Instr Sci (2009) 37:171–184
DOI 10.1007/s11251-007-9045-2

Instructors' perspectives on the utility of student ratings of instruction

Tanya N. Beran · Jennifer L. Rokosh

Received: 18 April 2007 / Accepted: 12 November 2007 / Published online: 4 December 2007

Abstract This study investigates instructors' attitudes about student ratings of instruction. The sample consisted of 357 instructors at a major Canadian university, where student evaluation is carried out in all courses at the end of each term. Instructors tend to agree that the student rating practice is an acceptable means of assessing institutional integrity, and is useful to administrators in making summative decisions. However, they consider the student evaluations only marginally valuable in their impact on enhancing instruction. Implications for the development of student ratings are discussed.

Keywords Student evaluation · Student ratings · Teaching instruction

Introduction

First used in Canadian and American universities in the mid-1920s, student ratings of instruction have become integral to accountability in higher education (Zabaleta 2007). The results from student rating instruments have been used to make critical judgments regarding teaching effectiveness in higher education. For example, student ratings have reportedly been used for instructional improvement, promotion and tenure decisions, program evaluation, and student course or instructor selection (Theall et al. 2001). There is considerable controversy surrounding student ratings, which concerns their use in personnel decisions by university administrators. The increased reliance on student ratings data in making personnel decisions such as promotion and tenure is concerning to faculty members who are not convinced of the reliability and validity of the ratings. This skepticism may be due, in part, to concerns raised in anecdotal reports about student ratings being "popularity contests" that are related to a number of negative consequences such as a reduction in faculty morale and grading standards (e.g., Armstrong 1998; Eiszler 2002; Gray and Bergmann 2003; Sproule 2000; Trout 2000). Given this concern, the purpose of

T. N. Beran (✉) · J. L. Rokosh
Division of Applied Psychology, University of Calgary, Calgary, AB, Canada T2N 1N4
e-mail: tnaberan@ucalgary.ca

the present study was to obtain empirical evidence about the extent to which instructors endorse these attitudes regarding student ratings.

Reliability and validity of student ratings

As the reliance on student ratings has increased over time, so has the amount of research attention on the psychometric properties of ratings, particularly reliability and validity. Although the empirical results are not always consistent across studies, researchers have generally supported the reliability and validity of student ratings as measures of teaching performance (Greenwald 2002). With regards to reliability, researchers have found that student ratings of an instructor are stable across items, raters, and time periods (Murray et al. 1990).

Regarding validity, researchers have attempted to demonstrate that student ratings of instruction are related to a variety of other credible indicators of effective teaching (Marsh 1987). Among the criteria examined are measures of student learning, alumni ratings, and ratings of teaching by colleagues, administrators, and outside observers (Feldman 1988; Marsh 1984; Murray 1983; Overall and Marsh 1980). Multisection studies have been used to assess the relationship between student ratings and student achievement in multiple sections of the same course taught by different instructors (Cohen 1981). Researchers correlate the section mean score for student ratings with the section mean for student achievement scores on a common exam. Results typically show substantial correlations between student ratings and student achievement, concluding that students generally give high ratings to instructors from whom they learn most, and generally give low ratings to instructors from whom they learn least (Abrami et al. 1990; Cohen 1981; d'Appolonia and Abrami 1997).

Furthermore, researchers have consistently demonstrated that students tend to form their judgments on the basis of what is taught, and are not overly influenced by extraneous characteristics, such as student characteristics (i.e., ability), course and setting effects (i.e., class size), and extraneous instructor characteristics (i.e., grading standards) (e.g., Arreola 1995; Centra 1993; d'Apollonia and Abrami 1997; Feldman 1993; Marsh and Roche 1997; Theall and Franklin 2001; Wachtel 1998). These studies provide support for the validity of ratings as a measure of effective teaching.

Utility of student ratings

Despite this large body of research supporting the reliability and validity of student ratings of instruction, there remain some unexplored questions about the use of student ratings. Indeed, "the body of literature supporting the validity of student ratings needs to be expanded to include studies of how student ratings are used on today's campuses and what happens as a result" (Ory and Ryan 2001, p. 41). Researchers have begun to explore the methodological validity of student ratings. Methodological validity, also termed "consequential validity" or "utility" refers to the use or application of a measure (Hellman 1998). For a measure to have a high degree of utility, it must provide the type of information necessary for the instrument to be used for its intended purpose. According to Messick (1995), validity refers not only to meanings and interpretations of assessment scores, but also to the inferences and social consequences that result from the evaluation. For student ratings to have adequate utility, they must be shown to be useful for all user groups, namely

the students who complete the evaluations, the administrators who use them for personnel decisions, and the instructors who are being evaluated. A recent study by Beran et al. (2005) offered some preliminary evidence of the "consequential validity" of student ratings of instructors for students, administrators, and instructors indicating that their use varies among these groups. Considering that ratings are used to judge instructors on their performance and have a direct impact on their careers, it is important to investigate how the results are used to improve teaching. Research has demonstrated that instructors' knowledge and attitudes about ratings may influence their use of ratings (Franklin and Theall 1989); therefore, it is important to investigate instructors' attitudes concerning the adequacy of this method of evaluation.

Instructors' attitudes and use of ratings

Contrary to anecdotal reports, which tend to emphasize instructors' negative views of student ratings, the empirical literature to date has revealed a more positive outlook. For example, Schmelkin et al. (1997) found positive attitudes among faculty members regarding their views on the usefulness of teacher evaluations in general. Likewise, Beran et al. (2002, 2005) reported a generally positive attitude among instructors regarding the usefulness of student ratings overall. Thus, perhaps instructors generally agree that there is potential value in student evaluation but may be more negative about the specific uses of evaluation results.

Many instructors support the formative application of student ratings, where ratings are used as a tool for instructional improvement; however, many remain skeptical about their summative application in personnel decisions, such as faculty retention, promotion, salary increases, or tenure. Nasser and Fresko (2002) reported that only slightly more than one-third of instructors felt that administrative heads or relevant committees should be entitled to receive ratings results. Furthermore, the percentage of instructors who indicated that course evaluations should have a critical impact on administrative decision-making was small, ranging from 8% to 23% (Nasser and Fresko 2002). These results suggest that while instructors displayed moderately positive views about student ratings and perceived them to be somewhat useful in improving instruction; in most cases they did not favor the distribution of evaluation results to others, or their prominence in summative evaluation. Comparatively, Schmelkin et al. (1997) reported slightly more positive attitudes of instructors regarding the summative use of ratings results. The majority of respondents did not single out any items as "inappropriate for evaluation" and 60% of respondents listed at least one item as "useful for personnel recommendations."

In sum, instructors typically agree that student ratings of instruction should be used in formative evaluation, but there is less agreement as to whether they should be used for summative evaluation. This may be related to a lack of knowledge about how the ratings data are being used, or concerns that the ratings data are being misused in personnel decisions. The purpose of the present study was to conduct a consequential validity study by obtaining empirical evidence about instructor attitudes regarding student ratings of instruction according to instructors at a major Canadian university where ratings were being used for teaching improvement, promotion and tenure decisions, program evaluation, and student course and instructor selection. This research raises the question of the value of student ratings. Is this a useful process for instructors to improve their quality of teaching, or is it simply, as Ory and Ryan (2001) warn, a routine process performed by students and instructors because it is mandated and has little or no effect on improving teaching

effectiveness? Specifically, the present study examined instructors' attitudes regarding student ratings, and how instructors use these ratings.

Method

Participants

This study was conducted at a major Canadian university of over 20,000 undergraduate students and 5,000 graduate students. The university employs approximately 1,800 full-time faculty and sessionals. A survey designed to evaluate instructor perceptions about the usefulness of student ratings was sent to all full-time faculty and sessionals (N = 1,800). A total of 357 faculty members (215 male—60%; 115 female—32%; 27 not specified—8%) completed the survey, yielding a response rate of 20%. Of these, 107 (30%) were Full Professors, 78 (22%) were Associate Professors, 72 (20%) were Assistant Professors, 76 (22%) were Instructors, and 24 (7%) did not specify. With respect to areas of instruction, the sample represented a variety of faculties and departments in the natural and physical sciences, arts, and professional faculties. The years of teaching experience ranged from 1 to 45 with an average of 15.8 years. Most of the faculty members had taught for 10 years. The final sample closely parallels the university population in terms of demographic variables. The demographic information for participants is summarized in Table 1.

Measures

Instructors were asked to complete a survey designed to assess their views on the usefulness of an institution-wide student rating instrument called the *Universal Student Ratings of Instruction Instrument* (USRI). The USRI is unique to the university where the study was conducted, and, thus, it should be noted that responses are specific to this university population. The USRI was first introduced at the university in 1992 and was designed with the intended purpose of assisting students in course selection, informing instructors about their teaching effectiveness, and assisting administrators in promotion and tenure decisions. The USRI is composed of 12 items that ask students to rate the course and instructor on a 7-point scale ranging from unacceptable to excellent. Examples of the items include: 'I learned a lot in this course,' and 'The course material was presented in a well-organized manner.' University policy mandates that students are required to complete these ratings at the end of every course that they attend. The average rating that instructors reported that they received from students for overall quality of instruction is 5.32 on a 7-point scale where a score of 1 is very low and a score of 7 is very high. The reliability of the USRI, according to Cronbach's α is approximately .92, indicating high internal consistency (Beran et al. 2002).

The faculty survey employed in the present study was developed by a review committee consisting of senior academics from across campus with experience in measurement and teaching. Item selection was based on the intended purposes of the USRI and anecdotal information about how student ratings are being used at the university. The faculty survey consists of three sections. The first section consists of 15 questions related to instructor attitudes about student ratings that focus on various procedural and utility aspects of student ratings (see Table 2 for a list of the dimensions surveyed). Instructors were also given the option of indicating when an item was not applicable to their teaching and were provided space in which to write comments specific to these 15 items. The second section

Table 1 Demographic characteristics of the sample ($N = 357$)

Variable	Frequency	Percentage
Gender		
Male	215	60
Female	115	32
Not specified	27	8
Academic rank		
Assistant professor	72	20.2
Associate professor	78	21.8
Full professor	107	30
Instructor	76	21.3
Not specified	24	6.7
Years Teaching		
≤5	59	16.5
6–10	62	17.5
11–20	84	23.5
21–30	60	16.7
31+	28	7.9
Not specified	64	17.9
Faculty		
Sciences	64	17.9
Humanities	50	14.0
Management	43	12.0
Social sciences	39	10.9
Engineering	36	10.1
Fine arts	15	4.2
Communications	12	3.4
Kinesiology	12	3.4
Medicine	11	3.1
Education	10	2.8
Nursing	6	1.7
Environmental design	5	1.4
Social work	4	1.1
Law	2	.6
Grad studies	1	.2
Not specified	47	13.2
Department		
Greek, Latin, & Ancient History	3	.8
English	17	4.8
Germanic Slavic & East Asian Studies	4	1.1
Philosophy	5	1.4
Religious studies	6	1.7
French, Italian, & Spanish	10	2.8
Biological sciences	16	4.5
Chemistry	10	2.8

Table 1 continued

Variable	Frequency	Percentage
Computer science	5	1.4
Geology/physics	9	2.5
Math & statistics	15	4.2
Physics & astronomy	8	2.2
Anthropology	2	.6
Archaeology	4	1.1
Economics	4	1.1
Geography	5	1.4
Linguistics	1	.3
Political science	3	.8
Psychology	12	3.4
Sociology	2	.6
Educational research	1	.3
Educational psychology	2	.6
Teacher preparation	3	.8
Art	4	1.1
Drama	3	.8
Music	7	2.0
Chemical engineering	8	2.2
Civil engineering	7	2.0
Geomatics engineering	4	1.1
Electrical engineering	5	1.4
Mechanical engineering	7	2.0
Biochemistry & molecular biology	1	.3
Neurosciences	1	.3
Continuing education	1	.3
Not specified	162	45.3

is composed of 8 items that measure the reported use of results to make changes or improvements to specific aspects of teaching (e.g., to improve the quality of teaching). Again, respondents are provided with a space in which to record any comments related to those 8 items. Instructors rated these 23 survey items on a 4-point scale ranging from 1 (strongly disagree) to 4 (strongly agree). Finally, the third section deals with demographic and professional information. Instructors were asked to record their sex, academic rank, faculty, department, and years of teaching experience.

Procedure

The survey, with a cover letter explaining its purpose, was mailed to all active full-time faculty members and sessionals, and once completed was mailed to the Vice-President Academic's office. The survey was administered at the end of a 3-year pilot project (1999–2002) on the implementation and use of the USRI. Over the course of the three years, results from the USRI were reported to instructors individually with printed feedback, sent

Table 2 Instructor mean, frequency, and percentage of agreement ratings ($N = 357$)

	Mean	Strongly disagree	Disagree	Agree	Strongly agree	Not applicable
USRI concepts easily understood	3.25	11 (3%)	20 (6%)	184 (53%)	127 (37%)	5 (1%)
Department head/dean uses USRI appropriately	2.92	27 (8%)	55 (16%)	126 (37%)	85 (25%)	46 (14%)
Support use of student ratings of teaching	3.22	25 (7%)	29 (8%)	142 (40%)	157 (44%)	2 (1%)
USRI results relevant to me	2.66	50 (14%)	91 (26%)	132 (37%)	73 (21%)	8 (2%)
Feedback easily understood	3.11	18 (5%)	35 (10%)	186 (52%)	109 (31%)	6 (2%)
Ratings are intrusive	2.26	64 (18%)	157 (45%)	75 (22%)	38 (11%)	14 (4%)
USRI is a waste of time	2.13	101 (29%)	144 (41%)	52 (15%)	47 (13%)	6 (2%)
USRI good benchmarking tool	2.33	68 (20%)	117 (34%)	119 (34%)	30 (9%)	11 (3%)
USRI results provide useful feedback to me	2.55	59 (17%)	94 (27%)	135 (38%)	57 (16%)	7 (2%)
USRI results consistent with my own assessment	2.81	29 (8%)	73 (21%)	159 (46%)	68 (20%)	17 (5%)
Students should not rate professors	1.90	117 (34%)	169 (48%)	31 (9%)	26 (7%)	6 (2%)
USRI results should be posted on the Web	2.07	119 (35%)	97 (28%)	85 (25%)	28 (8%)	14 (4%)
USRI should be used every course for every term	2.44	75 (21%)	100 (28%)	113 (32%)	57 (16%)	6 (2%)
Normative feedback on USRI results should be given	2.70	32 (10%)	78 (23%)	135 (40%)	53 (16%)	37 (11%)
Scheduling for administering USRI in class is problematic	2.16	63 (18%)	183 (52%)	51 (15%)	33 (10%)	19 (5%)

to administrators and made available to students through postings on the university's web site. Instructors received results that included the mean, frequency distribution, and standard deviation on each rating item for the instructor. Comparisons of the instructor's rating on each item with the corresponding mean and standard deviation for department and faculty instructors at the same level (i.e., junior level, senior level) were also shown. Where it did not compromise student anonymity, mean and standard deviation for each item were provided by gender, required/not-required course, major/non-major, student age, number of prior university/college courses taken, percentage of class attended, rated workload of class, and the student's expected grade in the course. The number of course enrollees and number of valid instruments received for the course were also reported. In addition, the mean student rating of the course workload, as well as the total number of times the instructor taught the course were indicated. Finally, an optional 60-word summary of the course written by the instructor(s) could also be included.

Results

Attitudes about ratings

To determine instructors' attitudes about student ratings, the frequency and percentage of responses are reported from the first section of the scale. These results, as presented in Table 2, indicate that the majority of instructors agreed or strongly agreed that the student

ratings concepts and feedback provided from the USRI are easily understood. Most instructors (n = 299, 84%) support the use of student ratings of teaching in general, while only 16% (n = 57) of instructors feel that students should not rate professors. In addition, many instructors agree that the USRI results are both personally relevant (n = 205, 58%) and useful (n = 192; 54%). Also, most report that USRI results are typically consistent with their own assessment (n = 227, 66%). Sixty-two percent (n = 211) feel that department heads and deans use these results appropriately. However, over half of respondents do not consider the USRI to be a good benchmarking tool (n = 185, 54%). More than half (n = 188, 56%) of the instructors indicated that normative feedback such as rankings should be provided, while most disagree with the results being posted on the web (n = 216, 63%). Instructor attitudes regarding the USRI were generally positive, where most indicated that the USRI is not a waste of time (n = 245, 70%), not intrusive (n = 221, 63%), nor is it problematic to schedule class time for the administration of the USRI (n = 246, 70%). Almost half of the respondents endorsed the use of the USRI for every course taught every term (n = 170, 48%).

To assess the dimensionality of the survey instrument, principal component analyses with Varimax rotation and eigenvalues exceeding unity were performed using instructor responses to the 15-item attitudes scale. As shown in Table 3, a total of 62.07% of the variance was explained by two factors after three iterations. Four items ("I understand the concepts of USRI as presented overall," "My department head/dean uses the USRI results appropriately in assessing my merit increment," "The USRI results should be posted on the web," and "Normative feedback should be provided") were omitted from this final model because their correlations to the other items in the scale were low. When an item had a loading of greater than .40 on two components, the higher loading was used to assign it to a component. The first factor consists of 8 items and appears to measure instructor attitudes about the validity and usefulness of the ratings instrument (USRI). Included were items such as understanding and usefulness of USRI feedback and value of USRI procedure. This factor was called *Attitudes about USRI*. The inter-item consistency (Cronbach's α) of the

Table 3 Means, factor loadings, eigenvalues, and percent of explained variance from principal component analysis with a Varimax rotation (N = 262)

Item	Mean	Attitudes about USRI	Attitudes about ratings
Support use of ratings	3.31	.23	.83
USRI results relevant	2.77	.84	.28
Understand USRI feedback	3.15	.66	.07
Ratings are intrusive[a]	2.81	.51	.66
USRI is a waste of time[a]	2.95	.67	.54
USRI is a good benchmarking tool	2.42	.73	.31
USRI provides useful feedback	2.65	.85	.23
USRI results consistent with own assessment	2.86	.68	.20
Students should not rate professors[a]	3.15	.13	.88
USRI used for every course	2.53	.58	.38
Scheduling USRI is a problem[a]	2.87	.41	.32
Eigenvalue		5.71	1.12
% of variance		37.44	24.63

[a] These items were reverse scored. Only respondents who answered all items are included in this analysis

items comprising this scale was .88. The three items loading on the second factor appear to measure attitudes toward student ratings in general, rather than from ratings from the USRI in particular. This component was called *Attitudes about Ratings* to indicate that they measure attitudes about ratings from any type of ratings method, and the reliability (Cronbach's α) of its items is .82.

Scale scores were then computed by using the sum of the items that loaded highly on each factor, and these values were used for between group comparisons. Accordingly, ANOVAs were conducted to determine if instructor attitudes, either specific to the USRI or about ratings in general, differed according to demographic variables. The analyses showed that all effect sizes are small, <3.0 (Cohen 1987), indicating no meaningful differences for sex, rank, faculty, or years teaching.

Use of ratings

The extent to which instructors considered the USRI to be useful for various aspects of teaching was also explored (see Table 4). USRI results were considered to be most useful for improving general teaching quality (n = 199, 57%), for refining overall instruction (n = 205, 58%), and for improving lectures (n = 191, 54%). According to instructor feedback, the USRI results are least often used for making decisions regarding course textbooks (n = 81, 23%), exams (n = 85, 24%), student assignments (n = 99, 28%), support materials such as audio-visual or reading (n = 121, 34%), or for refining instructional objectives (n = 137, 40%).

The dimensionality of the 8-item subscale assessing instructor use of ratings was analyzed using principle component analysis. A one-factor solution, accounting for 74.4% of the common variance, was obtained, with all 8 items showing a loading of .40 or greater under this factor. This factor was named *Use of Ratings* as the eight items that loaded on this factor reflect the degree to which instructors use the ratings. The inter-item consistency (Cronbach's α) of the eight items on this scale is .95. The mean of 17.88 (SD = 6.19) indicated that, as a group, respondents agreed that they were using the USRI.

The sum of all items that loaded highly on the *Use of Ratings* factor was calculated to obtain a single score for that factor. A series of univariate analyses (ANOVAs) were conducted to explore differences in the use of ratings results according to demographic characteristics. Effect sizes for sex, rank, faculty, and years of teaching were small,

Table 4 Mean and frequency of USRI usefulness for teaching purposes (N = 357)

USRI results used to:	Mean	Strongly disagree	Disagree	Agree	Strongly agree	Not applicable
Improve teaching quality	2.57	61 (18%)	80 (23%)	142 (41%)	57 (16%)	8 (2%)
Select textbooks	1.97	107 (30%)	140 (40%)	64 (18%)	17 (5%)	24 (7%)
Modify exams	2.01	96 (27%)	146 (42%)	71 (20%)	14 (4%)	25 (7%)
Plan assignments	2.06	92 (26%)	144 (41%)	87 (25%)	12 (3%)	17 (5%)
Improve lectures	2.49	69 (20%)	81 (23%)	145 (41%)	46 (13%)	10 (3%)
Select support material	2.17	86 (25%)	129 (37%)	99 (28%)	22 (6%)	13 (4%)
Refine overall instruction	2.56	59 (17%)	78 (22%)	159 (45%)	46 (13%)	8 (2%)
Refine instructional objectives	2.29	71 (21%)	127 (37%)	106 (31%)	31 (9%)	10 (3%)

indicating no meaningful differences in the use of ratings according to demographic characteristics of instructors.

Finally, correlations between the attitude scales and the use scale were analyzed using Pearson Product Moment correlation coefficients. There was a moderate positive correlation between instructors' attitudes about the USRI scale and their use of the ratings, $r = .66$, $p < .01$. Use was also positively correlated with attitudes about student ratings, $r = .45$, $p < .01$. These results indicate that instructors, who reported positive attitudes about student ratings and the USRI itself, were also likely to report using the ratings to improve many aspects of instruction. Also, attitudes about the USRI and ratings were highly correlated, $r = .67$, $p < .001$.

In conclusion, descriptive statistics reveal that instructors' responses to survey items are moderately positive and respondents report that the USRI is meaningful, non-intrusive, and useful. Responses suggest that the USRI results are most useful for improving quality of teaching in general, but the majority of instructors stated that the results are not used for modifying specific aspects of the course or instruction (e.g., modifying exams, text book selection). Generally, instructors who held positive attitudes about ratings in general also held favorable attitudes about the USRI and were likely to use them. Inferential statistics indicated that there are no meaningful differences in attitudes or reported use of ratings based on respondents' sex, rank, teaching experience, or faculty in which they teach.

Discussion

The purpose of this study was to examine the consequential validity of student ratings according to instructors at a major Canadian university. Results indicate that instructors generally hold moderately positive attitudes regarding student ratings and believe them to be useful for the general purposes of improving teaching quality or refining overall instruction, but consider these ratings as having limited use for improving specific aspects of instruction, such as selecting course materials and planning assignments and exams.

Positive attitudes

Examination of instructors' responses to quantitative survey items suggests that instructors have developed moderately positive attitudes toward student ratings of instruction in general. There was strong positive agreement with the statement "In principle I support the use of student ratings of teaching" and almost all instructors disagreed with the statement "students should not rate professors." Approximately one third of instructors strongly agreed that the USRI concepts and feedback are easily understood. Over half of the instructors agreed with the statement "My department head/dean uses the ratings results appropriately in assessing my merit increment," with one quarter of respondents strongly agreeing with that statement. In addition, instructors do not tend to consider ratings to be a waste of time, intrusive, or find it problematic to schedule class time for the administration of the USRI. Therefore, instructors agree that student ratings are an acceptable procedure at the university, and they are confident in their understanding of student ratings concepts and the feedback that they receive. Furthermore, they indicate that they are satisfied with the judiciousness with which their department heads/deans utilize ratings data for merit decisions. Thus, most instructors tend to regard the student ratings as potentially valuable.

Negative attitudes

Although almost half of the instructors seem to endorse the idea of student evaluations and many maintain that they incorporate their results into their own self evaluations, a significant proportion of respondents also gave strong negative responses. While these results appear contradictory, it is not unusual in student ratings research to uncover such variation. For example, Nasser and Fresko (2000) also found a significant lack of consensus in instructors' attitudes toward ratings despite the fact that they had attempted to develop course-appropriate measures.

In the present study, one survey item in particular elicited a strong negative response from instructors. Almost two thirds of instructors do not agree with student ratings results being posted on the web. Nasser and Fresko (2003) reported similar findings in their study, where over half of the faculty members surveyed were strongly opposed to student access of the results. Perhaps they were concerned that publishing the ratings is a violation of their right to privacy and confidentiality. Although only current students can access student ratings results posted on the university web site, some instructors may be unaware about restrictions to access, which may also explain, in part, instructors' apprehension about the ratings. It seems then that despite positive impressions about the general value of ratings, there are strong negative views from instructors about publishing ratings.

Equivocal reactions

While a few items elicited strong positive and negative reactions from instructors, instructors' reactions on the whole appeared ambivalent. For example, while one third of respondents believe that the USRI is not a good benchmarking tool, another third agreed that it was. Benchmarking, as it relates to student ratings, refers to the process of comparing an instructor's performance against the practices of other instructors or against a certain standard of excellence, for the purpose of improving performance. Qualities that are typically believed to be the benchmarks of good teaching include knowledge of subject, enthusiasm, sensitivity to student needs and progress, and preparation (Feldman 1988; Schmelkin et al. 1997; Sherman et al. 1987). Past research has found that faculty members disagree about the characteristics that constitute effective teaching (Schmelkin et al. 1997). Because teaching is a complex task consisting of multiple dimensions, not all faculty members will place the same amount of emphasis on the same aspects of teaching. Therefore, it is likely that there will always be some disagreement among faculty members regarding the adequacy of a single measure of teaching effectiveness.

While just over half of the instructors considered the USRI to be useful for various aspects of teaching such as improving general teaching quality, refining overall instruction, and improving lectures, just under half did not. When responding to whether instructors would use ratings for changes in teaching, only about a quarter to just over a third stated that they would (e.g., changing course textbooks, exams, student assignments, or instructional objectives).

There are many reasons for the limited utility for instructors. According to Centra (1993), it is not enough that instructors receive detailed information on specific behaviors requiring improvement, but they must also gain new knowledge from the information. This is based on findings that have shown improvement to be greatest when the students' evaluation has been very different from the professor's self evaluation. In the present study, over two thirds of instructors considered the student ratings to be consistent with

their own assessment. Thus, instructors themselves may have evaluated their instruction prior to receiving the ratings, and made changes based on their own, rather than on their students' evaluations. Also, unless teachers believe that the student ratings information they receive is salient and the source is respected, they may simply dismiss it. Although the majority of instructors indicated that students should rate instructors, they may doubt the validity of their own evaluations. If they do not personally value student ratings, they are unlikely to use rating information to make changes to their teaching. It is possible that instructors disregard student opinions on matters such as the currency or appropriateness of course material because they do not believe students to be knowledgeable to provide such feedback.

The USRI is an end-of-term tool for assessment used for formative information for teacher professional development and summative evaluation for accountability purposes. It is possible that instructors' apathy toward student evaluations reflects their mistrust of the measure because it is intended to serve both the formative purpose of improving instruction as well as the summative purpose of performance appraisal. Instructors may feel that instructional improvement is not considered to be the most important outcome of student ratings and that their needs are secondary to administrative needs.

Finally, Centra (1993) suggests that a failure to understand how to make significant changes is likely to be the most common barrier to instructional improvement. Even if instructors find the information from student ratings to be relevant, and they are motivated to improve their teaching effectiveness, they may be uncertain as to how to implement useful changes. While the majority of instructors indicated that the student ratings feedback is easily understood, and over half agreed or strongly agreed that the feedback provided by the student ratings instrument is useful, only a few instructors reported having made any substantial modification to individual aspects of instruction as a result of student feedback.

Limitations

Although the present study informs our understanding of the utility of student ratings for instructors, certain limitations should be considered. First, the data were collected at a single university, which may limit generalization to other universities or colleges. Perhaps the items on the ratings instrument do not reflect teaching qualities that instructors value, resulting in negative attitudes about student ratings. Also, the response rate among faculty members was low; thus, the sample may not represent the majority of faculty members' perceptions of student ratings.

Also, issues that can influence the utility of student ratings for faculty members might include the intended use of results (e.g., formative or summative evaluation), the potential distribution of results to other stakeholders, the perceived validity of student ratings, and procedural aspects (e.g., the extent to which students take the process seriously). Moreover, the instructors' perceived adequacy of the USRI as a measure of student ratings may have influenced their opinions about formative utility ("consequential validity"). For example, if instructors thought that the measure provided identifiable dimensions of teaching (e.g., Feldman 1989) and solid instructional factors (e.g., Marsh and Hocevar 1984) they may have held more positive opinions about the results. Future attempts to determine instructor attitudes about student ratings of instruction might include these questions.

Finally, the faculty survey did not include any items assessing instructors' ability to accurately interpret the ratings. Despite reporting that the ratings feedback is easily

understood, it is uncertain as to whether this information is *accurately* interpreted. Franklin and Theall (1990) demonstrated that inaccurate interpretation can lead to the possibility of misuse of ratings data. Furthermore, it is possible that instructors have misinformation about the posting of student ratings results and about the use of results by students and administrators. Knowledge of procedural use of the USRI was not measured and could have identified the role of knowledge on influencing attitudes. Therefore, future research on the usefulness of student ratings for instructors might include items that assess knowledge of procedures and interpretation.

To summarize, participants in this study seemed to agree with the *idea* of students rating instructors. Instructors said that they support the use of student ratings in formative assessment and report that the ratings provide useful information for improving teaching in general and refining overall instruction; however, there was little evidence of application of this ideal to their teaching practice. There was a moderate positive relationship between attitudes and use of ratings, indicating that an instructor who regards student ratings positively, is also likely to find the ratings useful for improving instruction. It is possible that since instructors find ratings to be of little practical value, their seemingly positive attitudes regarding student ratings actually reflects a neutral view point or a passive acceptance of the ratings in general. Regardless of the possibility that student rating forms do not provide the type of information instructors find useful for facilitating specific changes in instruction, they may prefer them to alternative means of evaluation, or see them as a tolerable administrative procedure.

References

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

Armstrong, J. S. (1998). Are student ratings of instruction useful? *American Psychologist, 53*(11), 1223–1224.

Arreola, R. A. (1995). *Developing a comprehensive faculty evaluation system*. Bolton, MA: Anker Publishing.

Beran, T., Violato, C., & Collin, T. (2002). *The Universal Student Ratings of Instruction Instrument at the University of Calgary: A review of a three-year pilot project.* Submitted to the office of the Provost and Vice-President (Academic) at the University of Calgary.

Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The utility of student ratings of instruction for students, faculty, and administrators: A "consequential validity" study. *The Canadian Journal of Higher Education, 35*(2), 49–70.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Fransisco, CA: Jossey-Bass.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Research in Higher Education, 51*(3), 281–309.

Cohen, J. (1987). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*(4), 483–501.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education, 28*(4), 291–344.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*, 583–645.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II – evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*(2), 151–211.

Franklin, J., & Theall, M. (1989). *Who reads ratings: Knowledge, attitudes, and practices of users of student ratings of instruction*. In Paper presented at the 70th Annual Meeting of the American Educational Research Association, San Francisco.

Franklin, J. L., & Theall, M. (1990). Communicating student ratings to decision makers: Design for good practice. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice*. New Directions for Teaching and Learning (No. 43). San Francisco: Jossey Bass.

Gray, M., & Bergmann, B. R. (2003) Student teaching evaluations. *Academe, 89*(5), 44–46.

Greenwald, A. G. (2002). Constructs in student ratings of instructors. In: H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement*. New York: Erlbaum.

Hellman, C. M. (1998). Faculty evaluation by students: A comparison between full-time and adjunct faculty. *Journal of Applied Research in the Community College, 6*(1), 45–50.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253–388.

Marsh, H. W. (1984). Student evaluation of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707–754.

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal, 21*(2), 341–366.

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.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Murray, H. G. (1983). Low inference classroom teaching behaviours and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 71*, 856–865.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology, 82*(2), 250–261.

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education, 27*(2), 187–198.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *New directions for institutional research* (Vol. 109, pp. 27–44). San Francisco, CA: Jossey-Bass.

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

Schmelkin, L. P., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education, 38*(5), 575–592.

Sherman, T. M., Armistead, L. P., Fowler, F., Barksdale, M. A., & Reif, G. (1987). The quest for excellence in university teaching. *Journal of Higher Education, 48*(1), 66–84.

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives, 8*(50). Retrieved on January 21, 2005, from http://www.epaa.asu.edu/epaa/v8n50.html

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a Witch hunt in student ratings of instruction. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *New directions for institutional research* (Vol. 109, pp. 45–56). San Francisco, CA: Jossey-Bass.

Theall, M., Abrami, P. C., & Mets, L. A. (Eds.). (2001). The student ratings debate: Are they valid? How can we best use them? *New Directions for Institutional Research* (Vol. 109, pp. 1–6). San Francisco, CA: Jossey-Bass.

Trout, P. (2000). Flunking the test: The dismal record of student evaluations. *Academe, 86*(4), 58–61.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23*(2), 191–212.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education, 12*(1), 55–76.