The Utility of Student Ratings of Instruction for Students, Faculty, and Administrators: A "Consequential Validity" Study

TANYA BERAN, CLAUDIO VIOLATO, DON KLINE & JIM FRIDERES
University of Calgary

ABSTRACT

Students, faculty and administrators at a major Canadian university were surveyed to investigate the utility or "consequential validity" of student ratings of instructors. Of the 1,229 (approximately equal number of males and females) students and alumni, about half (52%) indicated that they had never used the ratings, but of those who did use it, many (47%) reported using it several times to select courses and/or instructors. The majority (84%) of faculty members ($n = 357$) gave favorable responses about the usefulness of student ratings for improving quality of teaching. Paradoxically, even though faculty members were positive about the student ratings, they did not generally use them to make changes in their teaching. The majority (87%) of administrators ($n = 52$) stated that they use the student ratings for various purposes including decisions about faculty merit and tenure. Students, faculty and administrators considered the overall course instruction to be the most useful type of information derived from the student ratings. The results of the present study indicate that while the utility of data from student ratings of instructors is quite variable, there is evidence of "consequential validity" particularly from administrators.

RÉSUMÉ

Etudiants, professeurs, et administrateurs d'une université canadienne furent interrogés pour enqueter sur l'utilité ou "consequential validity" des évaluations des professeurs par les étudiants. Sur les 1,229 étudiants, environ la moitié (52%) ont répondu qu'ils n'avaient jamais utilisé ces évaluations, mais pour ceux qui les ont utilisées, beaucoup (47%) ont rapporté les avoir utilisées plusieurs fois pour choisir cours et professeurs. La majorité des professeurs ($n = 357$) ont donné des réponses favorables sur l'utilité des évaluations faites par les étudiants pour ameliorer la qualité de l'enseignement. Aussi, la majorité des administrateurs ($n = 52$) ont répondu qu'ils utilisaient ces évaluations a des fins diverses, y compris la qualité et la promotion des professeurs. Etudiants, professeurs, et administrateurs considèrent le jugement global sur le cours être le type d'information le plus utile derivé de ces évaluations d'étudiants. Les résultats démontrent l'existence de "consequential validity" en particulier des administrateurs.

Student ratings of instruction are widely employed in colleges and universities across Canada and the United States (e.g., Greenwald, 2002). Ali and Sell (1998) have noted that student ratings of instruction are one of the most thoroughly studied forms of personnel evaluation, and some aspects of their validity have been studied. Nonetheless, the extent to which such rating information is useful for university students, faculty or administrators remains unclear.

Most previous research has focused on psychometric properties such as reliability and validity of the student ratings instrument as indicators of the quality of teaching and the overall effectiveness of instruction by individual instructors. Reliability is generally adequate but evidence regarding the validity of student ratings has varied (see Arreola, 1995; Kulik, 2001; e.g., Trinkhaus, 2002). While the continued study of the validity and reliability of student ratings of instruction as measures of teaching effectiveness is laudable, a major issue that remains is the utility of the results for students, faculty and administrators. To what extent do students use the results from

these ratings for course and instructor selection for example? How do faculty use the feedback from these ratings? Do administrators such as department heads and deans employ the results of these ratings in decisions about hiring, retention and promotion, and to what extent is student rating information appropriate for such purposes? How are the results from these students' ratings used in universities for improving teaching? The major purpose of the present study was to examine student, faculty and administrator use of the results of an institution-wide or "universal" instrument intended to measure student ratings of instruction at a major Canadian university.

Validity as a Conceptual Framework

Extensive research has been conducted on the psychometric properties of student rating scales, particularly in regard to their reliability and validity (Greenwald, 2002; Hellman, 1998). Although the results are not consistent across all studies, researchers generally agree that student rating scales can measure aspects of teacher effectiveness (Ali & Sell, 1998; Aleamoni & Hexner, 1980; Arreola, 1995; Greenwald, 2002; Hellman, 1998; Marsh, 1987; Marsh & Bailey, 1993; Peterson & Kauchak, 1982). In his review of the research of student ratings of instruction, Greenwald (2002) concluded that most studies conducted between 1971 and 1995 adduced evidence of content and even criterion-related validity (e.g., peer ratings) for these instruments as measures of teaching effectiveness. Similarly, Kulik (2001) in his review reported evidence of criterion-related validity since student ratings are frequently similar to and correlate with results from other measures of teaching effectiveness (e.g., teaching awards, peer ratings).

Although student ratings may measure the quality of the course and instruction, it is not clear how the results of these student ratings are used. If they are intended to measure teacher effectiveness, then ratings could be used in formative evaluation to improve teacher effectiveness by providing feedback to instructors that may lead to behaviour change. They could also be used for summative evaluation of faculty for merit pay, hiring and retention of faculty and promotion decisions. Finally, if rating information is available to students, it could be used to guide course selection.

The validity of a measure can and does vary according to its purpose and use (Violato, McDougall & Marini, 1992). Whereas one purpose of a measure of student ratings may be to obtain an assessment of teacher effectiveness, another may be to provide users with information that can inform their decisions and behaviour in regard to courses and instruction. Each purpose affects the other. A measure cannot be used appropriately – and therefore lacks validity – if it does not measure what it purports to measure. Conversely, a measure will not be valid – quantify what it is intended to measure – if it is not used appropriately for its intended purpose. As each purpose affects the other, both have been referred to as validity (Messick, 1989).

Several types of validity have been specified by researchers (Beran, 2003; Hellman, 1998; Ory & Ryan, 2001). Hellman (1998) referred to the accuracy of a measure in quantifying a construct as "statistical validity", and the use of a measure as "methodological validity". Ory and Ryan (2001) referred to this latter type of validity as "consequential validity" whereby appropriate use of student ratings may lead to desirable or undesirable consequences. Methodological and consequential validity can be more generally referred to as "utility" in that both refer to a measure's application. The use of student ratings – in regard to their methodological or their consequential validity – has received little empirical examination in comparison to statistical validity (how well the student ratings measure teacher effectiveness). The major purpose of the present research was to conduct a consequential validity study by obtaining empirical evidence about how student ratings are used by students, administrators and faculty at a major university.

Consequences of Student Ratings Use

There appears to be substantial concern expressed in the academic community about allowing students access to ratings of instructors (Abrami, 2001). Although this information may facilitate student decisions about course selection, there is a concern that student ratings might also reflect retribution for low grades. While this information may be made public in printed or electronic form (e.g., on Web sites) with the intention of helping students make informed choices about courses or instructors, the extent to

which students actually access this information and for what reasons they use it is not well understood. It is possible, for example, that the statistical information in student ratings does not affect students' decisions about course selection. Although Coleman and McKeachie (1981) found that students registered in a highly rated course more often than a low rated course, Borgida and Nisbett (1977) found that students relied more on comments and anecdotes from other students than on published ratings when selecting courses. It is important to determine how often and for what purpose students use course ratings when they are provided with access to them.

Student ratings alone, however, do not appear to have a large impact on individual instructor teaching effectiveness. In a meta-analysis of the research on changing teaching behaviour after receiving student feedback, L'Hommedieu, Menges, and Brinko (1990) found a small overall effect size of .34 for the improvement of teaching based on feedback from student ratings. These authors concluded that this small improvement suggests little practical value for instructors. We, therefore, wanted to find out from instructors how useful, relevant, and appropriate they consider student ratings to be.

Institution-wide implementation of student ratings at universities may have been initiated for purposes of improving teaching effectiveness (i.e., formative evaluation), but they are also used for personnel decisions (i.e., summative evaluative functions) (Haskell, 1997). As a result, many concerns about administrators' use of student ratings have been expressed (Centra, 1993; Fries & McNinch, 2003; Murray, 1984; Theall & Franklin, 2001; Wagenaar, 1995), particularly if these ratings are the sole source of instructor evaluation in regard to decisions about hiring, promotion, retention, and/or tenure. Indeed, Haskell (1997) stated that student ratings have the second highest weighting value after publications when evaluating university faculty. It has also been suggested that evaluation committees may over-emphasize small rating differences since they are generally not familiar with research on student ratings and, therefore, misuse this information when making decisions affecting individual instructors (Abrami, 2001). Empirical evidence on the relative importance administrators place on student ratings in comparison to other sources of

teaching effectiveness, however, remains scarce.

In summary, several groups (students, faculty, administrators) may use student ratings for a variety of purposes. Determining the appropriateness of these uses requires feedback from the users themselves. Thus, part of an examination of the overall validity of student ratings of instruction (including consequential validity) includes knowing how the user groups actually utilize the information from the assessment. Empirical evidence bearing on this consequential validity of student ratings of instruction, however, is scarce. The major purpose of the present study, therefore, was to conduct a consequential validity study of student ratings of instruction by obtaining empirical evidence about how student ratings are used by students, faculty, and administrators. This actual use of student ratings was then compared with the university's intended purpose for them.

METHOD

The Universal Student Rating of Instruction Instrument

In 1992, a student rating system was introduced at a major Canadian university (undergraduate enrollment > 20,000; graduate > 5,000; full time faculty and sessional instructors > 1,800), with the intended purpose of assisting students in their course selection, informing instructors about their teaching effectiveness, and assisting administrators in promotion and tenure decisions. Based on these intended purposes, and anecdotal information about how student ratings are being used at the university, surveys asking about use of student ratings were developed by a committee of faculty members. Responses to these surveys were analyzed in the present study. They were administered at the end of a 3-year pilot project (1999-2002) on the implementation and use of the Universal Students Ratings of Instruction Instrument (USRI). This scale is composed of 12 items that ask students to rate the course and instructor. Examples of the items include: 'I learned a lot in this course', and 'the instructor is enthusiastic'. Students are asked to complete these ratings at the end of every course that they attend. Over the three years, results from the instrument were reported to instructors individually with printed feedback, and made available to students through postings on the university's Web site. The posted

results included the mean, frequency distribution, and standard deviation on each rating item for the course/instructor. The number of student respondents and course enrollees was also reported. Comparisons of the course/instructor rating on each item with the corresponding mean and standard deviation for department and faculty at the same level (i.e., junior level, senior level) were also shown. In addition, the mean student rating of the course workload, and the total number of times the instructor has taught the course were indicated. Finally, an optional 60-word summary of the course written by the instructor(s) could also be included.

Faculty and administrators were given similar information. The mean, standard deviation, and frequency distribution for each course, instructor and rating item were provided. Also the number of responses and course enrollees were reported as was a comparison of each course, instructor and rating item with the corresponding mean, standard deviation and decile for department, and faculty courses at the same level (e.g., junior or senior level). 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 classes attended, rated workload of class, and the student's expected grade in the course.

PARTICIPANTS

Students

At the time of this study, student ratings for the USRI had been reported to and used by students, faculty, and administrators for three years. At the end of the third year, participants completed surveys about the usefulness of the rating results. From a stratified random sample of classes that represent the various faculties and year of course, 1,700 students were given surveys. A total of 1,194 students completed the surveys (70% response rate). Also, 300 students from a random sample of alumni from the past three years of graduating classes representing all the university programs were sent questionnaires. A total of 35 alumni completed and returned questionnaires (12% response rate). Due to this low return rate, alumni responses were combined with those of current students.

These students and alumni are from various departments and faculties (n = 15; e.g., Education, Medicine, Law, General Studies, Social Sciences, Science, etc.). Of these 1,229 respondents, there was an even representation of males (n = 562, 46%) and females (n = 566, 46%). Another 8% (n = 100) did not specify their gender. The mean age of the respondents was 21.4 years with the most commonly reported age of 19 years and a range of 17 to 54 years. Most of the students were undergraduates (n = 1,067, 87%), and only 4% (n = 44) were graduates (9%, n = 118 students did not specify their status). Also, 86% (n = 1,056) were registered as full-time students, 4% (n = 54) were part-time students, and 9% (n = 119) did not specify. Over half of the respondents were in their first (n = 413, 34%) or second year (n = 257, 21%). There were 180 students in their third year (15%), 177 in their fourth year (14%), and 68 students in their fifth year or more (5%). Another 11% of respondents (n = 134) did not specify their year of study.

A student/alumni survey was administered to both current and previous students of the university. Using open-ended questions, students and alumni were asked to indicate the frequency and purpose of their use of the student ratings (e.g., 'Please indicate how you have used the information collected by the Universal Student Ratings of Instruction.'). A research assistant, who was unaware of the purpose of this study, categorized responses according to their similarity. In addition, respondents were asked to indicate the usefulness of several dimensions of the rating information including that for each of the 12 rating items on a 4-point scale (a list of these dimensions in summarized in Table 1).

Faculty

Surveys were sent to all full time faculty and sessional instructors (N = 1,800). A total of 357 faculty members (215 males – 60%; 115 females – 32%; 27 – 8% did not specify) completed these surveys yielding a response rate of 20%. The characteristics of the faculty respondents were similar to the greater population of instructors at the university. About a third of them (n = 107, 30%) were Full professors, 22% (n = 78) were Associate professors, 20% (n = 72) were Assistant professors, 22%

($n = 76$) were Instructors; 7% ($n = 24$) did not specify. They 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 average rating that faculty members reported that they received from students is 5.32 (on a 7-point scale). Self reported ratings ranged from 1 (very low) to 7 (very high).

Faculty members were asked to complete a 23-item survey regarding the usefulness of the student ratings for purposes of evaluating the quality of their teaching (See Table 2 for a list of the dimensions surveyed). All of these questions were presented on a 4-point scale with a higher score indicating that they strongly agree with the item. Examples of items included, 'In principle I support the use of student ratings of teaching', and 'I feel the Universal Student Ratings of Instruction is not intrusive'. Instructors were also given the option of indicating when an item was not applicable to their teaching.

Administrators

Of all the Deans and Department Heads who received surveys (N = 99), 52 completed and returned the survey (53% response rate). Of these respondents 27 (52%) were a Department Head, 6 (12%), a Dean, and 19 (36%) an Associate Dean. The majority of faculties were represented, although nearly two thirds of the respondents ($n = 33$, 63%) did not indicate their faculty or department.

Administrators completed a survey that asked them to consider the usefulness of the student ratings for various purposes (a list of these can be seen in Table 3). These closed-ended questions included, for example, 'Please rate the usefulness of information provided by the Universal Student Ratings of Instruction for making recommendations regarding faculty merit'. These questions were presented on a 4-point response scale with higher scores indicating greater usefulness. Administrators also were given the option of indicating that any item that was not applicable to their role.

RESULTS

Results from the three surveys are reported separately in this section. In addition to examining the frequency and mean for each response, between group differences were analyzed.

Student/Alumni Survey Results

Students and alumni were asked to indicate how often they used the student ratings, and how useful these results were. About half of the respondents ($n = 694$, 56%) indicated that they did not use the student ratings. Of the respondents who did use the information ($n = 535$, 43%), 31% ($n = 164$) stated that they used it to select a course, 64% ($n = 344$) stated that they used it to select an instructor, and 14% ($n = 73$) reported using it for other reasons such as simple curiosity. Respondents who used the information reported using it once ($n = 69$, 13%), twice ($n = 135$, 25%), or three times ($n = 77$, 14%) with an additional 47% of respondents ($n = 254$) indicating that they used the information four to ten times.

The degree of rated usefulness of several types of information generated by the USRI is shown in Table 1. Students indicated that knowing about the overall instruction of the course was the most helpful information ($M = 3.35$) given to them in comparison to the other items on the scale. Knowing about the detail of the course outline was considered to be the least helpful ($M = 2.70$).

Analyses were carried out to determine if student characteristics are related to ratings use. It was found that the frequency of using ratings information was not significantly correlated with age ($r = .04$, $p > .05$). Univariate analyses of variance for sex, $F(1, 1005) = .90$, $p > .05$, registration as full/part-time, $F(1, 993) = 2.82$, $p > .05$, and undergraduate/graduate status of students, $F(1, 993) = 2.44$, $p > .05$, however, revealed no significant differences. Year of program, however, was significant, $F(4, 977) = 5.83$, $p < .000$. Tukey's Honestly Significant Difference procedure (McCall, 1986) indicated that 5th year students ($M = 7.22$) used the ratings more often than 1st($M = 2.17$), 2nd($M = 3.46$), 3rd($M = 3.03$), and 4th ($M = 3.93$) year students.

Table 1
Mean, Frequency, and Percentage of Student and Alumni Ratings of USRI Usefulness
(*n* = 1,229)

Rating questions	Mean	Very useful	Somewhat useful	Not very useful	Not useful at all
Overall USRI usefulness	2.83	194 (17%)	667 (58%)	201 (17%)	96 (8%)
Overall instruction	3.35	572 (50%)	450 (39%)	91 (8%)	38 (3%)
Detail of course outline	2.70	215 (19%)	484 (42%)	338 (29%)	114 (10%)
Consistency of course with outline	2.80	245 (21%)	539 (47%)	261 (23%)	105 (9%)
Organization of content	3.15	414 (36%)	540 (47%)	147 (13%)	46 (4%)
Responses to student questions	3.12	402 (35%)	524 (46%)	164 (14%)	53 (5%)
Instructor's enthusiasm	3.21	498 (44%)	449 (39%)	144 (13%)	55 (5%)
Opportunities for assistance	3.12	426 (37%)	488 (43%)	179 (16%)	52 (5%)
Respect shown to students	3.16	467 (41%)	450 (39%)	171 (15%)	59 (5%)
Fairness of evaluation	3.30	570 (50%)	398 (35%)	129 (11%)	50 (4%)
Grading time	2.87	284 (25%)	513 (45%)	260 (23%)	86 (8%)
Amount learned in course	2.86	321 (28%)	454 (40%)	260 (23%)	110 (10%)
Helpfulness of support materials	2.77	229 (20%)	521 (46%)	292 (26%)	101 (9%)
Number of students completing USRI	2.71	259 (23%)	425 (38%)	305 (27%)	142 (13%)
Comparison of course rating to Department/ Faculty averages	3.14	419 (37%)	500 (45%)	141 (13%)	51 (5%)
Number of times instructor taught course	3.15	444 (39%)	479 (42%)	146 (13%)	62 (5%)
60-word instructor comments	3.02	363 (33%)	495 (45%)	154 (14%)	92 (8%)

Note. Percentages indicate the number of responses for each category over the total number of respondents who completed the questions to give an indication of the valid percent.

Table 2
Mean and Frequency of Purposes for Instructors Using USRI Results

Purposes	Mean	Not useful at all	Not very useful	Somewhat useful	Very useful	Not applicable
Improve teaching quality	2.57	61 (18%)	80 (23%)	142 (41%)	57 (16%)	8 (2%)
Select course 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 materials	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%)

Note. Percentages indicate the number of responses for each category over the total number of respondents who completed the questions to give an indication of the valid percent.

In summary, about half of the students and alumni indicated that they had never used the student ratings, but of those respondents who did use it, many reported using it several times. The overall instruction of the course was considered the most helpful information from the USRI, and students at the later stages of their programs used the ratings more often.

Faculty Surveys

Faculty members were asked about their opinions regarding the purpose and usefulness of the student ratings. As shown in Table 2, ratings are most often used for improving general teaching quality and instruction, and least often used to make decisions about course textbooks, exams, and assignments.

When rating the degree of usefulness of the ratings, the majority of faculty members stated that the scale's concepts ($n = 311$, 90%) and results ($n = 295$, 83%) are easily understood, used appropriately by department heads ($n = 211$, 62%), useful for teaching ($n = 299$, 84%), relevant to them ($n = 205$, 58%), and consistent with their own assessment ($n = 227$, 66%). Faculty provided generally positive responses with the majority indicating that the Universal Student Ratings of Instruction is not intrusive ($n = 221$, 63%), difficult to administer during class time ($n = 246$, 70%), a waste of time ($n = 245$, 70%), or inappropriate as a student assessment ($n = 286$, 82%).

In summary, responses to the survey items are very positive with the majority of faculty members stating that the USRI is useful, meaningful and non-intrusive. The majority of faculty members also stated that the student ratings are useful for improving quality of teaching in general, but fewer stated that the results are useful for changing specific aspects of their courses (e.g., text book selection, course assignments).

Administrator Surveys

When administrators were asked if they used the information provided by the USRI, 83% ($n = 43$) responded affirmatively, and 15% ($n = 7$) said that they did not. Administrators' rated usefulness in regard to different reasons for using the ratings is shown in Table 3.

Table 3

Mean and Frequency of Purpose for Administrators Using USRI Results

(n = 52)

Purposes	Mean	Not useful at all	Not very useful	Somewhat useful	Very useful	Not applicable
Faculty merit	3.26	1 (2%)	3 (7%)	23 (52%)	16 (36%)	1 (2%)
Tenure	3.10	3 (7%)	1 (2%)	25 (57%)	11 (25%)	4 (9%)
Promotion	3.12	3 (7%)	1 (2%)	26 (59%)	12 (27%)	2 (4%)
Identifying good/poor teaching	3.42	0 (0%)	2 (5%)	21 (49%)	20 (38%)	0 (0%)
Teaching awards	3.31	1 (2%)	5 (12%)	12 (28%)	18 (42%)	7 (16%)
Remediation of teaching problems	3.00	3 (7%)	9 (20%)	13 (30%)	15 (34%)	4 (9%)
Reappointment of sessional instructors	3.17	1(2%)	8 (18%)	16 (36%)	17 (39%)	2 (4%)
Tracking teaching	2.95	2 (4%)	5 (11%)	23 (52%)	7 (16%)	7 (16%)
Assigning courses to faculty	2.11	10 (23%)	13 (30%)	12 (27%)	1 (2%)	8 (18%)
Deciding on timetable	1.83	14 (32%)	14 (32%)	8 (18%)	0 (0%)	8 (18%)
Documenting overall quality of unit's teaching	2.71	5 (9%)	10 (23%)	21 (48%)	6 (14%)	3 (7%)
Analyzing trends in unit's teaching	2.55	4 (9%)	13 (30%)	17 (39%)	4 (9%)	6 (14%)
Promoting the unit	2.39	6 (14%)	13 (30%)	14 (32%)	3 (7%)	8 (18%)

Note. Percentages indicate the number of responses for each category over the total number of respondents who completed the questions to give an indication of the valid percent.

A close inspection of Table 3 reveals that the student ratings were most often used to identify quality of teaching, make decisions about teaching awards, faculty merit, tenure and promotion, and that they were least often used when deciding on the courses to timetable for faculty members.

When indicating the degree of usefulness of the types of information provided by the student ratings, administrators reported that ratings of the overall instruction of the course were the most useful aspect of the instrument (see Table 4). Knowing about ratings of the detail of the course outline and consistency of the course with the outline as well as helpfulness of support materials were considered to be the least useful. This result is consistent with students' feedback that the two items regarding course outlines are the least useful to them in making decisions about courses.

Administrators were also asked about the degree of emphasis they give to various measures used in their unit to evaluate teaching. The level of importance given to the USRI was 46%, followed by a faculty-wide rating instrument (30%), an open-ended comment form (26%), unit-specific rating instrument (17%), or teaching portfolio (15%). Thus, the student ratings were given the most consideration when evaluating teaching.

In summary, the majority of administrators stated that they use the USRI results for various purposes with a primary purpose of identifying the quality of teaching of individual faculty members as well as the overall effectiveness of their unit. Administrators also reported that ratings of the overall course instruction were the most useful type of information derived from the student ratings. Despite the use of alternative faculty teaching performance measures by deans and department heads, the USRI scores are given the most consideration when evaluating teaching instruction.

DISCUSSION

To advance our understanding of the consequential validity of student ratings, we asked students, administrators and faculty to report how often and for what purposes they used student ratings of instruction. Although student ratings were implemented with the intended purpose of assisting students in their course selection, according to the university's policy, only half of the students and alumni indicated that they had used

Table 4

Mean and Frequency of Usefulness of USRI Items for Administrators ($n = 52$)

Items	Mean	Not useful at all	Not very useful	Somewhat useful	Very useful	Not applicable
Overall instruction	3.42	2 (5%)	1 (2%)	17 (40%)	23 (54%)	0 (0%)
Detail of course outline	2.68	3 (7%)	15 (36%)	15 (36%)	8 (19%)	1 (2%)
Consistency of course with outline	2.63	3 (7%)	18 (43%)	11 (26%)	9 (21%)	1 (2%)
Organization of content	3.26	1 (2%)	6 (14%)	16 (37%)	19 (44%)	1 (2%)
Responses to student questions	3.22	1 (2%)	6 (14%)	17 (40%)	17 (40%)	1 (2%)
Instructor's enthusiasm	3.15	2 (5%)	7 (17%)	15 (36%)	17 (40%)	1 (2%)
Opportunities for assistance	3.07	1 (2%)	8 (19%)	19 (45%)	13 (31%)	1 (2%)
Respect shown to students	3.27	2 (5%)	6 (14%)	12 (29%)	21 (50%)	1 (2%)
Fairness of evaluation	2.93	2 (5%)	10 (24%)	18 (43%)	11 (26%)	1 (2%)
Grading time	2.95	1 (2%)	12 (29%)	15 (36%)	12 (29%)	2 (5%)
Amount learned in course	2.97	4 (10%)	8 (20%)	12 (30%)	15 (38%)	1 (2%)
Helpfulness of support materials	2.69	4 (10%)	12 (20%)	15 (33%)	8 (20%)	1 (2%)
Number of students completing USRI	3.00	3 (8%)	5 (13%)	18 (47%)	11 (29%)	1 (3%)

Note. Percentages indicate the number of responses for each category over the total number of respondents who completed the questions to give an indication of the valid percent.

the student ratings. This relatively low frequency may reflect a lack of student awareness about the availability of this information, particularly since it was used several times by those people who did use it. It is also possible that students are not clear about the importance of or how to access student ratings. In addition, some programs such as Engineering have very structured programs in that the students have to take specific courses in their program that are taught by only one professor. In such cases there would be little "use" in going to the USRI Web site. It is also possible that students place more weight on other course information (e.g., course description, course or instructor comments by other students, and feasibility for scheduling) as well as the constraints imposed by program requirements when selecting their courses. Also, students in their 5th or higher year of the program used the ratings more often than students in any other year, suggesting that these more experienced students may have high expectations for their courses. Such students may also be more selective in choosing courses to complete their degree, particularly if they have advanced low-enrollment elective courses to complete where instructor effectiveness is especially germane.

Although various types of information about the instructor and course were made available to students, knowing the rating for the overall instruction of the course was considered the most helpful information from the ratings. As students will not have had previous exposure to the course, they may only be interested in forming a general opinion before beginning the course.

Although faculty members provided strong positive feedback about the usefulness of the student ratings overall, few instructors stated that they actually use the information to change the course. It appears, rather, that they have developed a generally positive attitude about the ratings, finding them appropriate as a means of students providing feedback. This positive attitude may also be due to the generally high ratings (e.g., very good) that instructors receive (Beran, Violato, & Collin, 2002). Although they stated the results are useful for teaching, most instructors used them for general purposes of improving teaching quality or refining overall instruction, rather than for changing specific aspects of the courses (e.g., text book selection, course assignments). This general use may

explain why only a moderate effect size has been found for the usefulness of student ratings to improve teaching effectiveness (L'Hommedieu et al., 1990). Indeed, it appears as if student ratings may have a greater impact on teaching effectiveness when this information is accompanied by specific consultation with others (Cohen, 1980; McKeachie et al., 1980). Student ratings information may also be more useful if instructors obtain information that is relevant to their own courses rather than general information that applies to all university courses. Moreover, since faculty may suspect that characteristics of students (e.g., class attendance) and course type (e.g., lab, lecture) may be related to student ratings, faculty may dismiss the relevance of the ratings.

An alternate explanation is feasible. Considering that people are generally outwardly resistant to change when it is imposed upon them, student ratings may create a neutral reaction (such as general acceptance and tolerance of student ratings) but little acknowledged use of them. However, it is possible that instructors are actually considering the student feedback, accepting it, and recognizing the need for change. This acknowledgement may not be overtly evident on faculty surveys, however.

The majority of administrators stated they used the student ratings for summative purposes. Indeed, despite the availability and use of alternate department, faculty and university measures of teaching effectiveness, administrators depended more often on the student ratings rather than on other sources of information. Their ease of administration, scoring, numeric comparison, and interpretation may explain why they have become the preferred method of evaluation. Consistent with the intended purpose of informing promotion and tenure decisions, administrators are using student ratings as an essential source. Despite cautions by researchers of relying solely on student ratings (Ramsden & Dodds, 1989), and faculty concerns of misuse (Kulik, 2001), more than half of faculty members in the present study stated that their department heads used the student ratings appropriately. Thus, although student ratings have not been embraced by all faculty, the majority of faculty seemed to have little concern regarding how administrators were using student rating information.

Similar to faculty, administrators also provided positive feedback about the ratings. Just as instructors used the student ratings to make

general course improvements, administrators preferred the most general type of student ratings, which are the student ratings of the overall course instruction. Students, moreover, reported that the overall quality of instruction was the most useful information for them in selecting courses. Specific feedback about the course appears then to be less used by students, faculty, and administrators. This result is consistent with the suggestion that faculty may agree with "the idea" of evaluating teaching effectiveness (Murray, 1984, p. 127) but may also have concerns about some of the more specific consequences of their use.

While the results reveal that students, faculty, and administrators use the student ratings for different purposes, results also show that some groups use the ratings more often. In the case of administrators, nearly all respondents use the information. Likewise, most faculty members reviewed the results about their overall teaching activities. On the other hand, over half of the students do not use the results. Perhaps this is due to a lack of familiarity and accessibility to the ratings, program constraints that limit the utility of such information, and/or the dependence of students on their colleagues who access and share such information with them. Thus, it is difficult to determine the consequential validity of student use of the ratings as lack of student awareness will limit their use. In the present case, both faculty and administrators are presented with summary statistics reflecting student ratings. That is, information is sent directly to them, and these results are "normed" so that easy comparisons can be made. There is, therefore, little information cost (i.e., effort and time) to faculty and administrators in obtaining the information. In contrast, students must expend effort and time to access the results. Specifically, they must learn how to locate, log on to, and find each professor's ratings on the Web site. For privacy reasons, the Web site is designed to present professor's ratings separately, which precludes easy and direct comparison across professors. There is, then, more information cost to students in examining student ratings. Students may, therefore, decide to simply ask other students about individual professors. In addition, we do not know how many students who logged on to the Web site informed other students that the data are not useful.

With so many types of questions that can be asked to measure teaching effectiveness, it is possible that the USRI measure does not represent the majority of student rating scales used at other universities. Although the items are similar to the nine factors of the Students' Evaluations of Educational Quality Questionnaire (Marsh & Roche, 1993) that often appear in the research, other rating scales may be used differently. Additional limitations of this study include the low response rate particularly from alumni and faculty. It is important, therefore, that future research examine the consequential validity of other rating scales by determining their utility for additional samples of students, faculty and administrators. ❖

References

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *New Directions for Institutional Research.* (No. 109, pp. 59-87). San Francisco: Jossey-Bass.

Ali, D. L. & Sell, Y. (1998). *Issues regarding the reliability, validity and utility of student ratings of instruction: A survey of research findings.* Calgary, Alberta, University of Calgary: APC Implementation Task Force on Student Ratings of Instruction. Retrieved December 3, 2003, from http://www.ucalgary. ca/UofC/departments/VPA/usri/appendix4.html

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

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

Beran, T. (2003). The role of validity in psychological measurement for school psychology applications. *Canadian Journal of School Psychology, Special Edition, 18*(1/2), 223-243.

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.

Borgida, E., & Nisbett, R. E. (1977) The differential impact of abstract vs. concrete information on decisions. *Journal of Applied Social Psychology, 7*(3), 258-271.

Centra, J. A. (1993, April). The use of teaching portfolios for summative evaluation. Paper presented at the 74th Annual Meeting of the American Educational Research Association, Atlanta.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*, 321-341.

Coleman, J., & McKeachie, W. J. (1981). Effects of instructor/course evaluations on student course selection. *Journal of Educational Psychology, 73*(2), 224-226.

Fries, C. J., & McNinch, R. J. (in press). Signed versus unsigned student evaluations of teaching: A comparison. *Teaching Sociology*.

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.

Haskell, R. E. (1997). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives, 5*(6), 1-34.

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.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *New Directions for Institutional Research*. (No. 109, pp. 9-25). San Francisco: Jossey-Bass.

L'Hommedieu, R. Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology, 82*(2), 232-241.

McCall, R. B. (1986). *Fundamental statistics for behavioral sciences* (4th ed.). London: Harcourt.

McKeachie, W. J., Lin, Y-G., Daugherty, M., Moffett, M., Neigler, C., Nork, J., Walz, M., & Baldwin, R. (1980). Using student ratings and consultation to improve instruction. *British Journal of Educational Psychology, 50*, 168-174.

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. & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education, 64*(1), 1-18.

Marsh, H. W. & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal*, 30(1), 217-251.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (4th ed., pp. 13 - 103). New York, NY: Macmillan Publishing.

Murray, H. G. (1984). The impact of formative and summative evaluation of teaching in North American universities. *Assessment and Evaluation in Higher Education, 9*(2), 117-132.

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*. (No. 109, pp. 27-44). San Francisco: Jossey-Bass.

Peterson, K., & Kauchak, D. (1982). *Teacher evaluation: Perspectives, practices, and promises*. Salt Lake City, Utah: Utah University Center for Educational Practice.

Ramsden, P. & Dodds, A. (1989). *Improving teaching and courses: A guide to evaluation*. Centre for the Study of Higher Education, University of Melbourne. Parkville, Victoria.

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*. (No. 109, pp. 45-56). San Francisco: Jossey-Bass.

Trinkhaus, J. (2002). Students' course and faculty evaluations: An informal look. *Psychological Reports, 91*, 988.

Violato, C., McDougall, D., & Marini, A. (1992). *Educational Measurement and Evaluation*. Dubuque, IA: Kendal/Hunt.

Wagenaar, T. A. (1995). Student evaluations of teaching: Some cautions and suggestions. *Teaching Sociology, 23*(1), 64-68.