Using Standardized Student Evaluation Instruments To Measure Teaching Effectiveness In Lecture/Recitation Mode Classes

Gregory C. Potter, Rowan University
George C. Romeo, Rowan University
Da-Hsien Bao, Rowan University
Robert E. Pritchard, Rowan University

ABSTRACT

This paper investigates the variability of student teaching effectiveness survey evaluations among the various recitation sections when lecture/recitation instruction is utilized with the same instructor both delivering the lecture and teaching all of the corresponding recitation sections. The research results indicate that when an instructor teaches multiple sections using lecture/recitation instruction, then the meaningful measure of the instructor's teaching is the average of the student ratings for the various recitation sections. This study focuses on the variability of the students' responses to each item in the survey instrument as measured by its standard deviation.

Keywords: Student ratings, Assessment in higher education, Teaching assessment, Ratings of instruction, Student characteristics, Instructor characteristics, Effective teaching.

INTRODUCTION

Many institutions use one or more types of student evaluations of teaching (SETs) as a part of the recontracting/tenure and promotion processes. The practice, which is long-standing and the object of exhaustive research, has been subject to criticism with respect to the validity and reliability of the instruments, as well as their use by higher education administrators for critical personnel decisions.

This paper, while narrow in scope of investigation, reveals the possibility of bias in what at face value is a straightforward and seemingly uncomplicated SET application. The specific subject of the paper focuses on the interpretation of student survey results when standardized student evaluation instruments are used within the context of the lecture/recitation mode of instruction. More specifically, an investigation is undertaken of the variability of student survey evaluations when lecture/recitation instruction is utilized with the same instructor both delivering the lecture and teaching all of the corresponding recitation sections. The Educational Testing Service's (ETS) SIR II was used as the survey instrument in this study. The results, however, are applicable to other standardized student evaluation instruments as well.

The variability evident in this study underscores the complexity in analyzing what at face value is a surprising statistical result, namely, that the variance in the lecture sessions in many instances was not appreciably smaller than the variance of the recitation sections. Specifically, grouping students into common lecture sessions does not appear to reduce variances significantly; therefore, we should not be surprised if SIR results from a common lecture session and a number of recitation sections of the same instructor vary from recitation section to recitation section.

Most studies of student survey results focus on the mean student ratings instructors have received for each of the items included in the survey instrument. The reason for this attention to mean values is simple: instructors' teaching effectiveness is frequently judged based on those mean values. By contrast, this study focuses on the variability of the students' responses to each item in the survey instrument as measured by its standard deviation.

LITERATURE REVIEW

Although the use and acceptance of student evaluations of teaching (SETs) is widespread in higher education (Ahmadi and Cotton, 1998; Leamon and Fields, 2005; Shevlin et al, 2000), criticisms about objectivity, fairness, and how these instruments are used appear frequently. Wright (2006), for example, reports that student perceptions of fairness and grading and instructor demeanor are strongly related to student evaluation of professors, even though these factors may be unrelated to student teaching. He also suggests that the instructor who provides an "entertainment" experience in class likely will receive a more favorable evaluation.

Related to this phenomenon, Ellis et al (2003) cite extensive research showing a positive correlation between course grades students receive and the ratings they give instructors. This observation, these researchers comment, may explain why grades have risen over the years but aptitude scores have remained relatively stable. Paralleling this observation, Martinson (2000) suggests that a culture of consumerism permeates higher education, evidenced by an entitlement mentality among students and a resulting pressure for leniency on the part of instructors. Similarly, Spooren and Mortelmans (2006) cite research that calls SETs personality contests.

Also prevalent in the literature is the assertion that certain extraneous variables or biases may influence teaching measurement. Shevlin et al (2000) cite both theoretical and psychometric issues surrounding SETs that are unresolved. What, for example, are the nature and number of dimensions that represent teaching effectiveness? External variables, they note, such as student characteristics, lecture behavior, and course administration may confound measurement.

From another perspective, Ahmadi and Farhad (2001) question whether students can judge a class or its methods. Students may be confused about the purpose and value of ratings, often completing forms as quickly as possible. Steiner et al (2006) interpret such variables related to student perceptions as representing biases beyond the instructor's control.

Of the biases that influence SETs, the instructional environment, broadly speaking, is frequently cited. This typically includes courses characteristics (requirements, level, difficulty) (Algozzine et al, 2004; Yunker and Yunker, 2003) and the instructional setting (length of class period, time of day, number of students) (Campbell, Gerdes, and Steiner, 2005; McPherson, 2006).

The instructional environment also involves considerations of whether the class is taught in a conventional classroom setting or online. Terry (2007), in a study of online, campus, and hybrid modes of instruction, reports that student grades, retention results, and course evaluations are lower for online instruction modes compared with campus and hybrid modes.

Much of the SET literature is in agreement that SETs primarily subserve the promotion of learning (Emory, Kramer, and Tian, 2003; Sojka, Gupta, and Deeter-Schmelz, 2002). Although much of this literature takes issue with validity and reliability of instruments as a consequence of biases or influences beyond instructor control, a number of investigators envision a reconceptualization of the teaching-student relationship to promote better feedback and learning. McCormack (2005), for example, raises the question about the ethics of student evaluations. Everyday teaching practice, she remarks, involves making choices, whether to treat evaluations as instruments of organizational control or to view them as new opportunities for student feedback and the promotion of learning. Toward the latter view, Black et al (2004) speak to assessment for learning, which involves examination of the psychology of learning, including motivation and self-esteem issues.

RESEARCH DESIGN

This study was conducted at an AACSB accredited, regional university located in the Northeast from fall 2002 through fall 2006. Two tenured and one untenured accounting professors participated in the study. The study data consisted of the ETS SIR II survey results obtained from five different undergraduate accounting courses. All were taught using the lecture/recitation mode of instruction. In every instance, the professor who delivered the lecture also taught the recitation sections.

The accounting courses met twice each week. One meeting was held in a lecture hall and usually included 60 to 90 students. For the second meeting, students were divided into two or three recitation sections of approximately 20 to 30 students each. All of the recitation sections corresponding to each lecture session were taught on the same day of the week. Teaching the recitation sections on the same day of the week supported uniform material coverage and pace in all sections.

The three participating professors all used common-course syllabi. For each course this resulted in all of the professors following the same course objectives, covering the same material, using the same text, assigning the same homework, using similar tests and grading procedures, etc. Finally, each professor required the same level of work from the students enrolled in each of the lecture session/recitation sections that she/he taught.

Although students met collectively in a lecture session once each week, the SIR II instrument was administered in every recitation section. That is, each recitation section was assessed independently. This study includes the SIR II results from a total of 40 recitation sections, corresponding to 14 lecture sessions. Twelve lecture sessions had three corresponding recitation sections; the remaining two lecture sessions had only two corresponding recitation sections[1].

Seven broad categories including 30 of the SIR II survey items were selected for inclusion in this study. They are listed below:

A. Course Organization and Planning
B. Communication
C. Faculty/Student Interaction
D. Assignments, Exams, and Grading
F. Course Outcomes
G. Student Effort and Involvement
I. Overall Evaluation

The seven categories including the 30 survey items were selected because ETS reports the mean values for each section surveyed for both the categories and the 30 survey items. This information is included as a part of ETS's "Student Instructional Report II." The seven categories and 30 survey items are shown in Table I, Column (1). Since Category I (Overall Evaluation) contains only one survey item, the information for this category is the same as for survey item 40.

For each of the seven categories and 30 survey items, the variability (standard deviation) of the mean student responses corresponding to each of the 40 recitation sections was calculated. These values are shown in Table I, Column (2).

The standard deviation for the 14 lecture sessions was also calculated for each category and for the 30 survey items as follows. The calculation involved two steps. First, the individual standard deviations for the 14

[1] The courses/lecture sessions/recitation sections included: Cost Accounting (3 lecture sessions, 9 recitation sections); Intermediate Accounting I (3 lecture sessions, 9 recitation sections); Intermediate Accounting II (2 lecture sessions, 5 recitation sections); Principles of Accounting I (1 lecture session, 3 recitation sections); and Principles of Accounting II (5 lecture sessions, 14 recitation sections).

individual lecture sessions were calculated. For each survey item the standard deviations were calculated using the means of the corresponding two or three recitation sections. Second, the 14 standard deviations calculated for the 14 lecture sessions were averaged for each category and for the 30 survey items. The results are displayed in Table I, Column (3). The following should be noted:

1. The process of grouping by certain characteristics (commonality) always reduces variances, (i.e., by definition the process of grouping reduces the degree of randomness).
2. If the means for all of the (two or three) recitation sections corresponding to a lecture session were the same, then the standard deviation for that lecture session would be zero.

Finally, for each category and the 30 survey items, the standard deviation for the lecture sessions was divided by the standard deviation of the recitation sections. Table I, Column (4), displays the percent the lecture session standard deviations are of the recitation section standard deviations.

In summary, Table I provides comparative data for the seven categories and the 30 survey instrument items. The columns are organized as follows:

1. The first column lists the seven categories and the 30 SIR II survey instrument items.
2. The second column displays the standard deviation corresponding to each of the seven categories and 30 survey items for all of the recitation sections (n = 40).
3. The third column displays the standard deviation corresponding to each of the seven categories and the 30 survey items for all of the lecture sessions (n = 14).
4. The fourth column displays the percent the lecture session standard deviations are of the recitation section standard deviations (column three divided by column two).

RESEARCH RESULTS

The research results are shown in Table I, "Comparison of Variability Standard Deviations for SIR II Individual Questions: All Recitation Sections in Study versus Groupings of Recitation Sections into Lecture Sessions."

When the lecture/recitation mode of instruction is used, half of the class meetings are held at one time in a lecture hall. Thus, all the students included in the corresponding three (or two) recitation sections were exposed to the same set of experiences during one of the two weekly class meetings. That is, all of the students who attended a given lecture session received the same exposure during one-half of the total course meetings. Consequently, for each of the 30 survey items, one would expect the following:

1. There would be little variation among the mean values of the students' responses to each of the 30 survey items obtained in the two or three recitation sections corresponding to each of the 14 lecture sessions.
2. The standard deviation of the two or three mean values obtained in the two or three recitation sections corresponding to each of the 14 lecture sessions would be quite low.
3. For each of the 30 survey items, the standard deviation associated with the lecture sessions would be much lower than the comparable standard deviation associated with the recitation sections.

Contrary to this expectation, for most of the 30 survey items the research results indicate that this is not the case. For most of the 30 survey items, there was variation among the mean values of the students' responses obtained in the two or three recitation sections corresponding to each of the 14 lecture sessions. This resulted in the much larger than expected values shown in Table I, column (4). In fact, the standard deviations associated with the lecture sessions were less than fifty percent of the standard deviations associated with the recitation sections for only three of the 30 survey items (10 percent of the survey items).

Table I
Comparison of Variability Standard Deviations for SIR II Individual Questions:
All Recitation Sections in Study versus Groupings of Recitation Sections into Lecture Sessions

	Column (1)	Standard Deviations Column (2)	Column(3)	Column (4) (3)/(2)
	SIR II Category/Survey Items	**Recitation Sections (N=40)**	**Lecture Sessions (N=14)**	**Lecture Sessions as Percent of Recit Sect**
A.	**Course Organization and Planning**	**.20**	**.12**	**60%**
	The instructor's			
1.	explanation of course requirements	.23	.16	70%
2.	preparation for each class period	.23	.10	43%
3.	command of the subject matter	.20	.14	70%
4.	use of class time	.25	.16	64%
5.	way of summarizing or emphasizing important points in class	.25	.18	72%
B.	**Communication**	**.25**	**.15**	**60%**
	The instructor's			
6.	ability to make clear and understandable presentations	.31	.24	77%
7.	command of spoken English (or the language used in the course)	.41	.15	37%
8.	use of examples or illustrations to clarify course material	.28	.22	79%
9.	use of challenging questions or problems	.22	.17	77%
10.	enthusiasm for the course material	.33	.15	45%
C.	**Faculty/Student Interaction**	**.22**	**.16**	**73%**
	The instructor's			
11.	helpfulness and responsiveness to students	.22	.17	77%
12.	respect for students	.17	.14	82%
13.	concern for student progress	.31	.19	61%
14.	availability of extra help for this class (taking into account the size of the class)	.33	.27	82%
15.	willingness to listen to student questions and opinions	.22	.18	82%
D.	**Assignments, Exams, and Grading**	**.23**	**.17**	**74%**
16.	The information given to students about how they would be graded	.25	.16	64%
17.	The clarity of exam questions	.32	.25	78%
18.	The exams' coverage of important aspects of the course	.22	.17	77%
19.	The instructor's comments on assignments and exams	.29	.24	83%
20.	The overall quality of the textbook(s)	.30	.19	63%
21.	The helpfulness of assignments in understanding course material	.31	.23	74%
F.	**Course Outcomes**	**.27**	**.18**	**67%**
29.	My learning increased in this course	.28	.23	82%
30.	I made progress toward achieving course objectives	.29	.23	79%
31.	My interest in the subject area has increased	.34	.23	68%
32.	This course helped me to think independently about the subject matter	.29	.19	66%
33.	This course actively involved me in what I was learning	.27	.16	59%

G.	**Student Effort and Involvement**	**.18**	**.13**	**72%**
34.	I studied and put effort into the course	.20	.14	70%
35.	I was prepared for each class (writing and reading assignments)	.20	.14	70%
36.	I was challenged by this course	.22	.19	86%
I.	**Overall Evaluation**	**.28**	**.20**	**71%**
40.	Rate the quality of instruction in this course as it contributed to your learning (try to set aside your feelings about the course content).	.28	.20	71%

Each of the 30 SIR II survey items was analyzed, examining the recitation section standard deviation and comparing it to the corresponding lecture session standard deviation. By calculating the percentages displayed in Table I, Column (4), it is possible to determine if some of the 30 survey items are more or less susceptible to student bias. Refer, for example, to Table I, Survey Item 1, "The instructor's explanation of course requirements." Note that the ratio of column (4) to column (3) is .70 (70 percent). This means that by grouping the students into common lecture sessions, the standard deviation of the lecture sessions is 30 percent less than the standard deviation of the recitation sections (100 - 70 = 30 percent).

This 30 percent reduction is small. The fact that only 30 percent of the variability was reduced by grouping the students into common lecture sessions indicates the following: there were large variations among the students' perceptions in the two or three recitation sections corresponding to each of the 14 lecture sessions pertaining to how effective the three participating professors were in explaining the course requirements.

With all of the students enrolled in a particular lecture session, using the same course syllabus (that details the course requirements), with the same instructor, etc., it would seem reasonable to expect that there would be minimal variability in their responses to Survey Item 1. That is, one would expect that the mean survey values for Item 1 would vary very little among the two or three recitation sections corresponding to each of the 14 lecture sessions. If this expectation was correct, then the standard deviation calculated for the 14 lecture sessions would have been very small compared to the standard deviation calculated for the corresponding 40 recitation sections. Obviously, this was not the case.

Apparently, for Item 1, there were large differences among the mean survey values obtained in each of the two or three recitation sections corresponding to each lecture session. The students' perception of the effectiveness of "The instructor's explanation of course requirements" varied appreciably among the recitation sections corresponding to each lecture section. These differences in students' perceptions might be explained in a variety of ways including the following:

1. Class characteristics: size of the recitation classes, time of day the recitation classes were held, course level (sophomore, junior, senior).
2. Student characteristics: gender, major, expected grade, age, interest in the course, commitment to studying, reading ability.
3. Instructor characteristics: gender, age, teaching methods, teaching experience, rigorous or less rigorous, demanding versus lenient in grading.

Reviewing the entire 30 survey items reveals that there were only three for which the percentage of lecture session standard deviation was less than 50 percent of the recitation section standard deviation. These are listed in order of smallest to largest:

1.	Item 7	Command of spoken English	37%
2.	Item 2	Preparation for each class period	43%
3.	Item 10	Enthusiasm for the course material	45%

The variation in responses among the recitation sections corresponding to the 14 lecture sessions was relatively low for these three survey items. This indicates that, for these three survey items, there was a stronger level of consensus among the students enrolled in the two or three recitation sections corresponding to the 14 lecture sections than for any of the other 27 survey items.

The survey items with the highest percentage of lecture session standard deviation to recitation section standard deviation (greater or equal to 80%) include the following in order of largest to smallest:

1.	Question 36	I was challenged by this course	86%
2.	Question 19	The instructor's comments on assignments and exams	83%
3.	Question 12	Respect for students	82%
3.	Question 14	Availability of extra help for this class	82%
3.	Question 15	Willingness to listen to student questions and opinions	82%
3.	Question 29	My learning increased in this course	82%

For most of the survey items the percentage of lecture session standard deviation ranged between about 50 and 80 percent of the recitation section standard deviation. Examining the survey categories reveals that, by category, the percentage of lecture session standard deviation ranged from about 60 and 74 percent of the recitation section standard deviation.

CONCLUSION

This research indicates that the perceptions among students enrolled in different recitation sections associated with a common lecture session may vary appreciably. As the literature review suggests, a number of biases or influences no doubt are at work in this study. Clearly, this study, which reveals a tightly controlled, narrow universe of investigation, stands as an example of why administrators and others reviewing SET results cannot "fly just by the numbers." Zabaletta (2007) perhaps summarizes this well in the following statement from the conclusion of his study:

...student evaluations show a complex relationship between students and teachers. The components of this relationship are yet to be properly identified. Yet it appears to be very clear that such a relationship has little to do with recognized components of teaching performance such as knowledge, accomplishment of class objectives, expertise in the topic, communications skills, and overall teaching know-how (p.67).

In the particular instance of this study, differences among the various recitation sections suggests that when instructors who use the lecture/recitation instruction mode are being evaluated for purposes of recontracting/tenure, promotion, etc., it is important to note that differences in mean values of student survey items should be expected among the various recitation sections. This being the case, institutions should consider adopting the following two guidelines:

1. Suppose an instructor taught a lecture session and multiple recitation sections and the results for a particular measure (such as overall course evaluation), varied appreciably among the three recitation sections (perhaps ranging 4.5, 4.3, and 3.5 on a five-point scale). Then it is inappropriate to label the instructor as teaching "poorly" in the one recitation section with the low rating. The meaningful measure of the instructor's teaching is the average of the student ratings for the three recitation sections.
2. Suppose instructors are permitted to self-select student evaluations from a number of classes taught. That is, an instructor may select particular class evaluations to be submitted as a part of her/his recontracting/tenure and promotion processes (normally choosing the best class evaluations). If the instructor utilizes lecture/recitation instruction, then, to be meaningful, it is necessary for the instructor to submit the evaluations from all of the recitation sections corresponding to a particular lecture session.

Although the aforementioned guidelines may help ameliorate the potential biases that inhere in studies of this kind, the question is begged, what is an appropriate, fair, and meaningful role for such assessments in higher education? Potter and Pritchard (2007), in an article on assessing communications skills, underscore the complexity of that assessment, including the workload implications for faculty in the development of in-house, content-specific

instruments. Although the development of local assessment instruments and processes might be a daunting task and perhaps ultimately require the services of large-scale publishers, the risk of failing to capture the true picture of the classroom and student-teacher interactions by the use of generic, commercial instruments cannot be denied.

Perhaps McCormack (2005) summarizes the challenge facing higher education in the use of survey instruments by saying that "Evaluation as part of our everyday teaching practice requires us to make choices … [and] those choices, including some associated with students' feedback on our teaching, have ethical dimensions" (p. 465). Encouraging on-going dialog on those ethical dimensions perhaps may be one of the most important dialogs facing higher education.

REFERENCES

1. Ahmadi, R. & Cotton, S. (1998). Assessing students' ratings of faculty. *Assessment Update,* 10(5), 5-7.
2. Ahmadi, M. & Farhad, R. (2001). Business students' perceptions of faculty evaluations. *International Journal of Education Management,* 15(1), 12-22.
3. Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L, Ganesh, M. & Spooner, F. (2004). Student evaluation of college teaching. *College Teaching,* 52(4), 134-41.
4. Black, P., Harrison, C., Lee, C., Marshall, B. & William, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan,* 86(1), 9-21.
5. Campbell, H., Gerdes, K. & Steiner, S. (2005). What's looks got to do with it? Instructor appearance and student evaluations of teaching. *Journal of Policy Analysis and Management,* 24(3), 611-620.
6. Ellis, L., Burke, D., Lomire, P. & McCormack, D. (2003). Student grades and average ratings of instructional quality. *Journal of Educational Research,* 97(1), 35-40.
7. Emery, C., Kramer, T. & Tian, R. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education,* 11(1), 37-46
8. Leamon, M. & Fields, L. (2005). Measuring teaching effectiveness in a pre-clinical multi-instructor course: A case study in the development and application of a brief instructor rating scale. *Teaching and Learning in Medicine,* 17(2), 119-129.
9. Martinson, D. (2000). Student evaluations of teaching and their short term validity. *Journalism & Mass Communication Educator,* 54(4), 77-82.
10. McCormack, C. (2005). Reconceptualizing student evaluation of teaching: An ethical framework for changing times. *Assessment & Evaluation in Higher Education,* 30(5), 463-476.
11. McPherson, M. (2006). Determinants of how students evaluate teachers. *Journal of Economic Education, 37(1), 3-20.*
12. Potter, G. & Pritchard, R. (2007). Learning assurance: Communications abilities – instruments and processes. *Journal of the Academy of Business Education,* 8, 1-12.
13. Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education,* 25(4), 397-405.
14. Sojka, J., Gupta, A. & Deeter-Schmeltz, D. (2002). Student and faculty perceptions of student evaluations of teaching. *College Teaching,* 50(2), 44-49.
15. Spooren, P. & Mortelmans, D. (2006). Teacher professionalism and student evaluation of teaching: Will better teachers receive higher ratings and will better students give higher ratings? *Educational Studies,* 32(2), 201-214.
16. Steiner, S., Holley, L., Gerdes, K. & Campbell, H. Evaluating teaching: Listening to students while acknowledging bias. *Journal of Social Work Education,* 42(2), 355-376.
17. Terry, N. (2007). Assessing instruction modes for master of business administration (MBA) courses. *Journal of Education for Business,* 82(4), 220-225.
18. Wright, R. (2006). Student evaluations of faculty. Concerns raised in the literature and possible solutions. *College Student Journal,* 40(2), 417-422.
19. Yunker, P. & Yunker, J. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business,* 78(6), 313-317.
20. Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education,* 12(1), 55-76.