and Rokosh (2009) found that instructor's view students ratings as only marginally valuable in influencing instructional enhancement. According to McNulty, et al. (2010) and Lang and Kersting (2007), regardless of student evaluations of teaching, faculty do not adjust their teaching over time.

Given these findings, several policies regarding use of SETs for personnel decisions are warranted. First, given low response rates, and thus external validity, SETs should be utilized as only one of several measures of teaching effectiveness. Second, given the inability to demonstrate internal validity, SETs should not be the primary measure of teaching effectiveness. Third, whatever measures of teaching effectiveness are being utilized for personnel decisions should be universally applied to all faculty within a department including non-tenure track faculty.

The discipline of Communication has a long history of interest in instruction. The ancient rhetoricians, being itinerant teachers of public address, were keenly interested in the effects of speaker-audience and student-teacher interaction. This study focused on a department of Communication. This focus, perhaps, was a limiting factor in the results. More variability in student evaluations of teaching may occur when other disciplines are included in the research mix. However, the primary findings of this study, i.e., internal validity of the instrument, and external validity associated with response rates, should hold regardless of discipline. The other factors which might influence ratings outcomes should provide fertile ground for future research.

REFERENCES

Beran, T. N., & Rokosh, J. L. (2009). Instructors' perspectives on the utility of student ratings of instruction. *Instructional Science, 37*(2), 171-184.

Blalock, Jr., H. M. (1979). *Social statistics* (rev. 2nd ed.). New York: McGraw-Hill.

Brown, D. L. (1976). Faculty ratings and student grades: University-wide multiple-regression analysis. *Journal of Educational Psychology, 68(*5), 573-578.

Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment & Evaluation in Higher Education, 33*(5), 567-576.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495-518.

Cohen, P. A. (1983). A selective review of the validity of student-ratings of teaching - Comment. *Journal of Higher Education, 54*(4), 448-458.

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

Dee, K. C. (2007). Student perceptions of high course workloads are not associated with poor student evaluations of instructor performance. *Journal of Engineering Education, 96*(1), 69-78.

Dowell, D. A., & Neal, J. A. (1982). A selective review of the validity of student-ratings of teaching. *Journal of Higher Education, 53*(1), 51-62.

Dowell, D. A., & Neal, J. A. (1983). The validity and accuracy of student ratings of instruction: A reply to Peter A. Cohen. *Journal of Higher Education, 54*(4), 459-463.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*(4), 483-501.
Fraley, L. E. (1998). Adverse implications for university teaching concealed in economically driven policies. *Behavior Analyst, 21*(2), 289-305.
Gillmore, G. M., & Greenwald, A. G. (1999). Using statistical adjustment to reduce biases in student ratings. *American Psychologist, 54*(7), 518-519.
Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182-1186.
Harrison, P. D., Douglas, D. K., & Burdsal, C. A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education, 45*(3), 311-323.
Lang, J. W. B., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science, 35*(3), 187-205.
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), 1187-1197.
McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225.
McNulty, J. A., Gruener, G., Chandrasekhar, A., Espiritu, B., Hoyt, A., & Ensminger, D. (2010). Are online student evaluations of faculty influenced by the timing of evaluations? *Advances in Physiology Education, 34*(4), 213-216.
Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity, 43*(2), 197-209.
Quirk, M., Mazor, K., Haley, H. L., Wellman, S., Keller, D., Hatem, D., & Keller, L. A. (2005). Applied research: Reliability and validity of checklists and global ratings by standardized students, trained raters, and faculty raters in an objective structured teaching exercise (OSTE). *Teaching and Learning in Medicine, 17*(3), 202-209.
Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal, 34*(1), 91-115.
Rodin, M., & Rodin, B. (1971). Student evaluations of teachers. *Science, 177*(1), 1164-1166.
Spiel, C., & Gossler, P. M. (2000). The effects of biasing variables on students' ratings of courses. *Zeitschrift Fur Padagogische Psychologie, 14*(1), 38-47.
Sproule, R. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review, 21*(3), 287-294.
Steiner, S., Holley, L. C., Gerdes, K. & Campbell, H. E. (2006). Evaluating teaching: Listening to students while acknowledging bias. *Journal of Social Work Education, 42*(2), 355-376.
Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students evaluations of teaching. *Educational and Psychological Measurement, 65*(2), 272-296.

BUSINESS RESEARCH YEARBOOK

Business and Global Social Responsibility

VOLUME XIX 2012
NUMBER 2

MARGARET A. GORALSKI
H. PAUL LEBLANC, III

EDITORS

A Publication of the
International Academy of Business Disciplines

IABD

BUSINESS RESEARCH YEARBOOK

BUSINESS AND GLOBAL SOCIAL RESPONSIBILITY

VOLUME XIX, NUMBER 2
2012

Chief Editor
Margaret A. Goralski
Quinnipiac University

Associate Editor
H. Paul LeBlanc III
University of Texas at San Antonio

A Publication of the
International Academy of Business Disciplines

IABD

TABLE OF CONTENTS