

Tanya N. Beran
University of Calgary

and

Jennifer L. Rokosh
Golden Hills School Division

The Consequential Validity of Student Ratings: What do Instructors Really Think?

This study investigates instructors' perceptions about strengths and weaknesses of a student ratings instrument employed in their university. The sample consisted of 357 instructors in a major Canadian university where each term students are required to complete an evaluation at the end of every course. Qualitative analyses of their written responses indicate that most instructors held negative views about the ratings instrument, administration procedures, and use of results. They also reported concerns about biasing factors and the negative effect that ratings have on instructors. Few instructors provided positive comments about the validity of the ratings, the utility of ratings for the user groups, accountability, student representation, and cost efficient administration procedures. Moreover, only 25% considered ratings useful for improving teaching effectiveness.

Cette étude porte sur les perceptions qu'ont les professeurs des forces et des faiblesses d'un outil d'évaluation par les étudiants qui est employé dans leur université. L'échantillon consistait en 357 professeurs d'une grande université canadienne où les étudiants doivent compléter une évaluation de leur professeur après chaque cours. Les analyses qualitatives de leurs réponses écrites indiquent que la plupart des professeurs ont un avis négatif face à l'outil d'évaluation, aux procédures administratives et à l'emploi qu'on fait des résultats. Ils ont également fait part de leurs préoccupations relatives aux facteurs de préjudice et aux effets négatifs qu'ont les évaluations sur eux. Peu de professeurs ont fait des commentaires positifs sur la validité des évaluations, leur utilité pour les groupes d'utilisateurs, la responsabilité, la représentation des étudiants et la rentabilité des procédures administratives. De plus, seulement 25% d'eux estimaient que les évaluations jouaient un rôle dans l'amélioration de l'efficacité de l'enseignement.

According to Marsh (1987), student rating forms are arguably "the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research" (p. 369). Twenty years later, student evaluations have become widely used in universities and colleges and yet remain one of the most debated topics in higher education. Despite the substantial evidence for the reliability and validity of student ratings as indicators of teaching effectiveness (Cohen, 1981; d'Appolonia & Abrami, 1997;

Tanya Beran is an associate professor in medical education in the Department of Community Health Sciences. She is an international presenter and researcher in areas of education, measurement, and evaluation.

Jennifer Rokosh is a provisional psychologist in Calgary. She graduated from the University of Calgary in 2006 with a Master of Science in school psychology from the Division of Applied Psychology. Her research interests are program evaluation and assessment.

Greenwald, 2002; Marsh; Murray, Rushton, & Paunonen, 1990), little research has examined how ratings are being used. Ratings were initially introduced to provide instructors with student feedback for the purpose of improving their teaching. The extent to which instructors are using them for this purpose and the effect of a student ratings program on teaching from the instructors' perspectives is examined in the present study.

Characteristics of Effective Teaching

Critical to the evaluation of teaching effectiveness is a sound understanding of what *teaching effectiveness* means. Although numerous studies have focused on defining the qualities of good teaching, a generally accepted definition of effective teaching has not been identified (Kulik, 2001). Some researchers define effective teaching according to both the process (what teachers do) and outcomes (student learning, Centra, 1993). Alternative definitions refer exclusively to the instructional process (e.g., preparation of material, content knowledge). For example, Arreola (1984) regards teaching as encompassing three broad dimensions: content expertise, instructional delivery skills, and instructional design skills. Lowman (1984) specified effective teaching according to two dimensions: intellectual excitement and interpersonal rapport, whereby intellectual excitement encompasses clarity and presentation of current materials and interpersonal rapport includes showing interest in students as individuals, encouraging creative and independent thought, and being warm, open, predictable, and student-oriented. In a synthesis of 31 studies in which students and faculty were requested to specify characteristics of superior university teachers, Feldman (1988) identified nine characteristics of excellence such as stimulation of interest and speaking skills, intellectual challenges, encouraging independent thought, and motivating students to do their best. Moreover, they described effective teachers as those who show concern and respect for students and display knowledge of the subject matter (Feldman). Thus teaching is a complex task consisting of multiple dimensions (Abrami & d'Apollonia, 1991; Feldman, 1997; Marsh & Roche, 1997). Adding to this complexity is the fact that instructors who are being rated may themselves hold varying beliefs about what constitutes effective teaching.

Many student rating forms used in colleges and universities today are constructed based on studies such as those mentioned above, which have resulted in lists of "ingredients" for teaching excellence. However, Centra (1993) emphasizes that "good teaching is more complicated than any list of qualities of characteristics can suggest" (p. 41). Some traits, Centra suggests, are more readily measured, and, therefore, may be given more weight in an evaluation than necessary. Furthermore, some instructors exemplify teaching behaviors in varying degrees, displaying strengths in some aspects, but not all. In the end, successful teaching is highly dependent on the instructor's theory of how students learn combined with the instructor's beliefs about the teaching behaviors most likely to facilitate student learning based on that theory (Centra). A mismatch between these personal constructs of effective teaching and those constructs measured in student evaluation scales may lead to negative faculty reactions.

Consequential Validity of Student Ratings

Considerable research supports the reliability and validity of student ratings of instruction (Cohen, 1981; d'Apollonia & Abrami, 1997; Greenwald, 1997; Marsh, 1984; Marsh & Hocervar, 1991; Marsh & Roche, 1997). The extent to which student responses to evaluation forms reflect teaching effectiveness is based on a significant premise: that students handle the exercise responsibly. Given concerns that students seem to feel entitled to high marks because they pay tuition (Zimmerman, 2008), faculty may be concerned that they themselves will not be rewarded with high marks if their students are not. These ideas about ratings raise the question, then, of how instructors are using them.

Messick (1989) developed a unified conceptualization of validity that includes consequential validity. He stated, "the key issues of test validity are the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use" (p. 13). In other words, for a measure to be highly useful, it must provide the type of information required to be used for its intended purpose. The intended purpose for student ratings varies across user groups (Beran, Violato, Kline, & Frideres, 2006). Students use the ratings in the selection of courses and instructors, administrators use ratings as a summary measure of teaching effectiveness that is used in making decisions such as promotion and tenure, and instructors use the ratings to improve teaching, as well as course content and structure (Marsh & Roche, 1997).

Instructors' Attitudes and Use of Ratings

Whether instructors consider student ratings useful, and whether they actually use them to improve their teaching, is unclear from the research. Instructors' views on the general utility of student ratings have ranged from strongly supportive to extremely critical (Braskamp & Ory, 1994; Wachtel, 1998). Based on an abundance of anecdotal literature, the general impression seems to be that instructors are resistant to the use of student ratings primarily because of concerns about the quality and legitimacy of the data (Nasser & Fresko, 2002; Schmelkin, Spencer, & Gellman, 1997; Wachtel, 1998). Among the proposed reasons for instructors' opposition to student ratings is the concern that evaluations may be biased by characteristics of the instructor, course, or administration procedures. Although empirical research tends to refute these claims, labeling these as misconceptions or myths (Braskamp & Ory; Cohen, 1990; Seldin, 1993), many instructors remain concerned about the legitimacy of the ratings. The following two studies provide an example as to why these concerns persist. As explained in Cohen, courses that demand a high workload are not usually given low ratings. However, Cohen and Benson (1988) documented an exception to this among students in dentistry whereby high workloads were associated with low ratings. Thus trends and probabilities documented in research are likely to have their exceptions as heard through some published studies, specific cases, and anecdotal reports. These exceptions may significantly affect faculty attitudes, which may explain why ratings have been called "popularity contests" (Naftulin, Ware, & Donnelly, 1973), and are blamed for grade inflation (Greenwald & Gillmore, 1997). It has been suggested that they lead to a reduction in faculty morale and job satisfaction (Ryan, Anderson, & Birchler, 1980), deter innovation (Penny, 2003), and pose a threat

to academic freedom (Haskell, 1997). This latter suggestion is supported only by anecdotal data and implies that the use of student ratings may restrict instructors' comments; they may feel inhibited from challenging students' beliefs or discussing controversial subject matter for fear that students will express their disagreement through the instructor evaluation (Haskell).

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 about their views on the usefulness of teacher evaluations in general. At least 43% agreed with the statement "I frequently make changes in my classes from semester to semester based on student comments," whereas only 14% agreed that, "In general the evaluations do not provide any useful information" (Schmelkin et al.). Similarly, Beran and Violato (2005) reported a generally positive or neutral attitude among instructors about the usefulness of student ratings overall. Also, combined with consultative feedback, student ratings may be helpful in leading to teaching improvement (Cohen, 1980).

Due to a limited amount of current empirical research, it is uncertain whether instructors continue to hold these beliefs about ratings. Considering that a lack of confidence in student ratings may prevent instructors from modifying their courses or teaching styles (DeNisi & Kluger, 2000), understanding how instructors react to feedback might increase the potential for the feedback procedure to enhance performance (Moore & Kuol, 2005).

The present study examines the consequential validity of student ratings to determine the extent to which they beneficially affect teaching effectiveness. Instructors' use of ratings is determined by asking them specific, open-ended questions about how useful they consider the ratings to be. Given their concerns about the validity of the ratings, it is expected that instructors limit their use.

Method

Participants

This study was conducted at a major Canadian university of over 20,000 undergraduate students and 5,000 graduate students. A survey designed to evaluate instructors' perceptions about the usefulness of student ratings was sent to all full-time faculty and sessionals ($N=1,800$). A total of 357 instructors (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 sessional instructors, and 24 (7%) did not specify. The 357 instructors were part of a larger study that also examined administrators' and students' use of ratings. 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 demographic characteristics of the final sample are comparable to the university population and are summarized in Table 1.

Table 1
Demographic Characteristics of the Sample (N=357)

<i>Variable</i>	<i>Frequency</i>	<i>Percentage</i>
<i>Sex</i>		
Male	215	60
Female	115	32
Not specified	27	8
<i>Academic Rank</i>		
Assistant professor	72	20.2
Associate professor	78	21.8
Full professor	107	30
Instructor	76	21.3
Not specified	24	6.7
<i>Years Teaching</i>		
≤ 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
<i>Faculty</i>		
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
<i>Department</i>		
Greek, Latin, and Ancient History	3	.8
English	17	4.8
Germanic Slavic and East Asian Studies	4	1.1
Philosophy	5	1.4
Religious studies	6	1.7
French, Italian, and Spanish	10	2.8
Biological sciences	16	4.5
Chemistry	10	2.8
Computer science	5	1.4
Geology/physics	9	2.5
Math and statistics	15	4.2
Physics and astronomy	8	2.2
Anthropology	2	.6
Archaeology	4	1.1

Table 1 (continued)

<i>Variable</i>	<i>Frequency</i>	<i>Percentage</i>
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 and molecular biology	1	.3
Neurosciences	1	.3
Continuing education	1	.3
Not specified	162	45.3

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 this university, and responses may vary with other measures; however, its items are similar to those found in many such measures. The USRI was first implemented at the university in 1992. It is composed of 12 items that ask students to rate the course and instructor on a 7-point scale ranging from unacceptable to excellent. The items measure overall quality of the course, how much students learned (e.g., "I learned a lot in this course"), how well the course was organized (e.g., "Course content was delivered in a well-organized manner"), usefulness of support materials (e.g., "The course outline provided enough detail"), instructors' ability to communicate (e.g., "Students' questions were responded to appropriately," "Students were treated with respect"), and fairness of evaluation (e.g., "Evaluation methods for determining grades were fair"). These items are comparable to many of the characteristics defined as effective teaching (Arreola, 1984; Centra, 1993; Feldman, 1988; Lowman, 1984). According to university policy, all students are asked to complete these ratings at the end of every course. The average rating that instructors reported that they received from students 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 internal consistency of the USRI as measured by Cronbach's alpha is high at .92 (Beran & Violato, 2005).

A review committee consisting of senior academics from across campus with experience in questionnaire development constructed the faculty survey. It consisted of three open-ended questions asking instructors to report what

they perceived to be the strengths of the instrument, to list potential problems or concerns with it, and to provide suggestions to address any concerns (“Please identify what you perceive to be the strengths of the USRI?” “What do you think are the potential problems or concerns with the USRI?” and “What suggestions do you have to address these concerns?”). The last section requests demographic information including the instructor’s sex, academic rank, faculty, department, and years of teaching experience.

Procedure

The survey and covering letter explaining the purpose of the research were mailed to all instructors. It was administered after three years of using the USRI whereby students completed it after every course, and the results were reported to instructors, sent to administrators, and made available to students through postings on the university’s Web site. Feedback to the instructors consisted of the mean, frequency distribution, and standard deviation on each rating item. These ratings were compared with the mean and standard deviation for department and faculty instructors at the same level (i.e., junior level, senior level). The number of course enrollees and number of valid instruments received for the course were also reported.

Results

Instructors responded to questions about strengths of the USRI, its potential problems or concerns, and suggestions to address these concerns. Of the 357 participants in this study, 76% ($n=271$) provided at least one response, with 67% expressing a negative view ($n=182$). Interestingly, of those who indicated a clear dislike of the USRI instrument, 58% indicated openness to the concept of student evaluation and/or feedback in general.

Perceived Problems

A graduate student with training in qualitative analyses conducted an iterative reading of the responses to the open-ended question about the perceived problems of the USRI to determine several themes (Berg, 2006). Responses were reviewed and coded into response types. A new response type was created whenever a response did not fit into a previously established type. Negative responses were coded into categories reflecting the following six themes: problems associated with the USRI form, problems with administration and use of results, biasing factors, myths and misconceptions, negative effect on instructors and instruction, and other comments. Another 1% indicated that there were no problems (see Table 2).

A sizeable percentage (70%) of instructors expressed concern with the USRI instrument itself. For example, many noted that the USRI instrument is inadequate and provides little or no assistance in instructional improvement. Some noted that the form’s limited usefulness in instructional improvement is primarily due to a lack of written or qualitative feedback. One instructor elaborated, “USRI ratings on their own are not sufficient for making changes to instruction. To improve quality and content of the course, written feedback is also essential.” Many suggested that more trust is placed in individual or faculty rating forms because of the qualitative nature, according to such statements as, “If I had to sacrifice USRI or my faculty written questionnaire it would be the USRI. I gain more insight into how students feel (and why) from

Table 2
 Number and Percentages of Instructors Identifying Problems and Strengths
 (N=271)

	N (%)
<i>Problems</i>	
Poor design of the instrument (e.g., too general, quantitative, anonymous)	190 (70)
Procedural difficulties (e.g., abuse by students, publishing on Web)	151 (56)
Myth-based issues (e.g., students not qualified, popularity contest)	83 (31)
Ratings are biased (e.g., influenced by course difficulty, class size, student motivation)	79 (29)
Negative effect on instructors/instruction (e.g., decrease morale, course standards may be compromised)	30 (11)
Other	21 (7)
None	3 (1)
<i>Strengths</i>	
High validity of ratings (e.g., identify good/weak instructors, obtain course information, student perceptions)	30 (11)
High utility of ratings (e.g., for formative and summative purposes)	91 (36)
Accountability (e.g., holds instructors accountable)	11 (4)
Student representation (e.g., gives students a voice, allows students to vent)	24 (9)
Administration (e.g., ease of administration, universal)	35 (13)
Other	27 (10)
None	30 (11)

the faculty instrument." Others believe particular questions on the USRI are problematic. For example, some commented that the first item on the USRI about the student's perception of overall quality of instruction should be eliminated or moved to the end of the survey. As one instructor explained, "It is never matched by the average of all other questions rated—which clearly indicates that the overall rating does not reflect on what the USRI wants to rate, but rather on students' feelings about either their like or dislike of the course. I always had the average about 1.5 points higher than the overall." Another instructor stated, "The overall instruction question biases all subsequent responses—question ordering is key." Many instructors indicated that the universal nature of the USRI was a problem either because of issues of class size (e.g., small classes not included and larger classes more difficult to teach) or because it is not an appropriate means of evaluating certain types of courses (e.g., Web-based, team-taught). Finally, a number of instructors indicated that the anonymous nature of the USRI was problematic because it does not encourage student accountability. Some instructors felt that, "Some of the students are dishonest. Because they respond anonymously they can write anything," and, "The anonymous set up means disgruntled students see evaluations as a chance to get back at the professor."

The second most frequently cited comment about the USRI was about procedure (56%). Posting of USRI results on the Web was a concern for over 25 instructors. Among the most common objections to this practice is that it is an infringement on instructors' privacy, it could lead to misuse by students, and

that the USRI “should be a useful tool for faculty only.” One instructor elaborated, “I find the postings on the Web shocking. How many professions have a rating of performance made so public?” Issues related to abuse and misuse of the ratings results by administrative heads was another common theme in instructors’ responses. Some report that the USRI is the sole measure being used by their department heads in faculty evaluations. Others contend that administrators focus on only one question (the overall quality of instruction) when basing decisions on the USRI results. Finally, some believe that administrators are not informed in ratings research, nor do they receive appropriate training for reading and interpreting results.

Many find that the USRI is administered too frequently, taking up valuable class time, and resulting in “student rating fatigue.” One instructor concluded “the administration of USRI needs to be less frequent, less time-consuming, and easy to administer by department.” Other instructors feel it unfair to compare one instructor with another, that administration of the USRI takes place either too early in the term for students to have developed a fair opinion of teaching or too late in the term to enable professors to make meaningful changes for those students. Some report that not having access to the results is problematic; however, this might reflect a lack of understanding on the instructors’ part rather than a problem with the USRI, because all faculty members are provided with a copy of their ratings results.

A number of instructors identified flaws in the USRI that were coded as myth-based responses, the third category. These types of attitudes are commonly addressed in the literature as myths or misconceptions that are not supported by empirical literature, yet remain a significant and consistent concern among instructors and other stakeholder groups (Braskamp & Ory, 1994; Cohen, 1990; Seldin, 1993). For example, 13% of instructors included in this qualitative analysis consider the USRI to be an unfair measure as it is purely a “popularity contest” where the instructors who receive high ratings are those who are considered by students to be the most entertaining and popular instructors and not necessarily the most skilled or knowledgeable. Examples of this type of response include: “Good instructors does not equal popular instructors” and “It’s just a popularity poll—as has been documented, students rate most highly those instructors from whom they learn least.”

Another common myth among faculty members is that student ratings of instruction cause grading leniency. A statement reflecting this assumption was endorsed by 12% of respondents. Whereas some merely question the possibility of a grade bias, “It is possible that the higher your average on the midterm, the higher your score,” others are quite certain that giving out higher grades will result in better USRI scores “The obvious problem: students favour courses where they get easy grades” and, “People who water down content and give out less than appropriate workloads and higher grades get higher ratings.”

A number of instructors also feel that students are not qualified to judge teaching effectiveness or that they require time and experience in the workplace to reflect on the instruction received before they are able to render such judgments. Neither of these hypotheses is supported by literature, and thus these types of responses were also coded as myth-based. Almost 30% of

respondents questioned whether student ratings are biased by situational variables. Examples of the most frequently cited potential biases include characteristics of the student (e.g., class attendance, expectations, experience), course (e.g., class size, time course offered, faculty/department of course), and instructor (e.g., rank, physical appearance).

Finally, some instructors listed problems related to the negative effect that the USRI has, or can potentially have, on instruction. Some feel the process negatively affects instruction as it discourages creativity and innovation in the classroom. Some feel pressured to follow the course outline strictly, for if they stray from it in any way, it will be negatively reflected in their student ratings. Others feel that receiving negative results could reduce faculty morale. One instructor described the process as “humiliating and frustrating” and another expressed concern that it might be “intimidating to new teachers.” A number of instructors reported feeling that the student ratings procedure leads instructors to lower their standards to avoid receiving low ratings. For example, one instructor confessed, “I’m sure I could improve my ratings by being more lenient, and if I were a sessional, I would probably give in to the temptation.”

Perceived Strengths

Responses to the open-ended item soliciting strengths of the USRI yielded several themes. A new response type was created whenever a response did not fit within a previously established one. Positive responses were coded into categories reflecting the following six themes: validity of the scores, utility of ratings results, accountability, student representation, administration, and other comments. Another 11% indicated that there were no strengths (see Table 2).

Some instructors (11%) commented on the validity of the ratings, stating that they provide important information about the quality of instruction. Some emphasized that the USRI provides an understanding of the students’ perspectives of the instruction provided. Others indicated that the rating form has good content validity, commenting that the questions are appropriate and provide useful information.

The utility of ratings results was perceived to be a positive feature of the USRI by 36% of instructors who identified strengths of the USRI. Most identified the utility of the information for faculty members in improving instruction in particular. For example, one instructor noted, “The USRI can give useful feedback to professors so that they can choose the appropriate books for students, organize classes effectively, and teach in an inspiring manner.” Others identified the utility of the information for assisting students in course selection and administrators for making personnel decisions. Some consider the ratings instrument to “work in their favor” when it comes to personnel decisions. For example one instructor noted “as a sessional instructor, I appreciate that the department head can track my teaching efforts. I believe that part of the reason I continue to receive teaching contracts is because of my USRI results.” Others seem to recognize the usefulness of ratings for personnel decisions, but still do not feel satisfied with the process. For example, one instructor said, “I can see that the USRI is helpful to department heads, deans, etc., for assessment of teaching performance through students’ eyes but without any real understanding of what is going on.”

Some instructors (4%) referred to the notion of accountability in their responses. One instructor stated, "Tenure can be a very bad thing. Faculty get lazy, unresponsive, and un-engaged. The USRI will, hopefully, disabuse some instructors of the way they see students and their positions." Others focused more on accountability to the students directly, stating that the USRI "forces the instructor to recognize that the students expect and deserve good service" and "It introduces some measure of responsibility towards one's students."

The USRI, according to 9% of respondents, provides students with the opportunity to have their voices heard. These instructors typically commented that the format of the USRI (e.g., anonymous, universal) makes it an important tool in assessing student satisfaction. As one instructor commented, "It provides students with an opportunity to communicate messages difficult to deliver over other channels."

In conclusion, analysis of instructors' comments reveals that most of the instructors who provided written responses held negative views toward the USRI due to issues related to the structure of the USRI form, problems with administration procedures and use of results, biasing factors, other myths and misconceptions (e.g., students unable to judge instruction until post-graduation), and the negative effect that the USRI has on instructors and instruction. Fewer instructors provided positive comments about the validity of the ratings; the utility of ratings results for various user groups; instructor accountability; student representation; and straightforward, cost-efficient administration procedures. In regard as to whether instructors consider ratings useful for improving teaching effectiveness, only 25% of the total sample reported so.

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 although some instructors hold positive attitudes about student ratings and believe them to be useful for the general purposes of improving teaching quality or refining overall instruction, most report concerns about the ratings.

Negative Attitudes

Many negative reactions to student ratings became evident in the written responses. Indeed, more than half of the respondents reported concerns. When asked to identify perceived problems with the universal student rating form, a significant majority of participants in this study identified problems about the structure of the instrument. Specifically, many felt that the questionnaire produced a limited amount of useful information. A rating form is considered good when the information derived from it helps teachers identify their strengths and weaknesses (Centra, 1993). If a rating form is to serve a formative purpose, it requires a detailed and behaviorally oriented set of items to facilitate instructional improvement (McKeachie, 1986). The USRI does contain some items that are targeted at specific areas of instruction such as course materials, exams, and assignment planning. However, these items may not be specific enough for instructors to determine how to improve these areas. According to Centra, one reason why these "diagnostic items" do not lead to greater changes in teaching is because they are not specific enough.

Also in regard to the ratings instrument itself, many instructors indicated that the items were general and not applicable to their style of teaching or course design. For example, some stated that although student rating forms may be useful for the undergraduate lecture mode of teaching, they are not useful for higher-level active learning environments. As a result, many instructors reportedly devised their own methods for evaluation or indicated a preference for their departmental forms that in their opinion garner more useful information. Indeed, the most frequently cited suggestion for improvement to the USRI was to include a qualitative component where students are able to provide more detailed suggestions for instructional improvement. Thus in terms of policy implications, ratings instruments need to include behaviorally specific questions and allow room for students to write specific feedback alongside their ratings.

Some instructors expressed concern that comparing instructors with one another is unfair and can lead to a fall in morale and strained work relations. When using comparative data, half of the sample will necessarily fall below the 50th percentile no matter how small the differences are among scores. The literature speaks compellingly about the dangers of overemphasizing small variations in instructor evaluations, and administrators are advised against using ratings to compare one instructor with another by numerical means (d'Appolonia & Abrami, 1997; McKeachie, 1997). Furthermore, it is often stated that comparisons made between instructors in different faculties are not meaningful given the differences in disciplinary content, focus, and requirements. According to Rifkin (1995), it is best to employ a criterion-referenced system where faculty members are appraised according to a set of standards that encourages professional development, rather than a system that rank-orders instructors on a particular set of items. Theall and Franklin (2001) discussed the importance of including confidence intervals and guidelines for the interpretation of data should normative data be used. If the confidence intervals overlap, users of the data must be aware that it means that there were no significant differences and one cannot conclude that one rating was higher than another. To address a sense of unfairness with the ratings, it is also important that ratings systems be flexible to allow differences across faculties and units for what is regarded as good teaching (Cashin, 1996). These efforts will help instructors accept the ratings and use the added value they can bring to teaching.

Finally, consistent with past research, the present study indicates that many instructors continue to espouse beliefs about the validity of student ratings that have not been demonstrated in the research. In particular, many instructors believe students' evaluations to be biased by a number of factors including course difficulty, instructor popularity, grading leniency, prior student interest, and class size. Although research has consistently shown that most such background characteristics have a negligible effect on student evaluations, instructors continue to hold these beliefs. An important implication is the perception of bias by instructors. If the evaluations are perceived as biased, instructors may be hesitant to use them as important sources of information to facilitate modifications in teaching strategies and teacher behaviors in an effort to make teaching more effective. Approaches that can reduce these biases

include inviting input from instructors and fostering open communication between administrators and instructors about the student evaluation process (Cashin, 1996).

Positive Attitudes

Notwithstanding these concerns, many instructors indicated that ratings are an important means of having students' opinions heard, suggesting that students require and deserve an anonymous and consistent means of reporting their classroom experiences. Some instructors commented on the value of student ratings as a means of holding instructors accountable and encouraging self-reflection, whereas others find the standardized and universal form to be a clear and expedient means of providing feedback. Moreover, a quarter of the respondents did report using the ratings as means of improving instruction.

Limitations

Despite the insights gleaned through instructors' comments about student ratings, the present study is not without its limitations. First, the response rate was low, and perhaps only instructors with particularly negative (or positive views) took the time to complete the survey. Second, instructors were asked to report their perceptions of the student ratings instrument. Despite showing good validity and reliability, this measure may not adequately address key aspects of teaching that instructors value. Also, it may have provided a unidimensional rather than multidimensional assessment of teaching. Perhaps knowing this latter information would have been useful for instructors. Perceptions from instructors at other universities with different measures should be compared to determine if they were unique to instructors at this particular university.

Another limitation of the study, and one that may affect consequential validity for instructors, is their perception that the ratings are used for accountability purposes primarily. Although instructors at this university are not formally required to demonstrate instruction improvement, this perception of top-down direction may create resistance on the part of instructors to use student ratings. Thus measurement, in addition to administrative regimes, may be implicated in the consequential validity of student ratings.

In conclusion, the analysis of instructors' written responses suggests a wide variability in knowledge, beliefs, and attitudes about student ratings of instruction that can affect willingness to use ratings to improve teaching. Instructor evaluation, including the use of student ratings, is understandably a sensitive issue given that negative feedback may be provided. Our study reveals the critical importance of consistency between what instructors consider to be quality teaching and the measures used to assess them.

References

- Abrami, P.C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness—Generalizability of "N = 1" research: Comment on Marsh (1991). *Journal of Educational Psychology, 30*, 221-227.
- Arreola, R.A. (1984). Evaluation of faculty performance. In P. Seldin (Ed.), *Changing practices in faculty evaluation: A critical assessment and recommendations for improvement* (pp. 79-85). San Francisco, CA: Jossey-Bass.
- Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment and Evaluation in Higher Education, 30*, 593-601.

- Beran, T., Violato, C., Kline, D., & Frideres, J. (2006). The utility of student ratings of instruction for students, faculty, and administrators: A "consequential validity" study. *Canadian Journal of Higher Education*, 35(2), 49-70.
- Berg, B.L. (2006). *Qualitative research methods for the social sciences* (6th ed.). New York: Pearson.
- Braskamp, L.A., & Ory, J.C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco, CA: Jossey-Bass.
- Cashin, W.E. (1996). *Developing an effective faculty evaluation system*. Manhattan: Center for Faculty Evaluation and Development, Division of Continuing Education Kansas State University.
- Centra, J.A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass.
- 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.
- 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, P.A. (1990). Bringing research into practice. In M. Theall & J. Franklin (Eds.), *New directions for teaching and learning* (pp. 123-32). San Francisco, CA: Jossey-Bass.
- Cohen, P.A., & Benson, B.A. (1988). Workload and student course ratings in dental school. *Journal of Dental Education*, 52(2), 98-101.
- d'Apollonia, S., & Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist*, 52, 1198-1208.
- DeNisi, A.S., & Kluger, A.N. (2000). Feedback effectiveness: Can 360-degree appraisals be improved? *Academy of Management Executive*, 14(3), 129-139.
- 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. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R.P. Perry & J.C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368-395). New York: Agathon Press.
- Greenwald, A.G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, 52, 1182-1186.
- 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* (pp. 277-297). New York: Erlbaum.
- Greenwald, A.G., & Gilmore, G. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52, 1209-1217.
- Haskell, R.E. (1997). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives* 5(6), 1-35.
- 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* (pp. 9-25). San Francisco, CA: Jossey-Bass.
- Lowman, J. (1984). *Mastering the techniques of teaching*. San Francisco, CA: Jossey-Bass.
- McKeachie, W.J. (1986). *Teaching tips: A guidebook for the beginning teacher* (8th ed.). Lexington, MA.: Health.
- McKeachie, W.J. (1997). Student ratings: The validity of use. *American Psychologist*, 52, 1218-1225.
- 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. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11(3, special issue).
- Marsh, H.W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education*, 7, 303-314.
- 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, 1187-1197.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (4th ed., pp. 13-104). New York: Macmillan.
- Moore, S., & Kuol, N. (2005). Students evaluating teachers: Exploring the importance of faculty reaction to feedback on teaching. *Teaching in Higher Education*, 10(1), 57-72.
- 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, 250-261.

- Naftulin, D.H., Ware, J.E., & Donnelly, F.A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48*, 630-635.
- Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment and Evaluation in Higher Education, 27*, 187-198.
- Penny, A.R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education, 8*, 399-411.
- Rifkin, T. (1995). *The status and scope of faculty evaluation*. Washington, DC: Office of Educational Research and Improvement.
- Ryan, J.J., Anderson, J.A., & Birchler, A.B. (1980). Student evaluation: The faculty responds. *Research in Higher Education, 12*, 317-333.
- Schmelkin, L.P., Spencer, K.J., & Gellman, E.S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education, 38*, 575-592.
- Seldin, P. (1993). The use and abuse of student ratings of professors. *Chronicle of Higher Education, 40*(1), A40.
- 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* (pp. 45-56). San Francisco, CA: Jossey-Bass.
- Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23*, 191-212.
- Zimmerman, (2008, February). Course evaluation—Students' revenge? *University Affairs*.